UNIVERSIDAD NACIONAL DE LA PLATA

FACULTAD DE INFORMÁTICA

TESIS DOCTORAL

CLASIFICACION AUTOMATICA BASADA EN ANALISIS ESPECTRAL

CASO DE USO: PROCEDIMIENTOS CLASIFICATORIOS APLICADOS A OBSERVABLES DE LOS PROBLEMAS TAXONÓMICOS

Autor: Gregorio Perichinsky

Director: Ángel Luís Plastino

2007

UNIVERSIDAD NACIONAL DE LA PLATA

FACULTAD DE INFORMÁTICA

TESIS DOCTORAL

CLASIFICACION AUTOMATICA BASADA EN ANALISIS ESPECTRAL

CASO DE USO: PROCEDIMIENTOS CLASIFICATORIOS APLICADO A ASTEROIDES

Tesista: Gregorio Perichinsky
Lic. en Ciencias Físicas.

Director: Ángel Luís Plastino
Dr. en Ciencias Físicas.

2007

TABLA DE CONTENIDOS

DEDICATORIA

A mi familia

PREFACIO

Esta Tesis Doctoral surge como una necesidad de completar una trayectoria iniciada cuando en la Universidad Nacional de La Plata las Ciencias Duras tenían sus aposentos en la Facultad de Ciencias Físico Matemáticas, hasta que se crean las Facultades de Ingeniería, Ciencias Exactas, etc.

Eran épocas épicas tanto en la política estudiantil como en lo científico, en la defensa de la Universidad de la Reforma y en contra de cierto oscurantismo ligado al cientificismo.

En las "catacumbas del Departamento de Física" se entretejían la política del hombre digno del "Che Guevara" y el "aburrimiento" dentro de la duda metódica, el conteo de partículas radiactivas en un espectroscopio de coincidencias rápido lentas para terminar la tesis de diploma, sin horario y días feriados, la obsesión científica.

Luego el Doctorado en Ciencias Físicas con la Tesis en Física Teórica con el Prof. Dr. Ángel Luís Plastino, el amigo con el cual discutíamos también de política, y ... la **noche de los bastones largos** y la computación apareciendo en la Argentina a través del cálculo...y de nuevo las discusiones sobre tecnología, ciencia y la máquina-herramienta.

El honor y la vergüenza no son inherentes a ninguna condición. Hacer lo que corresponde, en ello consiste el honor: el decoro universitario.

La Computación, la Informática como un nuevo panorama en el desarrollo profesional y el aporte desafiante de la creación de un centro de estudios de la información, y los amigos colegas sugiriendo que intervenga... y el cambio hacia las Ciencias de la Computación, que no sabía que era, pero el "feeling" de que todas las disciplinas la necesitaban y necesitan que hasta podía ser una tecnología emergente para grupos interdisciplinarios.

Lo demás fue vertiginoso, el aprendizaje, la conducción del Centro de Estudios para el Procesamiento de la Información, la lucha por el reconocimiento de lo emergente por los de siempre, de nuevo... **la dictadura**, y luego... de nuevo casi volver a empezar en la investigación científica.

Primero el Departamento de Informática de la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata y luego el desafío de la Facultad de Ingeniería de la Universidad de Buenos Aires donde había que tratar que el

Departamento de Computación tuviera las “tres patas” que tantas veces dijera el Prof. Dr. Ángel Luís Plastino, la Docencia, la Generación de Conocimiento con la Investigación Científica y la Vinculación con la Sociedad, la Transferencia Tecnológica [Bunge, M. 1999] [Perichinsky, G. Investigation, 1995] [Nagel, E. 1968].

A pesar de mi mismo y pese a sus contradicciones voy a parafrasear a Ernesto Sábato “el proceso cultural es un proceso de domesticación que no puede llevarse sin rebeldía por parte de la naturaleza animal, ansiosa de libertad”.

El hecho es que la imagen de un hombre no es que se vea bueno ni malo, ni grande ni trivial, pero sí que está elaborado para satisfacer “exageradas expectativas” de la grandeza humana.

Las cosas que parecen más justas y simples son, en definitiva, las que se revelan más oscuras y difíciles.

Tener que formalizar todo lo bueno y todo lo malo, tener que hacer el “Doctorado en Ciencias de la Informática”, es parte de mi trayectoria.

La interdisciplina, en un problema epistemológico [Gianella, A. E. 2000] [Klimovsky, G. 1994] que combina simultáneamente su vaguedad con la importancia filosófica que posee, es el problema de la reducción, vinculado a cierta postura filosófica es el Reduccionismo Metodológico, que implica la afirmación de que objetos o ámbitos de cierta naturaleza pueden, al fin, definirse o caracterizarse en términos o en componentes que corresponden a otro ámbito, de naturaleza distinta.

Se puede ubicar este problema dentro de la Explicación Científica [Hempel, C. G., 1996] además del Ontológico y Semántico, el Reduccionismo Metodológico, significaría una reducción semántica del lenguaje de teorías al lenguaje de otras, con la resultante de que, unas sean derivadas de las otras, con la dependencia deductiva de unas con relación a otras.

Y así de simple y de oscuro se lo “conté” a Plastino, y apareció la interdisciplina y la necesidad de generar conocimiento, **introduciendo la Informática en consuno con otras disciplinas...**frente a un problema de astronomía en su Laboratorio de física y proyecto PROTEM.

Si quieren saber como, continúen.

El autor

RECONOCIMIENTOS

La amistad es la negación de esa soledad irremediable a la cual está condenado cualquier ser humano.

A Ángel Luís Plastino por acompañarme en esta aventura del conocimiento con el silencio crítico y cálido del amigo de tantos años.

A Rosa Orellana por el aporte de su ciencia astronómica y por soportar con la humildad del que sabe mi lamentable obligado desorden.

A Antonio Quijano que siempre me distinguió con su respeto y amistad.

Al grupo inicial, incorporaciones y a los actuales del proyecto I015 de UBACYT (FIUBA), y en el laberinto de la relación encontraron como resolver la utopía, Ciencia para el Departamento de Computación y Disimular debilidades, dándome soporte, marco y acompañamiento, a pesar de distanciamientos, contradicciones y oportunismos, tal vez humanamente justificables.

A Arturo Servetto que supo responder con amistad, respeto y apoyo en todas las desventuras por pretender lograr más de lo que se puede, laboratorios, carrera de Ingeniería Informática, transformar una Especialidad en Carrera de postgrado y proyectos de investigación y extensión (FIUBA).

A los más de veinte asistentes y miembros del Laboratorio de Sistemas Operativos y Bases de Datos y de proyectos acreditados en UBACYT, bajo mi dirección durante 15 años.

A los Docentes, Técnicos y Administrativos del Instituto LIDI y de la Facultad de Informática por ser el respaldo imprescindible.

Al Decano de la Facultad de Informática Prof. Ing. Armando De Giusti por su impulso, apoyo y comprensión sinónimos de amistad de tantos años.

MENCIONES

Por el apoyo y reconocimiento de la Facultad de Ciencias Astronómicas y Geofísicas y en particular de la Prof. Dra. R. B. Orellana, del Departamento de Mecánica Celeste (Astrometría) - Universidad Nacional de La Plata – Buenos Aires – Argentina.

Por parte del Prof. Dr M. H. Hamza y del Prof. Dr Vladimir L. Uskov de la International Association of Science and Technology for Development – IASTED – (Calgary – Alberta - Canadá) por reconocer mi trayectoria y mi participación en el Comité Técnico sobre Software, Educación y Bases de Datos, y la Mención Especial y Reconocimiento en Rhodes - Greece.

Al Prof. Dr. Jorge Muniz Barreto, Professor Titular de la Universidade Federal de Santa Catarina - Departamento de Informática e de Estadística Leboratório de Conexionismo e Ciências Cognitivas y Editor Responsable de la Revista Eletrônica de Sistemas de Informação – RESI – Brasil, y haber Mencionado Especialmente el trabajo y con la referencia del Observatorio Astronómico de Río de Janeiro: TAXONOMIC EVIDENCE APPLYING INTELLIGENT INFORMATION ALGORITHM AND THE PRINCIPLE OF MAXIMUM ENTROPY: THE CASE STUDY OF ASTEROIDS FAMILIES.

Al Prof. PhD KISS Imre, Secretario Responsable Editor del JOURNAL OF ENGINEERING y en cuyos Anales publicaron el trabajo: TAXONOMIC EVIDENCE OF CLASSIFICATION APPLYING INTELLIGENT DATA MINING. GALACTIC AND GLOBULAR CLUSTERS y tuvo la deferencia de ponerme como Referencia Científica. Faculty of Engineering Hunedoara - University "Politechnica" Timisoara – HUNEDOARA. Rumania.

A la ACM –ASSOCIATION FOR COMPUTING MACHINERY y del SIGAPP, por proponerme la Membresía en el ACM Special Interest Group on Applied Computing. Mención personal a Rosemary Rodríguez, por el ACM second anniversary (Babbage), reconocimiento.

Agradezco a la Dirección de Relaciones Internacionales (DRI) del CONICET (y UBA) por haberme Propuesto como Evaluador del proceso de evaluación científico-tecnológica del CYTED, en el Programa Iberoamericano de Ciencia y Tecnología para el Desarrollo, con mención a Sandra Mazoteras, Madrid – España.

PROLOGO

La Investigación Científica metódica le ha permitido al hombre incrementar su conocimiento en forma exponencial. La aplicación del método científico, basado en la observación, la experimentación y la verificación es la fórmula con la que se ha logrado progresivamente una comprensión cada vez más clara, cada vez más precisa y cada vez más amplia. Esto es porque el investigador, a través del método científico, va logrando un conocimiento verificable, al cumplir con las etapas de la investigación científica y lo crucial de la etapa de contrastación de hipótesis.

El conocimiento no es infalible, por lo tanto puede ser discutido, ratificado o rectificado, pero siguiendo las pautas metodológicas originales o especificando debidamente las razones de su modificación; discusión sobre el planteo del problema, formulación de la hipótesis, fijación de los objetivos, metodología de trabajo y preguntas que deberán ser respondidas en las conclusiones. Ciencia es la búsqueda racional del conocimiento, al manejar variables en forma diferenciada en la investigación de laboratorio y la investigación de campo.

“Es un largo viaje del intelecto que nació con el hombre y es materia viva evolutiva”.

Cuerpo de doctrina metódicamente adquirido: objetivo en los hechos, interpretativo en las leyes, deductivo en las hipótesis sobre bases epistemológicas, especulativo en las teorías soportadas por una base empírica. Esta base empírica del conocimiento es lo que permite la especulación objetiva, verificable y refutable, para el instrumentalismo y el realismo, al encadenar términos y enunciados mixtos para formar teorías [Bunge, M.1983].

Es difícil formar grupos y crear leyes abarcativas de problemas siendo más eficiente agruparlos en conjuntos disjuntos, de acuerdo al grado de profundización en el tema, en el modelo hipotético deductivo, resolviendo el dilema o par <conocimiento, práctica> [Perichinsky, G. 1995], mediante la

fundamentación, la predicción y la explicación de los fenómenos, operaciones esenciales de las que se ocupa la ciencia [Hempel, C.G., 1996].
Se puede comenzar por decir que la Ciencia es el Conjunto de conocimientos obtenidos mediante la observación y el razonamiento, sistemáticamente estructurados de los que se deducen principios y leyes generales y Teorías, cuando los objetos y entidades permiten encontrar regularidades [Klimovsky, G. 1994] [Perichinsky, G. 1995] [Gianella, A.E., 2000].

Para el desarrollo de la tesis, “un problema, el marco teórico y el papel asignado a las hipótesis y la realidad”, dan origen a las tareas de Investigación y la explicación, es la motivación principal del enunciado verdadero, utilizándose leyes y datos (explicar lo falso no tiene sentido).

En los múltiples lemas (trilema) se abarca la ERA científica del siglo XVII (con referencias a su evolución previa) y la ERA Pre-Moderna y Moderna. Además en la bibliografía relacionada aparecen la física, la astronomía, las ciencias naturales y las ciencias sociales como ejemplares de los tópicos tratados. Se ve que el avance de los paradigmas no es continuo ni espasmódico, su evolución avanza de acuerdo a la solución del dilema o par <conocimiento,práctica>, confluyendo los hitos históricos, en los primeros años y mediados del siglo XX (luego refinados) y, como metáfora planteada, de una teoría como sistema de ecuaciones, con términos empíricos y términos teóricos.

La Investigación está conformada por Procedimientos que contienen Acciones y Efectos que tienen por fin ampliar el conocimiento científico, tanto en el “laboratorio” como en el “campo”, al realizar actividades intelectuales y experimentales de modo sistemático con ese propósito sobre una determinada materia, en laboratorio, campo y experimentación ex post facto [Klimovsky, G. 1994] [Gianella, A.E., 2000].

La experimentación de laboratorio permite un mayor control de las variables, creando artificialmente las situaciones y casos a analizar.
El paradigma formal establece la relación entre elementos, como control de un

único elemento o variable que actúa sobre otra, y el resto de los elementos se mantienen constantes. El inconveniente es que resulta difícil crear condiciones artificiales sin alejarse de la realidad.

Para tratar varios componentes, y determinar propiedades y relaciones entre ellos se realiza la Experimentación Factorial.

Para lograr condiciones cercanas a la realidad se realizan Experimentos de campo, tomando componentes reales y estudios en escala. Tomar más componentes es acercarse a los fenómenos, le da relevancia respecto de otros tipos de experimentos. Por ello la investigación de campo es una metodología usada en ciencias sociales. Los investigadores ingresan en un grupo u organización o institución, toman contacto directo con los procesos e interacciones sociales de esos grupos.

Los componentes son más controlados en la investigación de campo, pues se utilizan mediciones y escalas para el registro de las conductas; se obtiene una información completa de los fenómenos.

Experimentos ex post facto son aquellos en los cuales no se manipulan las variables, debido a impedimentos a veces de índole ética y otras veces de tipo técnico. Los fenómenos involucrados pueden ser sociales, económicos, históricos y astronómicos, con variables cualitativas y cuantitativas. Las primeras responden a criterios clasificatorios, las segundas permiten alguna correspondencia de orden numérico, y pueden a su vez dividirse en variables ordinales y métricas, sujetas a escalas de medición.

En el método experimental se requiere gran cantidad de casos para evaluar un número reducido de propiedades, para asegurar la validez, la objetividad y la confiabilidad experimental mediante un diseño técnicamente adecuado; pueden usarse tanto en métodos exploratorios como para contrastar hipótesis.

Una analogía en un investigador o una comunidad científica, conforma una “Base Empírica”, en la cual un Objeto, Entidad o Situación es un “Dato”, o captado es una “Observación” [Klimovsky, G. 1994]. Objetos “Directos” o empíricos en la Zona Empírica, e “Indirectos” o Teóricos en la Zona Teórica.

La dinámica del conocimiento surge del análisis del modo de trabajo en las ciencias, en la búsqueda del conocimiento, la clasificación es la dinámica del

conocimiento, siendo para el observador una acción, o procedimiento o conjunto de procedimientos, donde se realiza la acción.
Dicho análisis es descriptivo y normativo, actividad de reglas metodológicas, proceso de investigación de los científicos individual o colectivamente, interpretación que verifica las hipótesis, es un modelo tanto formal como fáctico, o teoría de parte de la realidad de un sistema determinado.

Comienza por un hecho desconcertante o estado de ánimo de desorientación y perplejidad, falta de medida en las acciones o palabras del aspecto, que presenta un problema para el observador, cuestión a aclarar, pues son un conjunto de hechos, circunstancias o proposiciones, que dificultan la consecución de algún fin. Propuesto por Aristóteles, la Episteme de los griegos, fundamentó el conocimiento mediante reglas metodológicas, hasta que los empiristas ingleses de los siglos XVII y XVIII, cambian los conceptos del conocimiento como fruto de la experiencia, aportando el método experimental inductivo.

Los momentos del Novum Organum (y sus ídolos) [Bacon, F. 1605-1620] [Klimovsky, G. 1994] [Gianella, A.E., 2000], indican: (1) el relevamiento u observación directa o experimental de los datos, (2) planteo del método inductivo de generalización de los patrones de comportamiento, de los datos relevados, que al conectarse sistemáticamente conforman conjuntos parte o niveles, leyes de máximo nivel que se pueden denominar teorías. El (3), deducir o predecir hechos o fenómenos al aplicar las leyes y teorías, en una precedencia dinámica.

Para el neopositivismo o positivismo lógico del siglo XX [Popper, K. 1961-1963-1974-1985], el método inductivo es un reduccionismo de la investigación, es una precedencia errónea, dado que los datos se relevan en función del problema a resolver y su generalización, generando teorías sin partir de hipótesis de trabajo o una heurística de nivel superior.

Se llega así, al modelo Hipotético-Deductivo de investigación, cuyas etapas del sistema dinámico y holístico conforman un grafo de precedencia que se

podrían denominar fases de la investigación [Perichinsky, G. 1995] [Bunge, M. 1969] [Samaja, J. 1993], por ser dinámico y no etapas mecánicas, con Momentos, Subetapas, Tareas o Acciones, constando hasta de doce fases o etapas en distintos autores.

Se consideran etapas básicas [Gianella, A.E. 2000]: Problema, Hipótesis, Marco Teórico, Procedimientos deductivos, Consecuencias contrastables, Procedimientos de contrastación, Evaluación de los resultados y según el resultado se verifica la hipótesis o se genera una nueva hipótesis refutando la anterior. Lo que da origen a tareas de Investigación tales como el problema, el marco teórico y el papel asignado a las hipótesis y la realidad y no, un mero relevamiento de datos.

El conocimiento es un producto de valor estratégico, del cual a las organizaciones generadoras del mismo suele interesar su producción, acopio y transmisión pero no necesariamente su aplicación a la resolución de problemas de la sociedad. El ambiente de desarrollo de los sistemas más modernos se basa en tecnologías y modelos teóricos desarrollados hace más de dos décadas. En este marco, la comercialización del conocimiento imprime carácter al tipo de productos informáticos requeridos por el mercado.
Siguiendo a [Bunge, M.1999] y otros; “el enfoque de mercado de la ciencia básica está condenado al fracaso porque, a diferencia de la tecnología sin método científico, que es un herramental técnico, la ciencia y la tecnología creativa no están en venta. Ésa es la razón por la que su financiamiento debe ser público y no privado. Mantengamos así la investigación básica si deseamos que siga creciendo y enriqueciendo la cultura y alimentando la tecnología. Expulsemos a los mercaderes del templo de la investigación desinteresada y salvémosla de parecerse a la caricatura economista que probablemente alguien tenga en mente, que además no es ciencia desde el punto de vista cultural [Perichinsky, G. 2005]”.
Los artefactos científico-tecnológicos no son mercancías del mercado autodefinido por el neoliberalismo y la gestión como actividad comercial.

Se está en presencia de una revolución tecnológico-cultural cuyas bases residen en la transformación controlada de información en conocimiento a gran escala, la que a su vez conduce a la creación automática de nuevo conocimiento. Es muy probable que estos cambios produzcan una nueva división del trabajo en la cual los países centrales serán aquellos que hayan dominado la tecnología de generar conocimiento en forma automática, en tanto que los nuevos dependientes quedarán relegados a la producción y transformación de materia y energía. Esta revolución generará cambios mucho más profundos y trascendentes que los de la revolución industrial.

Esta Tesis Doctoral presenta en su concepción una visión integradora y original de la aplicación de distintas teorías científicas de diversa extracción epistemológica a la resolución del problema de extraer nuevo conocimiento a partir de un gran volumen de datos. Desde esta perspectiva, la originalidad de su aporte es esencialmente informática. Por otra parte, al presentar la construcción del instrumento que permite resolver el problema de referencia, es esencialmente científico tecnológico. Por analogía con la taxonomía se realizaron aportes sobre un desarrollo propio de bases de datos relacionales dinámicas, como motor y reservorio de datos clasificados, necesarios para utilizar en tecnologías informáticas emergentes como “data mining”, “datawarehousing” y “data mart”. Estas pueden ser definidas como el proceso iterativo de selección, exploración, y modelado de grandes cantidades de datos para mostrar relaciones desconocidas previamente y contrastación de Teorías Clasificatorias mostrando su “Robustez”. En este proceso se requiere análisis, manipulación de datos y herramientas de visualización que permitan facilitar el descubrimiento de modelos que ayudan a los usuarios a explorar hipótesis sobre sus datos.

Por lo expresado, debe ser una razón de orgullo institucional que la Facultad de Informática de la Universidad Nacional de La Plata haya sido el entorno propicio para la primer Tesis Doctoral que explica la Taxonomía Computacional en Ciencias Informáticas con un nuevo Criterio Clasificatorio.

Por último, como mencionara oportunamente el Dr. Ramón García Martínez, para este prólogo debemos hacer una clara referencia a su autor. Su dilatada trayectoria en el campo de la Informática en puestos de relevancia en la Universidad Nacional de La Plata, la Universidad de Buenos Aires, en la Administración Pública y en temas impactantes en la Extensión, le ha ganado el reconocimiento de sus pares como uno de los mentores de cambios importantes de la cultura informática de las antecitadas organizaciones y del país. Como si este aporte no fuera suficiente, esta tesis corona una vida dedicada a la investigación, la enseñanza y la promoción de la Informática en todas sus áreas, que le valieron una Categoría I de la Comisión Regional de Categorización Metropolitana, máxima en investigación científica y tecnológica.

La Plata, 10 de Diciembre de 2007.

Laboratorio y Proyectos Acreditados en UBACYT
DEPARTAMENTO DE COMPUTACIÓN
FACULTAD DE INGENIERÍA
UNIVERSIDAD DE BUENOS AIRES

EXORDIO

Esperemos lo que deseamos, pero soportemos lo que acontece.

Marco Tulio Cicerón

I

1. EXORDIO

Esta tesis aborda la definición de un método numérico basado en invariantes para la clasificación automática de objetos a partir de la información de sus caracteres, focalizado en la búsqueda de las invariantes con base en una aplicación original metodológica de los principios de superposición e interferencia en el análisis de espectros, en congruencia analógica con la taxonomía numérica, por su relación lógica y con fortaleza metodológica.

Se demuestra un nuevo criterio para dar validez al método en casos no resueltos hasta ahora por la ciencia.

Este exordio como principio o preámbulo de la tesis, tiene especialmente por objeto excitar la atención; y para su desarrollo, “el problema, el marco teórico y el papel asignado a las hipótesis y la realidad”, dar origen a las tareas de Investigación y la explicación científica.

Siendo la motivación principal del enunciado verdadero, utilizando leyes y datos, es necesario conceptualizar la problemática epistemológica (primera parte 1.1.) y un programa de investigación científica (PIC) como sucesión de teorías emparentadas semántica y sintáticamente, que se van generando en distintas disciplinas por observaciones intrigantes, que se captan históricamente y llaman la atención pues se comportan en forma desconcertante o funcionan de una manera diferente a la esperada, constituyendo familias de fenómenos intrigantes (segunda parte 1.2.).

1.1. De la Epistemología.

1.1.1. Etapas de la investigación científica.

Se consideran las siguientes etapas básicas de la investigación científica [Gianella, A.E., 2000]: Problema, Hipótesis, Marco Teórico, Procedimientos deductivos, Consecuencias contrastables, Procedimientos de contrastación, Evaluación de los resultados y según el resultado se verifica la hipótesis o se genera una nueva hipótesis refutando la anterior. Lo que da origen a las tareas

de Investigación son el problema, el marco teórico y el papel asignado a las hipótesis y la realidad y no, el mero relevamiento de datos.

De los problemas se formulan preguntas y hay que intentar responderlas o explicarlas, trascendiendo el contexto del conocimiento del estado de una disciplina, respecto a la realidad relativa.

La implicación y la generalización de los problemas y preguntas se pueden ordenar en grados: gradación [Perichinsky, G. Investigation,1995].

En las investigaciones el marco teórico, con sus componentes de una o más teorías, homogéneas o heterogéneas, está presente a través de sus hipótesis, condicionando a los interrogantes o preguntas que se formulan, e interesarse por algo o por avanzar en determinada dirección.

Mediante una hipótesis o conjetura una vez formulado con claridad el problema, que se va a investigar, y se procederá a buscar su solución.

El conjunto de ideas conectadas requieren Teorías e Hipótesis de trabajo formado por relaciones de compatibilidad e implicación, que pretenden comprender y explicar un determinado dominio de la realidad. Las hipótesis de trabajo [Klimovsky, G. 1994] se estratifican en tres niveles: (1) Descripción de individuos u objetos (artefactos) de bajo nivel que describen, analizan, registran, enumeran y atribuyen propiedades. Los objetos toman relaciones de la Base Empírica formada por conjuntos de entidades, fenomenologías, propiedades y relaciones programadas; (2) Nivel intermedio de observaciones que generalizan, correlacionan, subsumen y clasifican; es un nivel preteórico y, (3) Hipótesis de máximo nivel y observaciones, que explican, predicen, comprenden, sistematizan, inventan soluciones y metodologías.

Esta tarea resulta tener muchas veces valor heurístico, es decir, contribuye a estimular la creatividad del científico.

Los mecanismos de producción de ideas y de resolución de problemas, "lógica del descubrimiento", recurren a procedimientos de la lógica formal, el cálculo de predicados o la teoría de conjuntos y se suelen utilizar sistemas lógico-matemáticos, premisas encadenadas, en tanto marco teórico.

1.1.2. Consecuencias observacionales.

Surgen consecuencias contrastables como enunciados inferidos deductivamente de las hipótesis, susceptibles de confrontación con la experiencia. El lenguaje es observacional, no teórico.

Si fueran enunciados acerca de propiedades, hechos o relaciones ya conocidos, las hipótesis los explican.

Los procedimientos de contrastación de las consecuencias observacionales, son una etapa crucial de la investigación científica. Las formas que puede adquirir esta etapa tienen modalidades muy diferentes, propia de las distintas técnicas de investigación, la observación sistemática, la experimentación, la administración de testes (pruebas), y la realización de encuestas y grabación de entrevistas y la recolección de datos estadísticamente procesados.

Se aplican procedimientos experimentales en los que se analizan las modificaciones en los valores de la variable independiente en relación con la dependiente, empleadas en ciencias experimentales. Los momentos de aplicación son, (1) el diseño del experimento, (2) la realización del experimento y ambiente de laboratorio o de campo, y (3) el registro y evaluación de los resultados obtenidos.

En lugar de la experimentación se realizan observaciones sistemáticas o una experimentación ex post facto.

El enunciado de una ley es una hipótesis general empíricamente confirmada, inmersa en una teoría, sistema hipotético-deductivo representante de una regularidad objetiva [Bunge, M., 1969, 1983, 1999].

Los enunciados están cargados teóricamente y solo son aceptados como un acuerdo de la comunidad científica [Imre Lakatos, 1999].

Todas las proposiciones, son premisas (datos) o consecuencias (teoremas) al tomarse en conjunto, de una teoría bien organizada en tanto es axiomática, están sistemáticamente unidas por la relación de deductibilidad (sintaxis) y algún tópico común (semántica). Un modelo es una teoría de mediano alcance pues no tiene clase de referencia ni enunciados que sean leyes [Bunge, M. 1999]. Las teorías que contienen leyes probabilísticas y las que requieren cláusulas “ceteris paribus” no son refutables, estas últimas no requieren más factores pues las hipótesis auxiliares se van a cumplir, complementariamente, en el fenómeno en estudio, de todas maneras siempre se puede sustentar la teoría (factual) [Poincaré, H. 1908] [Quine, William van Orman, 1975-1981].

En la evolución de la constratación, respecto al modelo epistemológico hipotético-deductivo, surgen tres modelos planteados como más racionales [Popper, K 1963-1974] [Imre Lakatos 1999] denominados falsacionismo o

refutacionismo [Klimovsky, G. 1994]: Falsacionismo dogmático, Falsacionismo ingenuo es el hipotético-deductivo simple y Falsacionismo sofisticado es el hipotético-deductivo complejo, modelos propuestos para la investigación científica.

El falsacionismo y el neopositivismo [Popper, K 1963] [Imre Lakatos 1999] al considerar el carácter racional de la investigación científica, rechazan el justificacionismo del conocimiento, donde las afirmaciones siempre son demostradas empíricamente, condición de contrastación o verificación.

En el Falsacionismo Dogmático [Imre Lakatos 1999] se formulan hipótesis nuevas y se contrastan rigurosamente, para un esquema de Programación de Investigación Científica (PIC). Es metacientífico, pues se usa su modelo para probar la tesis del modelo hipotético deductivo.

Contiene una demarcación neta, entre enunciados teóricos por una parte y enunciados observacionales o básicos, por otra, son demostrables por la experiencia, es decir su verdad o falsedad. Es decir una teoría es científica si tiene una base empírica, entendida como el conjunto de los falsadores potenciales de la teoría, verificables por la experiencia.

Para estructurarse científicamente el falsacionismo debe evolucionar metodológicamente, planteando el modelo del Falsacionismo Ingenuo, uno conservador, de la comunidad científica [Poincaré, H. 1908] y otro revolucionario, con cambios en las teorías refutadas [Deum, P., Ryle, G., Einstein, A.].

1.1.3. Desarrollo de la ciencia.

A pesar de la evolución de que representa el falsacionismo metodológico ingenuo sobre el falsacionismo dogmático, ambos comparten algunas tesis que impiden explicar la historia real de la ciencia. Los únicos verdaderos descubrimientos son las refutaciones de hipótesis científicas.

La historia de la ciencia ha mostrado que las cosas no sucedieron de acuerdo con estos criterios tan simples; por lo cual hay que reemplazarlos por versiones refinadas de los mismos principios:

i. “Las contrastaciones son, al menos, un triple enfrentamiento entre teorías rivales y experimentación.”

ii. “Algunos de los experimentos más interesantes resultan de la confirmación más que de la falsación.”

El desarrollo de la ciencia, surge de la competencia de una secuencia de teorías que comparten un núcleo duro (hard core), formado por hipótesis. Un programa de investigación científica (PIC) es una sucesión de teorías emparentadas T_1, T_2, T_3,…,T_n, que se van generando unas a partir de las otras. Tienen en común un conjunto de hipótesis fundamentales que forman su núcleo duro, al cual se lo declara "irrefutable" por decisión de la comunidad científica. El núcleo duro de todo programa de investigación se halla resguardado por un cuerpo de hipótesis auxiliares que forman un “cinturón protector” alrededor del núcleo, a fin de lograr un ajuste entre teorías y resultados experimentales.

Para producir un procedimiento habitual en la investigación científica, el falsacionismo ingenuo evoluciona hacia el falsacionismo sofisticado.

Una evolución se produce al reemplazar una teoría por otra en el mismo programa de investigación. Otra evolución es un criterio que establece que una teoría T está falsada si y sólo si se ha propuesto otra teoría T', con las siguientes características:

i. T' tiene más contenido empírico que T, es decir predice nuevos hechos, hechos que son improbables a la luz de T o incluso prohibidos por T.

ii. T' explica los aciertos previos de T, es decir todo el contenido no refutado de T está incluido en T'.

iii. Alguna parte del contenido excedente de T', respecto de T, está corroborado.

Hay continuidad en el conocimiento y se presupone la comunicación entre científicos que trabajan en programas de investigación diferentes. El falsacionismo comparte la creencia de que no puede haber una distinción entre términos teóricos y observacionales, no está de acuerdo en que ello implique la inconmensurabilidad de las teorías científicas rivales. Oscilando con modelos para los cuales la historia y la existencia de una base empírica son fundamentales si implican la inconmensurabilidad [Kuhn, T.S. 1980] [Feyerabend, P. 1981].

En el modelo del Falsacionismo sofisticado se utiliza el esquema estratificado de tres niveles [Klimovsky, G. 1994] visto antes, las hipótesis auxiliares y ad-hoc (post-hoc), espera que se aprueben las hipótesis auxiliares con el tiempo, pues juega un papel fundamental en la

consideración de los llamados «experimentos cruciales». Crucial es un experimento que refuta una teoría. El programa (PIC) puede recuperarse mediante un desarrollo creativo de su heurística positiva. Visto retrospectivamente podrá afirmarse que el experimento en cuestión había sido crucial, y se lo pretendió refutar.

Una heurística es positiva si se generan hipótesis que protegen a la teoría. Una heurística es negativa cuando se genera un cinturón de hipótesis auxiliares que protegen al núcleo duro. Se toman hipótesis auxiliares de la base empírica, que reemplazan a las hipótesis de la teoría central del Programa de Investigación Científica [Klimovsky, G. 1994].

1.1.4. Modelo hipotético-deductivo.

En el modelo hipotético-deductivo, dinámico y holístico [Bunge, M 1999], se rechaza todo ese desarrollo de modelos, con lo cual coincido [el autor], pues los enunciados básicos no pueden verificarse por observación o experimentación, rechaza la lógica inductiva, ya que, cualquier ley universal tendrá probabilidad cero de demostrarse, pues por inducción infinita la investigación científica no logra su verdad. Además no demarca la Ciencia de la No Ciencia, entre afirmaciones teóricas y observacionales.

Se distingue en la Historia Interna de una disciplina o teoría científica a la que incluye a las variables que pueden cambiar a la teoría, si las cuestiones metodológicas lo indican, de la Historia Externa de los elementos EBCP de Bunge, Económicos, Biológicos, Culturales y Políticos (Vertehen und geisteswissenschaften ≡ Comprensión Intuitiva, Humanística y Social).

Surge así un argumento que relativiza la refutación, que es la falibilidad de los datos observacionales, o falta de certeza que puede existir respecto de la verdad o falsedad de los enunciados empíricos.

Si los resultados obtenidos en la contrastación empírica son favorables la hipótesis está corroborada y adecuada al problema y se produce la aceptación de la hipótesis, pero esa aceptación tendrá siempre un carácter provisorio.

Si los resultados obtenidos fueran refutatorios, hay que determinar cuál o cuáles premisas son las responsables de ese resultado adverso, y se produce el rechazo de la hipótesis. Es la hipótesis en cuestión, o alguna de las auxiliares que intervienen que deberán evaluarse en forma independiente. Si ellas no fueran refutadas, es falsa la hipótesis inicial, y habrá que abandonarla

y proponer en su lugar una nueva, o se podrá intentar corregir modificándola en algún aspecto, en su alcance o en alguno de sus términos.

1.1.5. Enunciados, Instrumentalismo, realismo y conductismo.

Para construir los enunciados es necesario construir un lenguaje científico [Althusser, L 1965], que transforme todos sus términos, incluso los lógicos, en términos específicos o técnicos, porque los términos del lenguaje cotidiano son inadecuados y; construir oraciones que puedan ser útiles para expresar conocimientos como lo señalado por el "instrumentalismo", en el problema de los términos teóricos, formando expresiones complejas que permitan describir un estado de cosas, observable o no. Los términos además de empíricos y teóricos pueden ser presupuestos, específicos, lógicos, designativos, ordinarios y científicos.

El instrumentalismo considera que muchos términos teóricos no son designativos, a pesar de ser específicos, por lo cual habría que decir que no son ni empíricos ni lógicos.

El conductismo usa la palabra "constructo" (del inglés “construct”), para insinuar que un término teórico es en realidad una construcción basada en aspectos puramente empíricos. Así, en epistemología, es abstracta una teoría constituida únicamente por enunciados teóricos puros, no pudiéndose deducir nada de los mismos, aplicables a la experiencia o a la práctica, ni realizar explicaciones ni predicciones sobre lo que acontece en la base empírica. Son enunciados teóricos "mixtos", con términos teóricos y empíricos, o "enunciados puente", vinculando el ámbito puramente teórico del discurso, a lo observable o práctico, localizado en la base empírica, son las "reglas de correspondencia".

El instrumentalismo prefiere artificios de carácter lingüístico para vincular observaciones entre sí.

El realismo considera que los términos teóricos se refieren a entidades, aunque no sean observables, probar la verdad o la falsedad de los enunciados teóricos, sin acudir a observaciones o a métodos estadísticos.

La estructura de los enunciados es de tres niveles, (1) Enunciados empíricos básicos (singulares); (2) Enunciados empíricos generales o generalizaciones empíricas. Derivando a los Universales, Existenciales, Mixtos y Estadísticos o probabilísticas; y (3) Enunciados teóricos (generales Puros y Mixtos).

Un científico, para formular hipótesis o conjeturas, usa el método sorprendente

y hasta decepcionante que usa un artista cuando se le ocurre una obra de arte, el poder de imaginación y de creación de que dispone. Imaginar qué puede haber "detrás" de una apariencia fenomenológica, explicar el comportamiento de ésta apariencia o de un fenómeno, inventar hipótesis y después controlarlas. Los conceptos teóricos se definen operacionalmente, característica lingüística y lógicamente es el "contexto abierto del lenguaje".

La actividad de la epistemología como investigación de la ciencia, es una situación dialéctica con los científicos; de aprendizaje mutuo. Para Albert Einstein, los procedimientos operacionales, no ocultan el significado de los términos teóricos, ligados a la noción de teoría no a la definición operacional, pero se aplican las técnicas operacionales para introducir conceptos.

En el instrumentalismo y en el realismo, es legítimo usar términos teóricos cuando los requieran.

Los positivistas, introducen términos teóricos si aumenta la predictibilidad de las teorías, mientras que una divergencia tiene carácter semántico y es abismal. Por ello el instrumentalismo afirma que los términos teóricos son instrumentos verbales sin referencia y sin significación, complementarios de lo empírico para construir deducciones lógicas, tiene enunciados con términos teóricos que no son genuinamente hipótesis, porque no se los supone verdaderos. Son seudo-hipótesis o principios de un sistema axiomático, que en el tercer nivel son enunciados puros o de reglas de correspondencia, cuyos términos teóricos serían simples rótulos sin significado y de carácter deductivo. Si en la teoría se dispone de datos y de seudo-hipótesis, es posible realizar deducciones que conduzcan a consecuencias observacionales, por lo cual los términos teóricos son instrumentos mediatizadores, como enzimas y catalizadores, que permiten construir una "reacción deductiva". Permite obtener observaciones previstas a partir de observaciones ya obtenidas, lo cual acrecienta el conocimiento de la base empírica, por ello se denomina "instrumentalismo" a esta posición.

El realismo admite que los términos teóricos tienen un sentido puramente instrumental, pero hay casos en que no es así, aludiendo a entidades no observables, considerando real al referente del término teórico.

El realismo es muy atractivo pues permite obtener un conocimiento que trasciende el de la base empírica. La ambición del realista es conocer cómo es

el mundo en sus fundamentos ontológicos o lo que existe más allá de lo accesible a nuestros sentidos e instrumentos.

El aspecto informativo de una teoría, por su carácter hipotético, denota que las entidades existen, con hipótesis definitorias, que tienen la misión de definir, mientras que las restantes serían meras hipótesis.

Una metáfora de una teoría puede ser imaginada como un sistema de ecuaciones en que los análogos son los términos empíricos u observables, aquéllos cuyo significado es conocido. Los términos teóricos, serían los desconocidos o incógnitas, de los cuales se tienen que satisfacer ciertas condiciones.

El científico discute los hechos tal como se le presentan a través de un paradigma que los constituye y articula, pero éste niega la realidad [Kuhn, T.S. 1980] [Feyerabend, P. 1981]. Cuando se produce una revolución científica la "realidad" deja de ser lo que era para transformarse en algo nuevo, pues los hechos, articulados por el antiguo paradigma, desaparecen en tanto tales y son reemplazados por los "nuevos hechos" que ahora ofrece el nuevo paradigma.

Las tesis “fuertes” [Kuhn, T.S. 1980] sostienen que el concepto de "verdad", entendido en un sentido absoluto y aristotélico, es totalmente inútil en ciencia, lo cual es muy grave [el autor].

La posición filosófica de la teoría del conocimiento y lo que ofrece, más que una ontología, es una tesis gnoseológica (coherentismo). Cada comunidad científica, al constituir su paradigma, decidirá implícitamente cuál es su ontología. En el realismo que considera una realidad en la que existe el "objeto en sí" o "noúmeno" no es accesible al conocimiento directo. Así, para la ciencia y la constitución del conocimiento, esta realidad está vedada y no cumple ningún papel interno a la ciencia misma [Kant, I. 1973] [Kuhn, T.S. 1980].

1.1.6. Evolución científica.

La evolución científica, por acomodación y equilibrio, “no se puede concebir, como un acercamiento por aproximaciones sucesivas a la realidad.”

“La historia de la ciencia y, en particular, la de la tecnología, es una larga y clarísima descripción de cómo los medios técnicos y los procedimientos de la ciencia para mejorarla muestran un progreso, aumento de eficacia y operatividad, pese a que a lo largo del tiempo los paradigmas se sustituyen unos a otros.” [Klimovsky, G. 1994]

Es un proceso que se denomina acomodación y, a diferencia de la asimilación (semejante al de ciencia normal), es característico de las etapas de cambio en los procesos evolutivos, que finaliza cuando se alcance un nuevo estado de equilibrio, en el cual el organismo recobra las facultades de asimilación.

A medida que la ciencia se desarrolla y se formulan teorías que se suceden unas a otras por los procesos de asimilación, acomodación y equilibrio, los objetos de los que habla cada teoría se asemejan cada vez más y se aproximan a lo que configuraría el "objeto real", nunca alcanzable.

El problema de la explicación [Gianella, A.E. 2000], es la motivación principal para la formulación de teorías científicas, capaces de explicar sucesos que intrigan a los científicos, quienes quieren comprender. El centro de gravedad de la epistemología y la metodología es la operación de contrastación y de predicción. El inductivismo no explica ni predice, propone un tipo de inferencia que permite obtener generalizaciones con datos y muestras.

Son tres operaciones esenciales de las que se ocupa la ciencia: fundamentación, predicción y explicación.

Fundamentar un enunciado es indicar las razones por las cuales se le puede considerar verificado, y por el método hipotético deductivo, está "suficientemente corroborado". La predicción, se refiere a consecuencias observacionales, no se sabe si el enunciado es verdadero, pero la predicción ofrece elementos que tratan de anticipar si en el futuro, ocurrirá de la manera descrita. La predicción es más débil que la fundamentación, no prueba la verdad y ni siquiera equivale a una corroboración, hay que verificar y establecer que lo predicho se ha cumplido y sea admitido como conocimiento, la observación desempeña un papel esencial.

La explicación es una deducción cuya conclusión describe el hecho intrigante con premisas de datos y leyes.

Son dos tipos de premisas, los datos o condiciones iniciales de la situación y leyes, premisas de las cuales se deduce el enunciado que describe el hecho intrigante que se quiere explicar.

Para explicar un hecho son necesarios datos y leyes. Las leyes por sí solas no permiten deducir aspectos fácticos singulares. A la inversa, aunque se conozcan datos, sin leyes no será posible realizar la deducción.

La estructura lógica del esquema de explicación como en el método hipotético

deductivo, lo lleva a proponer lo que se denomina el "principio de simetría entre explicación y predicción". Si se hace una predicción y esta se cumple, entonces, automáticamente, se transforma en explicación.

La Predicción es un acontecimiento que habrá de ocurrir, en sentido epistemológico es la conexión deductiva entre conocimientos que ya se poseen. En la práctica científica hay que realizar predicciones por medio de teorías y leyes y renunciar a las profecías. También es necesario discriminar entre auténticas explicaciones y seudo explicaciones. Estas últimas argumentan para dar una explicación, porque hay ausencia de datos o se entró en un círculo vicioso.

El modelo estadístico de explicación, es un caso particular, donde las leyes son enunciados estadísticos o probabilísticos que establecen una regularidad en sus términos no universales.

La Explicación genética, no usa leyes, sino hechos pertinentes encadenados por precedencia.

Las Explicaciones teleológicas son modelos que explican un hecho presente, con algo que ocurrirá en el futuro (telos, significa "fin" u "objetivo").

En el funcionalismo [Parsons, T. 1966] [Manilowski, B. 1986], se adscribe a una sociedad con comportamiento homeostático, sistema funcional, donde la alteración de variables o factores que caracterizan su funcionamiento, produciría un proceso que le permitiría recobrar su estructura. Explicar por causas y por razones [Deum, P., Ryle, G., Einstein, A.], que es cuando se deduce usando premisas-leyes y leyes causales.

1.1.7. Reduccionismo.

Cuando se encara con vaguedad un problema importante, se reduce, es una postura filosófica denominada reduccionismo. Es cuando se tratan objetos o ámbitos de cierta naturaleza, que pueden definirse o caracterizarse en términos o en componentes, que corresponden a otro ámbito de naturaleza distinta.

Se advierte la conexión entre reducción y explicación, si existe un procedimiento para reducir una disciplina a otra y, una teoría a otra de una disciplina anterior, donde las leyes de la disciplina que ha sido reducida, se transforman en hipótesis derivadas de las teorías de mayor alcance. Las leyes fundamentales de una disciplina serán explicadas por las leyes o las teorías de

la disciplina básica a la cual se reduce la primera. Los atractivos epistemológicos y metodológicos de la reducción, es que una disciplina quedará además de reducida, explicada, sobre la base de las teorías exitosas de la disciplina fundamental.

La tesis del “reduccionismo ontológico”, es que al conectar las leyes de una disciplina con las de otra, los objetos que trata una disciplina serán reducidos, y su apariencia es de entidad simple, no de una estructura compleja o sistema, cuyas propiedades deben comprenderse en términos de las entidades de otra disciplina.

El "reduccionismo semántico", no reduce entidades a otras entidades, sino que el lenguaje de una disciplina, que se quiere reducir, sea traducido al lenguaje de otra disciplina básica.

Es un problema semántico y sintáctico, donde, el reduccionismo de carácter semántico, sugiere la posibilidad de dejar de hablar con un vocabulario y terminología, para decir lo mismo de la disciplina anterior.

Una clase interesante de reducción del problema de explicación, por su conexión es el “Reduccionismo metodológico”, al reducir una teoría básica a otra reducida, que implica una reducción semántica del lenguaje de una teoría básica al lenguaje de otra reducida, con el resultado que, al hacerlo, se descubra que una teoría es derivada de otra. Primero, la reducción semántica por utilizar ambas teorías vocabularios diferentes, segundo la dependencia deductiva de una con relación a la otra y tercero, porque la "máquina de deducir", no nos permitiría acceder a las locuciones de la teoría reducida a partir de las de la teoría básica.

Existe una correlación entre lo que sucede con las entidades de ambas teorías, la regla de correspondencia. Se conectan dos vocabularios, no de manera semántica, sino formulando hipótesis acerca de cómo se correlacionan situaciones descritas, por expresiones en el vocabulario de la teoría reducida con otras que emplean la teoría básica. Por lo tanto la teoría básica, la teoría reducida y el conjunto de reglas de correspondencia vinculan expresiones del vocabulario [Nagel, E. 1968].

Por consiguiente, aunque la noción de explicación científica es más profunda y general que la de reducción, reducir y explicar se vinculan en el sentido que, una teoría queda explicada por aquella a la que metodológicamente se reduce.

1.2. De la Tesis.

Para abordar un método numérico basado en invariantes para clasificar objetos en forma automática, a partir de la información de sus caracteres, focalizado en la búsqueda de invariantes, el análisis espectral, la taxonomía computacional y la teoría de la información, del desarrollo de la ciencia y la generación de conocimiento, surge la competencia de una secuencia de teorías, de un programa de investigación científica (PIC) (ver 1.1.3.), que tiene en común un conjunto de hipótesis fundamentales, a fin de lograr un ajuste entre teorías y resultados experimentales.

Se demuestra la validez del método en casos no resueltos hasta ahora por la ciencia.

En las investigaciones, el marco teórico con sus componentes de una o más teorías, homogéneas o heterogéneas, está presente a través de sus hipótesis, condicionando a los interrogantes o preguntas que se formulan, e interesarse por algo o por avanzar en determinada dirección.
Mediante una hipótesis o conjetura, una vez formulado con claridad el problema que se va a investigar, se procedió a buscar su solución.
De los problemas surgieron las preguntas que se formularon y que han podido ser respondidas o explicadas, trascendiendo el contexto del conocimiento del estado de una disciplina, respecto a la realidad relativa.

La implicación y la generalización de los problemas y preguntas se pueden ordenar en grados: gradación [Perichinsky, G. Investigation,1995].

La gradación básica para la aplicación, en casos de uso, de una metodología original, de los principios de superposición e interferencia, para la generación y análisis de espectros, en congruencia analógica con la taxonomía computacional numérica y el teorema de Tchebycheff, los paradigmas de las bases de datos y herramientas emergentes de la inteligencia artificial, para

verificar la fortaleza del método y además, la teoría de la información y la máxima entropía (PME).

De acuerdo a todo lo expresado surge un conjunto de explicaciones de los problemas a resolver en las distintas eras científicas.

1.2.1. **La pre ERA científica (Aristóteles)**

Se puede decir que comienza unos 2.000 años antes que Newton, en su Liceo y con su discípulo Teofrasto, que lo sucedió en la dirección del Liceo (peripatos), fue el fundador de la botánica. En el Liceo había una biblioteca, un zoológico y un jardín botánico. Tenía colecciones de mapas y de minerales, y varias aulas y talleres donde se estudiaba e investigaba.

Se realizaban simposios, y en ellos surgió el nombre de la Física, la Meteorología, la Economía, la Poesía, la Etica y la Política. También se enseñaba lógica, biología, medicina, astronomía, historia y sociología. Surgiendo así la competencia de una secuencia de teorías, de un programa de investigación científica (PIC) (ver 1.1.3.), y su gradación frente a problemas.

La clasificación (taxonomía), milenios antes que Carlos de Linneo, de los vegetales, árboles, arbustos, matas y hierbas es mucho más racional que otras que se usaron antes del siglo XVII.

La organización de género, orden, división, reino, especie, familias y clases, con el principal propósito de dar esta jerarquía y plantear relaciones de evolución entre individuos.

El geocentrismo y el sistema de las esferas planetarias ya impuesto por la Academia de Platón, fueron planteadas por Aristóteles, quien estimó erróneamente el diámetro de la Tierra, pero en el Liceo se planteó el primer sistema heliocéntrico y se adelantó a los modernos.

Lo mismo ocurrió con los estudios y explicación de los ciclos y epiciclos, en el Liceo de Aristóteles aplicaron y se explicaron en astronomía y biología. Por otra parte los peripatéticos crearon el museo de Alejandría.

El más nuevo de los problemas de la Física es al mismo tiempo el más antiguo, pues no existe experiencia alguna más primitiva, tanto en un niño como en toda la raza humana, que la sensación de recibir luz y calor del sol. Hay que fundamentar y explicar como llegan a través de los espacios vacíos, interestelares. Los griegos dieron una respuesta muy simple y satisfactoria, explicaron sin indagar y validar experimentalmente. La explicación de los griegos era que el sol y todos los cuerpos que irradian luz y calor debían arrojar corpúsculos pequeñísimos cuyo choque contra el ojo o la piel producía las sensaciones de luz y calor.

Por otra parte, el arco iris ha sido uno de los fenómenos externos más tempranos observado por el hombre primitivo, “bonito” y efímero, que atrajo su misticismo, era incapaz de explicarlo y se inclinó por una importancia sobrenatural, y aunque familiar en los casos más simples de refracción, no la conectó con el arco iris-espectro en su forma natural. Tanto en Alejandría, Claudio Tolomeo (130), en su tratado de óptica, describió la medida de ángulos de incidencia y ángulos de refracción, como después el científico árabe, Alhazen (Abu Alí al-Hasan ibn al-Haytham) (1038), en observaciones similares, discutidas, de geometría Celeste, rechazadas por Nicolás Copérnico (1500), quien era consciente de la refracción atmosférica y de su efecto en la posición clara de objetos astronómicos, pero, no descubrió la ley de refracción, al igual que los científicos más tempranos. Más tarde, incluso Johannes Kepler, el descubrimiento real de que los senos de los ángulos de incidencia y refracción son una proporción constante, surge del trabajo de Willebrord Snell de Leyden (1591-1626).

1.2.2. **La ERA científica (Newton siglo XVII).**

La clasificación (taxonomía) tomando como referencia a Michel Adanson como su iniciador, a través de sus estudios del Estado Operativo, con la cantidad de información y caracteres de los individuos, y de Carlos de Linneo, hasta nuestros tiempos (hace 300 años), con un punto débil en la ausencia de significados cuantitativos en términos clasificatorios, siguieron y explicaron más las jerarquías existentes y rangos y relaciones doctrinarias en la clasificación como el Esencialismo, Cladismo, Evolucionismo y Feneticismo con la similitud

o fenotípica y los orígenes o genotípica que bloquearon el avance de la similitud y de allí el planteo evolucionista de Charles Robert Darwin.

El empirismo [John Locke, siglo XVII] epistemologicamente indica que todo conocimiento depende de la experiencia y toda teoría debe verificarse experimentalmente. Al aplicarlo en la clasificación no filogenética el empirismo lleva al feneticismo y salvo casos no verificables, puede aplicarse el operacionismo, que consiste en la descripción de operaciones que conducen a establecerlo mediante una enumeración de pasos a seguir, tales como las Relaciones Fenéticas o de similitud, cronísticas o temporales, de parentesco o filogenéticos y las espaciales o geográficas.

Darwin explica el fenómeno de la evolución por un mecanismo de mutaciones aleatorias sucesivas. Los individuos sufren a continuación la selección natural: los mejor adaptados sobreviven y se reproducen y los otros desaparecen, siguiendo a la Filosofía zoológica (1809) e Historia de los animales invertebrados (1815-1822), de Jean-Baptiste de Lamarck.

Para Aristóteles y en casi toda la pre ERA científica los individuos eran inmutables. De todas maneras hay que llegar al siglo XX para que se reconozcan estas hipótesis.

La obra de Sir Isaac Newton pues representa una de las mayores contribuciones a la ciencia. La contribución más específica de Newton a la descripción de las fuerzas de la naturaleza fue la explicación de la fuerza de la gravedad, la más débil, dedujo la ley de la gravitación universal, y la formulación de las tres leyes del movimiento, estableciendo las bases de la dinámica y las ideas tempranas del Color [Sawyer, R.A 1963].

La teoría corpuscular fue aceptada hasta el año (1800).

Al pasar del experimento a la especulación, pues todas eran de la misma naturaleza y origen, causadas por la vibración de partes de los cuerpos, como los colores de películas delgadas, como pompas de jabón coloreadas, no corroborada, serán explicadas por la interferencia.

Como una consecuencia de esta visión, Thomas Young (1802) intentó, para interpretar la línea amarilla en el espectro como un efecto de interferencia, y David Brewster y J. H. Gladstone (1860) usaron la misma explicación para las líneas de Fraunhofer.

La duda fue de Christiaan Huygens en el siglo XVII o siglo de las luces, quien, partiendo de los fenómenos observados de la transmisión de las ondas de agua sobre la superficie de un estanque o de las ondas sonoras a través del aire, sostuvo que la luz podría ser alguna perturbación vibratoria transmitida por algún medio que llena todo el espacio interestelar, que denominó éter luminoso o transportador de la luz.

Evolucionó, por otra parte, debido a las leyes de Newton de la mecánica y de gravitación universal y de la Mecánica Celeste debido a las ecuaciones de Johannes Kepler; y al aceptar, Newton, la teoría corpuscular, la teoría del éter o teoría ondulatoria, tuvo pocos adeptos, hasta que los fenómenos de interferencia, escapaban a cualquier explicación basada en la teoría corpuscular, mientras que eran explicados por la teoría ondulatoria.

Los hallazgos experimentales fueron expresados por las ecuaciones diferenciales en derivadas parciales de James Clerk Maxwell. Las ecuaciones de Maxwell relacionan los cambios espaciales y temporales de los campos permitiendo calcularlos en cualquier momento. Al resolver las ecuaciones se "*predice"* un nuevo tipo de campo electromagnético producido por cargas eléctricas en movimiento acelerado. Este campo se propaga por el espacio con la velocidad de la luz en forma de onda electromagnética (radiación). En 1887, Heinrich Hertz generó esas ondas por medios eléctricos, sentando las bases para la radio, el radar, la televisión y otras formas de telecomunicación: *"Concluyen en esa época con la corroboración y la respectiva explicación."*

Durante el siglo XIX la evidencia se hizo más notoria, y a fines de siglo la teoría corpuscular fue abandonada por: (1) Fenómenos de interferencia y superposición; (2) La Experimentación de la velocidad de propagación de la luz en distintos medios es distinta; (3) Haberse verificado las ondas hertzianas, semejantes a las ondas luminosas; (4) Verificarse que la velocidad de la luz es independiente de la velocidad de la fuente (radiador).

Para 1900, la teoría ondulatoria se convierte en una fortaleza, y el éter es un portador para las ondas electromagnéticas, así su existencia depende de las ondas en el vacío. La interferencia y la superposición se redujeron a una suma y resta de intensidades eléctricas y magnéticas.

En la Tesis se utilizan estos principios para conceptos Taxonómicos de los sistemas clasificatorios, como analogía en teorías emparentadas como se expresara en la primera parte, y al generar espectros de objetos y familias taxonómicas.

1.2.3. **La ERA de la ciencia Moderna.**

Comienza en el siglo XX (Einstein) y en los primeros 30 años, aparecen los principales conceptos en todas las disciplinas.

La taxonomía tradicional inclusive la post Darwiniana evolucionó en conceptos y procedimientos.
Las nuevas formas sistemáticas, el desarrollo de la genética (en particular de la genética de poblaciones, con Julian Sorell Huxley, John Scott Haldane, Ernst Mayr, Ronald Fisher, Sewall Wright y Theodosius Dobzhansky), la biogeografía y la paleontología aportan la base matemática y experimental a la teoría de Darwin constituyendo el neo darwinismo y la taxonomía (1920-1950).
Los avances en la genética (Gregor Johann Mendel), la citología y la variación geográfica condujeron a considerables progresos en el entendimiento de los mecanismos de la evolución de las especies y de las infraespecies, pero constituyó poco en la comprensión de la naturaleza y evolución de las más altas categorías y de la estructura taxonómica en general.
Son algo más que simples generalizaciones descriptivas (se intentó hacer todas y no se hizo ninguna bien) [**el autor**].

Los esfuerzos hechos a partir de asumir las bases filogenéticas, en forma sistemática, fueron un corsé para las observaciones taxonómicas y sus bases clasificatorias, que en realidad sirvieron para la descripción de patrones variacionales que de hecho existen en la naturaleza y conceptos e ideas de todos los niveles.
Se deben establecer criterios para definir categorías y operaciones, para no caer en discusiones científicas sin sentido.

En los años 1950 en adelante se produce un punto de inflexión cuando H J LAM define al taxón y George Gaylord Simpson y Blackwelder definen a la taxonomía numérica desarrollando su teoría y metodología.
En 1962 Sneath, Sokal y Rohlf, a partir de las metodologías, logran clasificaciones precisas y publican luego los principios de la taxonomía numérica.

Se describe la manera en la cual varios grupos de cosas existentes surgieron (o se originaron) es muy diferente a asignarle valores cuantitativos a esos grupos.
El enfoque de la clasificación o taxonomía numérica comprende un aspecto filosófico de la teoría de la clasificación o fenética y otro de técnicas numéricas, que son los pasos operativos para aplicar dicha teoría.
En ningún caso la respuesta es total y en muchos casos no puede ser respondida. Hasta que las respuestas puedan ser dadas adecuadamente nuestros esquemas clasificatorios nunca pueden ser satisfactorios o naturales. Pueden ser un poco mejor que mnemotécnicas (artificio para mejorar la memoria por medios artificiales o codificación), simples esqueletos o estructuras de las cuales sostenemos, suspendemos, (enganchamos o enlazamos) algunos fragmentos del conocimiento.
El rápido y explosivo desarrollo de la taxonomía numérica y el gran interés por este campo, produjo no solo nuevo material a ser estudiado sino también una estructura científica con su propia perspectiva. La influencia de estos métodos en otras disciplinas hizo que surgieran nuevos conceptos y técnicas para sistematizarla. Como la técnica cladística para caracteres continuos de Edwards y Cavalli-Sforza, y de caracteres discretos de Camin y Sokal, con investigación intra-OTU de Crovello y de clustering de Williams y Lance (1966-1968), con métodos generalizados de cladística numérica de clusters solapados, con centroides, de clases, de convergencia más rápida de MacNaughton-Smith, Jardin y Sibson (1964-1971).
Los propósitos fueron obtener teórica y prácticamente procesos clasificatorios y contrastar la visión convencional con los conceptos que sean susceptibles de evolución.

La evidencia taxonómica que implica la selección de objetos (organismos) de estudio, la selección y definición de caracteres taxonómicos y los criterios homológicos.

En general se trata de mantener un simbolismo uniforme para los caracteres, unidades taxonómicas operacionales (OTU) y taxones (taxones ≡ taxa: grupos de OTU´s).

La taxonomía numérica es el agrupamiento de unidades taxonómicas por métodos numéricos en TAXONES (TAXA) en base a los estados de sus caracteres.

Una clasificación es superior cuando tienen más leyes científicas que contribuyan más a la formulación de hipótesis explicativas, siendo un principio organizador del conocimiento, estable, pues no hay una modificación drástica por incorporar nueva información, robusta, ya que no es alterada por la incorporación de nuevos objetos y predictiva, pues todo nuevo objeto tiene las mismas propiedades de las entidades del grupo.

En *análisis de cluster*, que se aplica a la formación de conceptos en aprendizaje automático, pocos investigadores de Inteligencia Artificial han trabajado Michalski y Stepp, 1983, Stepp, 1987; Fisher y Langley, 1986, realizaron estudios de análisis de clusters pero no comprensibles para la Inteligencia Artificial.

Aldenderfer y Blashfield, 1984, para resolver el problema de la matriz de datos y de similitud derivada, cuando tiene menos información aconsejan utilizar el método de 'Monte Carlo'. Jaynes, 1986; propuso la teoría de la decisión, posteriormente lo hicieron Cheeseman, Kelly, Self, Stutz, Taylor y Freeman, 1988. Con el Teorema de Bayes se tienen valores confiables que con una clasificación Bayesiana se garantiza un clustering. Duda y Hart, 1973, junto con Pearson elaboran teorías de aplicación de invariantes y la normalización de los datos de la matriz original, para no tener valores despreciables.

Para las instancias de la matriz de similitud se trabaja con la distancia taxonómica, pudiendo utilizar así la distancia de Hamming y el *coeficiente de coincidencia de Jaccard* (Romesburg, 1984, p. 143) y finalmente el coeficiente de Cower de Sokal y Sneath, aunque siempre se termine utilizando la simple distancia Euclídea, o de la Teoría de la Información de Gluck y Corter (1985)

utilizan la categoría de utilidad, la sumatoria de algún valor de instancias dividido por la cantidad de instancias, de Fried y Holyoke (1984-1988).

La gran cantidad de información e instancias llevó a la utilización de computadoras y la necesidad de generar algoritmos eficientes, como Gennari et al., 1989.

Finalmente, usando distancias Euclídeas (o con métrica de Manhattan) puedo computar la matriz de Distancia Taxonómica o de Similitud o de Semejanza o Matriz de Coeficientes de Similitud, Matriz mediante la cual deseo encontrar la estructura taxonómica de dimensiones (t x t) donde t es el número de OTU´s.

Los clusters son los conjuntos de OTU´s en el hiperespacio, fenotípicos en término de patrones.

El centro del cluster o centroide representa un objeto promedio, que es simplemente una construcción matemática, que permite la caracterización de la Densidad y la Varianza, y el Radio y Rango del taxón.

Se postula que los coeficientes de correlación o de asociación pueden ser relacionados con las distancias.

En un hiperespacio se pueden representar las posiciones de los t OTU´s en un sistema de coordenadas, si dichas posiciones son cercanas, la distancia disminuye hasta hacerse cero si coinciden, así la distancia puede ser vista como el complemento de la similitud.

A partir de los dominios normalizados se calculan la diferencia media entre caracteres, tomándose el valor absoluto de la diferencia pues esta puede ser negativa, y la distancia taxonómica donde se pueden considerar las métricas de Minkowski y de Manhattan.

El método de clustering cuyo acrónimo es SAHN resume lo encontrado anteriormente: Sequential, Agglomerative, Hierarchic and Nonoverlapping.

En las técnicas en las cuales la estrategia es la distorsión del espacio parece como si el espacio, en la inmediata vecindad de un cluster se ha contraído o dilatado. Si volvemos al criterio de admisión para un candidato que se une a un cluster existente, este espacio vecino es constante sobre todo en el método pair-group.

El tratamiento dinámico e integrado de los dominios permite una fácil normalización, atributo - dominio - valor, y la implementación en el modelo de Base de Datos Relacional Dinámica y su utilización en Taxonomía Numérica.

La contribución teórica - empírica es la aglomeración de objetos formando clases producidas por pasos del método obteniendo clusters y dominios con valores normalizados y la densidad y el rango en términos del radio del conjunto puede ser visualizado como una INVARIANTE CARACTERÍSTICA de los OTU´s.

De acuerdo a los problemas astronómicos, caso de uso fundamental de esta tesis, pues examinando la distribución de los asteroides con respecto a sus elementos orbitales, en particular su movimiento principal, la inclinación y la excentricidad, se observan condensaciones en distintos lugares que parecen al azar, pero hay algunos casos en los cuales tener en cuenta solo las leyes de la probabilidad no es tan evidente (Hirayama,K. a partir de 1918).

Los asteroides están demasiado agrupados por tener inclinaciones cercanas o los planos de las orbitales tienen prácticamente el mismo polo, por ello es que se podía aventurar que existen familias de asteroides asociados.

Así para J. R. Arnold, 1969, la distribución de elementos orbitales en cinturones de asteroides no es al azar mostrando la existencia de familias se aproximan a clusters para ciertos valores especiales.

Según Arnold siguiendo la ley de Poisson el número de elementos de un conjunto debe ser menor que un cierto número esperado, con la cual no se concuerda en esta tesis pues los eventos no siguen esta ley por contradecir los grandores físicos, las características fenotípicas de caracteres o atributos de los asteroides y finalmente su genotípica u origen común.

Toda esa conclusión parece ser arbitraria pues debe prevalecer el concepto conservativo de la masa es decir la densidad y la estabilidad del entorno.

Condiciones de vecindad cercana deben ser tenidas en cuenta y las familias de alta densidad son las más estables y menos azarosas.

Investigadores han llegado a la conclusión que el problema de la clasificación de los asteroides en familias está claramente definido y prácticamente resuelto, visión simplista que en tal estado esta tesis no comparte.

Los criterios de rechazo de un miembro de una familia no son claros, son arbitrarios o directamente no se exponen en los trabajos y por lógica consecuencia no son automáticos.

Se puede observar que el crecimiento en observaciones entre 1969 de Arnold, y 1990, Carusi, Masaro, 1978, Williams, 1979-1989 y Knezevic, Milani, .1990, traen discrepancias.

Las discrepancias surgen de los métodos de cómputo de los elementos propios, del criterio de rechazo de los objetos a ser clasificados, del tamaño de la muestra, de los métodos de identificación de familias y los criterios de rechazo de un miembro de una familia.

Los métodos de cómputo de los elementos propios (no efemérides) tratan la eliminación de las perturbaciones seculares de planetas sobre el verificado Cinturón principal de asteroides.

El algoritmo permite calcular un código de calidad (QC) que indica cuantas iteraciones hay que realizar para que converja.

La cantidad de asteroides que se pueden recalcular es con alrededor de 55 iteraciones y siempre que la inclinación no sea grande al igual que la excentricidad (Jet Propulsion Laboratory, California Institute of Technology).

Todo este desarrollo aparece poco claro y arbitrario, no hay un sustento formal en la relación convergencia cantidad de iteraciones y el número de asteroides.

Las familias de Hirayama se estabilizaron con valores propios de decenas de miles de asteroides nuevos descubiertos.

Se produce así un nuevo punto de inflexión, segundo, al poderse desarrollar a partir de 1990 y 1994, algoritmos de cómputo que permitían obtener clusters en forma automática sin intervención arbitraria de investigadores condicionado al algoritmo ni nivel jerárquico de fenogramas o dendrogramas, donde se sacan primero los fenogramas, el investigador decide el nivel jerárquico y luego se encuentran los clusters o familias [Crisci, López Armengol, 1983, página 69].

También se utilizaron herramientas de la inteligencia artificial como la teoría de onditas (WAVELET), Algoritmos Genéticos y redes neuronales (RRNN).

Zappala, Cellino, Farinella y Knezevic (1990-1994) y Bendjoya & Cellino con Hergenrother (1992-1996), tienen un criterio que es importante donde una clasificación de los asteroides mejorada es nombrada en las familias dinámicas, mientras analizando una base de datos de asteroides numerados cuyos elementos propios se han computado en un nuevo segundo-orden, cuarto-grado de la teoría de perturbación secular, y verificada su estabilidad en términos

largos. El criterio multivariado usa la técnica de análisis de datos agrupándose en orden jerárquico. Fue aplicado para construir para cada zona del cinturón de los asteroides un "dendrograma", gráfico, en el espacio de los elementos propios, con una distancia en función relacionada a la necesaria velocidad incremental del cambio orbital después de la eyección del cuerpo del padre fraccionado.

Las familias se identifican entonces por la comparación con los dendrogramas similares, los derivados de una "casi randómica" distribución de elementos que comparan la estructura para una escala gruesa (bruta) de la distribución real.

Los parámetros de importancia asociados con cada familia, medidos como resultados de las concentraciones aleatorias, (como para transformar zonas anisótropas e in-homogéneas en zonas homogéneas e isótropas de las zonas intra-espacios (inter-gaps) en el cinturón de asteroides, modificando los atributos mecánicos como el semi-eje mayor y la inclinación) y los parámetros de robustez (estabilidad), se obtuvo repitiendo el procedimiento de la clasificación después de variar los elementos de velocidad en cantidades pequeñas al volver a computar las zonas reales de los cálculos con el cambio artificial de los coeficientes de la función de distancia.

Para tomar promedios de variación de distancias estaban armadas las designadas estalactitas, mientras tomando la anchura y la profundidad en la función de la velocidad modificada. Siendo un criterio innovador es importante analizarlo aunque no está claro la técnica del arreglo, mientras agrupándose, dentro de las zonas y los promedios de variación de velocidades, como antes de mencionó, y por otro lado se ignoran las familias de hasta cinco elementos y con solapamiento, todos son la síntesis de una instrumentación arbitraria.

Las familias más importantes y confortables son las de costumbre que juntas constituyen el 14% del cinturón principal conocido, de la población; pero 12 familias más confiables y confortables que se encontraron a lo largo del cinturón, la mayoría partió parcialmente de clasificaciones anteriores miembros son los taxonómico diferentes de los precedentes, de aquéllos con menos de cinco miembros no serán definitivamente diferentes (algo que no implica que ellos necesariamente y genéricamente serán "irreales").

Después de más de 100 años de las familias de Hirayama y los avances en Taxonomía Computacional llego a un nuevo criterio, que produce junto con intentos con Teoría de Onditas, Algoritmos Genéticos y Redes Neuronales a un

tercer punto de inflexión, que comienza en 1998 pero sigue con el nuevo siglo, XXI.

Con estas motivaciones, un Criterio Espectral, en el Análisis de Clasificación, he decidido lograr el análisis espectral, las clasificaciones se extendieron a la base de datos de los elementos propios de asteroides en las familias. Reconozco que los trabajos de Zappala son muy importantes (clasificación automática y método jerárquico), y un punto de inflexión en los tempranos 90´s pero es diferente el acercamiento porque trabajo en taxonomía computacional, en un hiperespacio taxonómico, y no en un criterio de composición y precedentes físicos y cosmoquímicos. Zappala y otros usan una metodología confundiendo, ambos, al tratar con sólo una variable de velocidad, un espacio transformado no claramente unívoco.

La decisión es lograr la clasificación en familias, que extienden el uso de la base de datos de elementos propios de asteroides, con un criterio de análisis espectral futuro. Incorporando así un conjunto actualizado y más grande de elementos oscilantes que se derivaron de la teoría de perturbación secular cuya exactitud (específicamente, la estabilidad en el tiempo) se ha verificado extensivamente por la integración numérica a largo plazo; en forma automática, y perjudicar la técnica de análisis de datos en los grupos del no-azar, no se usa en el espacio de elementos propios como en el criterio de Zappala y cuantitativamente la importancia estadística de estos grupos; con la robustez de las estadísticas para las familias importantes con respecto a las variaciones aleatorias pequeñas de elementos propios, todos basados en un análisis de Taxonomía Computacional.

No considero la transformación isotrópica y los conjuntos homogéneos, mientras cambiando los valores de la excentricidad y el semiejes al volver a computar los valores de las zonas de entre-espacios del cinturón de los asteroides en las velocidades en promedio, o los grupos eliminados de 5 o menos objetos y familias que se solapan, todos los cuales considero están fuera de un criterio Computacional.

Estos clusters constituyen familias, mediante el análisis estructural, basado en sus características fenotípicas, exhibiendo sus relaciones, en lo que se refiere a grados de similitud, entre dos o más OTU´s.

Entidades formadas por dominios dinámicos de atributos, cambien según los requisitos taxonómicos: la clasificación de objetos para formar familias o clusters.

Se representan aquí los objetos de Taxonomía por la aplicación de la semántica del Modelo de Base de Datos Relacional Dinámica.

Se obtienen familias de OTU´s empleando como herramienta i) las distancias Euclideas y ii) las técnicas del vecino más cercano. Así la evidencia taxonómica se reune para cuantificar la similitud para cada par de OTU´s (método pair-group) obtenido de la matriz del datos básica.

La contribución principal de la tesis presente es introducir el concepto de espectro de OTU´s, basado en los estados de sus caracteres. El concepto de espectros de familias surge, si el principio de superposición se aplica a los espectros de los OTU´s, y los grupos se delimitan a través del máximo de la relación de Bienaymé-Tchebycheff que determina Invariantes (centroide, varianza y radio).

Aplicando la técnica de dominios independientes dinámicamente integrados, para computar la Matriz de Similitud, y, con el recurso de un algoritmo iterativo, se obtienen familias o clusters.

Un nuevo criterio taxonómico es de ese modo formulado.

Una aplicación astronómica ha funcionado.

Así, un nuevo acercamiento a la Taxonomía Computacional se presenta, que ya ha sido empleado en la referencia a la Minería del Datos (Data Mining), para verificar la robustez del método.

Machine Learning es el campo dedicado al desarrollo de métodos de cálculo donde el aprendizaje subyacente procesa y se aplica, a los sistemas de aprendizaje basados en computadora, en problemas prácticos de Sistemas Complejos y Dinámicos. Data Mining intenta resolver esos problemas relacionados a la búsqueda de modelos interesantes y las regularidades importantes en las grandes bases de datos. Utiliza métodos y estrategias de otras áreas, incluso de Machine Learning. Al aplicar esas técnicas para resolver un problema de Data Mining, se dice que ésta es Inteligente.

En la tesis se analiza los TDIDT (Top Down Induction Trees), la familia de inducción y en particular al algoritmo C4.5. El intento es determinar el grado de eficacia logrado por el algoritmo de C4.5 cuando es aplicado en datos para

generar a modelos válidos de datos en problemas de clasificación con la Ganancia de Entropía (PME).

El algoritmo de C4.5 genera los árboles de decisión y la decisión gobierna a los datos pre-clasificados. "Divida y gobierne" es el método que se usa para construir árboles de decisión. Este método divide los datos de entrada en subconjuntos según algunos criterios preestablecidos. Entonces funciona en cada uno de estos subconjuntos, que los dividen de nuevo, hasta que todos los casos presentes en un subconjunto pertenezcan a la misma clase.

El algoritmo de Árboles de decisión por Inducción se desarrolló como un método de aprendizaje dirigido, para los árboles de decisión, los conjuntos deben tener un grupo de atributos y una clase. Los atributos y clases deben ser discretas, y las clases deben ser no solapadas. Las primeras versiones de estos algoritmos permitieron simplemente dos clases: positivo y negativo. Esta restricción se eliminó en descargos posteriores, pero la restricción de clases no solapadas se conserva. Las descripciones generadas por ID3 cubren cada uno de los conjuntos de entrenamiento.

El algoritmo de C4.5 es un descendiente del algoritmo de ID3, y resuelve muchas de las limitaciones de su predecesor. Por ejemplo, el C4.5 trabaja con atributos continuos, dividiendo los posibles resultados en dos ramas,: una para valores menores que un número dado, y otra para valores mayores que dicho número. Es más, los árboles son menos espesos porque cada hoja cubre una distribución de clases y no clasifica en particular como los árboles de ID3, esto obliga a refugiarse en un árbol menos profundo y más comprensible. C4.5 genera un árbol de decisión que divide recursivamente los datos, según la profundidad de la primera estrategia. Antes de hacer cada partición, el sistema analiza todas las posibles pruebas que pueden dividir el conjunto de datos y pueden seleccionar la prueba con la ganancia de información superior o la proporción de ganancia de entropía superior (ME). Para los atributos discretos, considera una prueba con los posibles resultados, de la cantidad de posibles valores que el atributo puede tomar. Para el atributo continuo, una prueba binaria se ha realizado en cada uno de los valores que el atributo puede tomar.

El principio de máxima entropía (PME) se aplica ampliamente, no sólo en la física, sino también en la meteorología, genética y en general en los procesos de cualquier naturaleza, dónde se quiere obtener información que empieza con un conjunto de datos incompleto, o usando la cantidad más pequeña en las

suposiciones anteriores. Para el caso de sistemas físicos, el PME proporciona una formulación alternativa de la Mecánica Estadística, elegante y compacta, formulada por Jaynes, que presenta al PME como un método canónico para construir la densidad principal, relativa a variables cuyos valores medios se conocen "a priori". Sin embargo esta manera constructiva no permite asegurar que la densidad, contraria principal, puede reproducir valores medios de otras magnitudes, del sistema en estudio, que no son parte de la información "a priori". Esta aserción falsa fue la más grande critica a la re-formulación de su ME referida al PME. Después se logra sacar esta limitación (con una formulación cuántica aplicable también al caso clásico), dando un método específico para encontrar a todos los operadores "excelentes" del sistema, esto es todos los operadores necesarios para describir la dinámica del sistema completamente. La densidad del operador de esta manera construida, es válida para la temperatura diferente de cero y fuera de equilibrio. Los Sistemas Dinámicos utilizan en los últimos años esta metodología.

Para Albert Einstein, los procedimientos operacionales, ligados a la noción de teoría, se aplican a las técnicas operacionales para introducir conceptos.
Los modelos funcionales, sistémicos, holísticos y homeostáticos hipotético-deductivos, mantienen una acomodación en estado de equilibrio donde unos paradigmas suceden a los otros, de acuerdo a la historia interna y externa de las disciplinas y sus teorías.
El reduccionismo metodológico permite solucionar problemas dentro del instrumentalismo y el realismo mediante construcciones de experimentos de laboratorio y de campo, integrando disciplinas en forma semántica y sintáctica.

En los estudios de Max Planck (1900), sobre los espectros de radiación del cuerpo negro (black body), a partir de consideraciones teóricas, para resolver las anomalías de la distribución de la energía, corrigió los argumentos que explicaban que era continua, considerándola discontinua, formada por corpúsculos, cuantos o fotones, haciendo coincidir los resultados experimentales con los teóricos.
La osadía de Albert Einstein, superó las dificultades del éter fibroso, en 1905, lo expuso en forma revolucionaria, vinculando los resultados de Planck.

Einstein supuso que la energía emitida por cualquier radiador no sólo se conserva en *cuantos*, al trasladarse por el espacio, como se había supuesto, sino también que una fuente dada podía emitir y absorber energía radiante únicamente en esas unidades.

Einstein encontró una ecuación que predecía correctamente todos los hechos observados, y en Investigaciones en el *Ryerson Laboratory* (1904-1915) y otros, surge una prueba abrumadora, de que la ecuación de Einstein es una ecuación exacta, de validez general, constituyó el progreso más conspicuo de la física experimental y del futuro de la ciencia, por la dualidad corpúsculo ondulatorio.

Se reconcilió la teoría ondulatoria con la corpuscular frente a esos fenómenos nuevos y perturbadores, suponiendo una estructura fibrosa del éter e imaginando toda la energía electromagnética, trasladándose a lo largo de líneas de fuerza, concebidas como verdaderas cuerdas, que se extenderían a través del espacio (ya en el siglo XXI se estudian fenómenos de campos y cuerdas, no éter fibroso, no aceptadas hacia fines del siglo XX, 1984).

1.2.4. **Estructura de los capítulos.**

En el presente capítulo (1) al igual que en el Prólogo, veo necesario, poner a la Epistemología como LA INVESTIGACIÓN EN CIENCIAS [Perichinsky, 2007, Epistemología], pues expone que el investigador, a través del método científico, va logrando un conocimiento verificable, al cumplir con las etapas de la investigación científica y lo crucial de la etapa de contrastación de hipótesis, mostrando a su vez que “es un largo viaje del intelecto que nació con el hombre y es materia viva evolutiva”, y que como cuerpo de doctrina metódicamente adquirido: objetivo en los hechos, interpretativo en las leyes, deductivo en las hipótesis sobre bases epistemológicas, especulativo en las teorías soportadas por una base empírica.

Esta base empírica del conocimiento es lo que permite la especulación objetiva, verificable y refutable, para el instrumentalismo y el realismo, al encadenar términos y enunciados mixtos para formar teorías.

Es difícil formar grupos y crear leyes abarcativas de problemas siendo más eficiente agruparlos en conjuntos disjuntos, de acuerdo al grado de

profundización en el tema, en el modelo hipotético deductivo, resolviendo el dilema o par <conocimiento, práctica> [Perichinsky, 1995, Investigation], mediante la fundamentación, la predicción y la explicación de los fenómenos, operaciones esenciales de las que se ocupa la ciencia [Hempel, 1996].

En el Capítulo sobre el Estado del Arte (2) se plantean los campos cuya convergencia a la resolución del problema de clasificación automática se plantean en esta tesis: clustering, base de datos, Data Mining como testeo del método de la propuesta y taxonomía y redes neuronales.

En el Capítulo sobre la Descripción del Problema (3), los problemas fundamentales no resueltos son la clasificación de objetos en familias o clusters mediante un método automático de formación de regiones, el cual es un proceso de agrupamiento de objetos en clases teniendo en cuenta sus relaciones y atributos comunes, a los efectos de realizar estudios de sus características y las propiedades estructurales y de su comportamiento relativo, en la tesis se da solución a los mismos y la utilización de Data Mining para corroborar su fortaleza. La utilización del Teorema de Tchebycheff, para el manejo de invariantes, y de la Espectroscopia como se explica en el Capítulo (5) para llegar a los espectros de objetos y de familias, sin tener que utilizar algunas maniobras arbitrarias o no claramente explicadas, en otros criterios.

En el Capitulo sobre la Solución Propuesta (4) se introduce la matriz de datos, se define el proceso su construcción y se describe el proceso de normalización asociado. Se presenta la matriz de similitud y se definen los espectros de OTU´s y familias. Se analiza la caracterización, la dispersión y la normalización del rango de la dispersión. El refinamiento de invariantes utilizando el teorema de Tchebycheff, la distribución de masa y la aplicación de equiprobabilidad para la Entropía Máxima (PME). Se muestra la algoritmia asociada que incluye el análisis de conservación del espacio y la distorsión ínter cluster por vecindad. La Representación Taxonómica en Base de Datos, Dominios Estandarizados de OTU´s en Bases de Datos Relacionales Dinámicas y Data Mining para testear la fortaleza de la solución propuesta y la Algoritmia.

En el Capítulo sobre Fenomenología Física (5) se plantean los principios de interferencia y superposición y las analogías asociadas, se vincula el análisis espectral y la dinámica de fasores con el hiperespacio taxonómico y su dinámica vectorial de distancias normalizadas.

En el Capítulo sobre Sistemas Complejos y Dinámicos (6) se explican y se describen los Sistemas Complejos y Dinámicos, con sus herramientas de Mecánica Estadística y los métodos de Ciencias de la Computación desde la Mecánica Clásica y Cuántica, y se plantea la Teoría de la Información, la Entropía el Principio de máxima Entropía, la Distancia de Hamming y la Ganancia Entrópica con su aplicación en Data Mining, la Metodología y la Minería de Datos Inteligentes (Data Mining). Para los Algoritmos de Quinlan ID3·y C4.5, para TDIDT (Top Down Induction Trees) familias de Inducción y árboles de decisión. En Ciencias de la Computación se explica la Taxonomía Computacional.

En el Capítulo sobre la Aplicación (7) se plantea como un caso de experimentación la clasificación automática de cuerpos celestes, en particular de familias de asteroides. El Principio de Máxima Entropía, la Metodología de Ganancia de la Información relativa a la Entropía y la Minería de Datos Inteligentes (Data Mining) y el Testeo para mostrar la fortaleza del método. Se realiza un enfoque de la Ingeniería de Requerimientos, desde de la Ingeniería de Software, para el caso de experimentación se exponen los planteos y conjeturas o hipótesis de Hirayama y desde Arnold hasta Zappala, y su problemática de identificación de cuerpos celestes y de elementos propios no perturbados. Se plantean los matices de la implementación para la atención del caso de estudio en la matriz de datos, la matriz de similitud y la estructuración. Finalmente se presentan los resultados desde el "***nuevo criterio"*** de Análisis Espectral de Taxonomía Computacional, para la familia María según la Clasificación de Hirayama.

En el Capítulo sobre Algorítmia (8) se describe la versión general, de los algoritmos originales desarrollados, para llevar adelante en forma automática la

clasificación automática basada en análisis espectral aplicando la identificación de taxones por normalización e invariantes.

En el Capítulo sobre las Conclusiones (9) se identifican los aportes originales de la tesis a los distintos campos de conocimiento y se plantean futuras líneas de investigación.

En el ANEXO I equivalente al Capítulo (10), se muestran los Elementos de la Entidad de la Matriz de Datos y las Instancias de la Matriz de Datos.

En el ANEXO II equivalente al Capítulo (11), como en el Plan de Tesis se plantean los trabajos en Redes Neuronales que se realizan por Metodologías Automáticas y por Interacción, en el Laboratorio PROTEM, del Departamento de Física de la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata, (Programa PROTEM del CONICET), se expone una Introducción, un poco de historia, la Neurona Artificial, la Red de Neuronas Artificiales, el Clasificador Básico o Perceptrón, el Algoritmo de Aprendizaje y la Separabilidad Lineal, el Perceptrón Multicapa , las Memorias Asociativas y la Red de Hopfield, Aprendizaje no Supervisado y Redes Constructivas, ART y Refuerzo

Finalmente, el ANEXO III equivalente al capítulo (12) donde se realiza una clasificación aplicando el Nuevo Criterio de la Tesis, a un conjunto de familias en Botánica, para corroborar, la clasificación realizada en "Introducción a la Teoría y Práctica de la Taxonomía Numérica", género Bulnesia y sus Especies (Zygophyllaceae) [Crisci, López Armengol, 1983, página 30], en base al método SAHN y deduciendo criterios con fenogramas y coinciden los clusters y familias, dando lugar a la primera contrastación del nuevo criterio [**el autor**].

ESTADO DEL ARTE

La información se transmite ,se consume, gratis, la tecnología se usa para acelerar el consumo, no siempre gratis, la ciencia, lenta, potencia el conocimiento por su método deductivo, se evita, no se compra.
El Autor, en base a una idea de Oscar Wilde.

XXXIV

2. ESTADO DEL ARTE

2.1. ESTADO DEL ARTE EN CLUSTERING

2.1.1. INTRODUCCION

Investigadores en diversos campos han estudiado la tarea básica de *clusterizar* instancias en clases. Aunque instanciaciones específicas de esta tarea difieren de un campo a otro, la siguiente es una presentación general del problema:

- Dado: un conjunto de instancias, cada una descripta por algún número de pares atributo-valor;
- Hallar: un conjunto de clases que agrupe esas instancias.

Por ejemplo, supóngase súbitamente ubicado en la jungla de un planeta desconocido. Como un agente de aprendizaje, uno inmediatamente comenzará a crear conceptos para clasificar y organizar las instancias observadas (en plantas, animales o cualesquiera objetos percibidos). Este ejemplo enfatiza la naturaleza *no supervisada* del problema - el aprendiz está tratando de imponer una estructura a su alrededor sin ninguna realimentación.

Una dificultad fundamental para la tarea de clustering es que requiere algunas formas de evaluación de un conjunto de clases potenciales. Por ejemplo, una *función de evaluación* mide la categoría de un conjunto de clases con respecto a los datos. La creación de una función de evaluación está muy ligada a la definición de algún criterio de *similitud* entre las instancias. Por otro lado, los atributos (y tipos de atributos) que describen las instancias afectan las medidas de similitud. Además, existen diferentes algoritmos que crean clases a partir de instancias. Aunque algunos algoritmos requieren funciones de evaluación específicas, a menudo el investigador puede probar un conjunto de diferentes funciones con un único algoritmo y viceversa.

En este trabajo, plantearé un gran número de técnicas de clustering bajo un marco de referencia como un espacio de métodos posibles. Mi impresión es que esto ofrecerá una mayor comprensión que simplemente listar diferentes métodos de distintos campos o que tratar de definir la ‘mejor’ técnica o la ‘óptima’. A pesar de que el marco de referencia que desarrollo en este estudio

probablemente no cubra todos los métodos de clustering, creo que trae a la luz algunas características internas interesantes y que describe un amplio rango de métodos posibles.

Está dirigido a graduados y/o investigadores en aprendizaje automático o inteligencia artificial en general, que no han estado al tanto de trabajos fuera de su propio campo. Existe un amplio campo de investigaciones en estadística y en biología, generalmente conocido como *análisis de cluster*, que se aplica al trabajo en aprendizaje automático, si uno lo tiene en cuenta para las diferentes ramas de estas disciplinas. Aunque unos pocos investigadores de Inteligencia Artificial han trabajado en este área [Michalski y Stepp, 1983a; Stepp, 1987; Fisher y Langley, 1986], en él se realizaron estudios de análisis de clusters no comprensibles para la Inteligencia Artificial. En particular, el análisis de cluster es muy similar al estudio de *formación de conceptos* en aprendizaje automático. Un objetivo de este trabajo es enfatizar esta similitud y mostrar cómo investigadores en aprendizaje automático pueden beneficiarse con el conocimiento sobre análisis de cluster.

Comienzo este informe presentando puntos de vista del problema de clustering mirado desde varias perspectivas diferentes, empezando con un enfoque desde el aprendizaje automático. En la tercera sección describo la dificultad e importancia de elegir una medida de similitud o una función de evaluación; esta sección incluye también algunas de las medidas más importantes y usuales. Continúo con una descripción de algoritmos que utilizan estas medidas y concluyo con una discusión acerca de la dificultad de validar o evaluar una técnica de clustering.

2.1.2. PARADIGMAS Y TRAYECTORIAS

Más que intentar dar una definición más precisa sobre la tarea de clustering, en su lugar describiré el problema desde cuatro paradigmas diferentes: aprendizaje automático, biología, estadística y teoría de la decisión. De esta manera, los objetivos y trayectorias de los investigadores de diferentes campos serán explícitos.

Mis propios caminos para llegar al clustering provienen del paradigma de aprendizaje automático y la terminología utilizada en este estudio proviene de

ese campo de la literatura. Con ambos, para dar una idea acerca de la diversidad de la terminología y para hacerlo accesible a lectores provenientes de otros paradigmas, la siguiente es una breve lista de términos técnicos 'traducidos' (los términos utilizados en este trabajo son los que figuran al final de cada renglón en letra itálica).

- **un objeto, unidad taxonómica operacional (OTU), evento o caso es una *instancia* (de una metaclase).**
- **los caracteres, características o variables que describen una instancia son *atributos* (que describen a los objetos).**
- **las métricas de distancia, medidas de asociación o coeficientes de similitud que comparan instancias son *medidas de similitud* y**
- **el criterio de optimización es una *función de evaluación* relacionada con la ley de clausura.**

Debido a que los investigadores han abordado el problema de clustering desde tan diferentes perspectivas, no sólo se trata el problema descrito con diferente terminología, sino que también la tarea en sí misma varía levemente. Esto es apenas sorprendente, la biología tiene objetivos muy diferentes a los del aprendizaje automático. Puesto que deseo comparar otros esfuerzos distintos a aquellos del aprendizaje automático, comienzo con este paradigma.

2.1.2.1. CLUSTERING EN APRENDIZAJE AUTOMÁTICO

Desde la perspectiva del aprendizaje automático, el clustering se ve como un problema de formación de conceptos. Como gran parte de la inteligencia artificial, este campo se enfoca hacia una analogía entre la computación y los procesos humanos. Por lo tanto, el proceso de clustering y los conceptos generados por ese proceso poseen implicancias en el estudio (del ser humano) y en la organización y representación del conocimiento (del ser humano).[1] A pesar de que el aprendizaje automático no siempre hace explícita esta conexión,

[1] Estos enfoques son desde el campo estrechamente relacionado de la psicología cognoscitiva, donde el objetivo explícito es estudiar los procesos cognoscitivos humanos. La formación de conceptos ha sido estudiada por psicólogos cognoscitivos tales como Smith y Medin (1981), Mervis y Rosh (1981), Barsalou (1987), Corter, Gluck y Bower (1988) y Anderson (1988).

generalmente existe por lo menos una débil analogía con el sistema humano de estructuración de aglomeración de objetos (clustering).

A modo de ejemplo de una aplicación para formación de conceptos considere un robot explorador que percibe una secuencia de diferentes pelotas. Incluso si el robot está equipado con un sistema perceptivo que reduce cada instancia a un conjunto de pares atributo-valor, todavía deberá crear y organizar un útil conjunto de conceptos sobre estas pelotas. Por ejemplo, luego de observar unas pocas pelotas de béisbol, deberá crear un concepto para ellas y ser capaz de reconocer una nueva pelota de béisbol como un miembro de esa clase y no como una instancia de alguna otra clase (pelota de voleibol, pelota de tenis, etc.).

Una característica que distingue la formación de conceptos es que las clases aprendidas deben ser por *intensión*, más que por extensión. Por ejemplo, la clase pelota de béisbol debe ser una "descripción conceptual" de las pelotas de béisbol vistas, más que simplemente una lista de todas las instancias componentes [Michalski y Stepp, 1983b]. Este énfasis en definiciones de conceptos por intensión significa que las funciones de evaluación que comparan clases son más apropiadas para la formación de conceptos que las medidas de similitud que comparan instancias.

Un segundo aspecto de la formación de conceptos es que las clases aprendidas están generalmente organizadas en una jerarquía de conceptos. Esto es, los conceptos aprendidos están organizados en una forma más general jerárquica, conceptos inclusivos hacia la parte superior y más específicos, conceptos exclusivos hacia la parte inferior. Esto refleja la naturaleza jerárquica del conocimiento en dominios típicos de aprendizaje automático. Por ejemplo, pelotas de fútbol y de voleibol tienen más similitud entre ellas que con pelotas de béisbol o de crosse (raqueta). Una jerarquía natural para estos cuatro tipos de pelotas sería poner las pelotas de fútbol y de voleibol juntas en una clase más general, "blanda, grande" y las pelotas de crosse y béisbol en una clase "dura, pequeña".

Una tercera característica de la formación de conceptos es que el aprendizaje se produce en forma *incremental*. Como el robot observa cada pelota sucesiva, la debería agregar a su conocimiento inmediatamente; los conceptos aprendidos se actualizan con cada nueva experiencia sin reprocesar instancias previas. En

contraste, un sistema no incremental debe recibir el conjunto entero de instancias antes de generar un conjunto de clases. Este tipo de sistema es incompatible con los objetivos de la formación de conceptos pues uno puede no conocer el 'conjunto' completo de instancias y uno puede necesitar utilizar los conceptos aprendidos en cualquier instante de tiempo. Por ejemplo, el robot deberá ser capaz de utilizar su conocimiento en cualquier punto durante el aprendizaje y deberá seguir aprendiendo, sin tener en cuenta la cantidad de pelotas encontradas. Estos problemas son quizás más obvios para los aprendices humanos, quienes observan una secuencia de instancias sin fin.

Investigadores en aprendizaje automático están generalmente interesados en algoritmos robustos más que en métodos de clustering de propósito específico. Un investigador, por lo tanto, aplicará su método a una amplia variedad de dominios, incluyendo a menudo grandes conjuntos de datos con ruidos. Hallar un único método de clustering que trabaje con un gran número de variados dominios está motivado por la evidencia psicológica de que existe al menos un algoritmo semejante: el sistema humano de clustering.

Finalmente, si un sistema aprende, uno debería ser capaz de medir sus progresos en alguna tarea a desempeñar. Esta es una tarea utilizada para evaluar (y cuantificar) la habilidad del sistema antes y después del aprendizaje. Con este tipo de medida numérica, el éxito de una formación de conceptos puede ser evaluado sobre un número de dominios diferentes o un conjunto de sistemas diferentes pueden ser comparados con un conjunto de datos dados.

Como he descrito el problema, el de clustering es *no supervisado*. Sin embargo, existe también una gran cantidad de trabajos en aprendizaje automático sobre formación de conceptos no supervisado generalmente conocidos como "aprendiendo de ejemplos". Si bien esta es una tarea que está relacionada, las diferencias entre estos dos problemas son muy importantes. La formación de conceptos supervisada aprende a determinar a cuál, dentro de un conjunto de clases conocidas, pertenece una instancia, mientras que el aprendizaje no supervisado impone una estructura de conceptos (aquellos que no son conocidos a priori) en el conjunto de instancias.

2.1.2.2. CLUSTERING EN BIOLOGÍA

Históricamente, las primeras aproximaciones de la computación hacia el clustering surgieron en el campo de la biología. En particular, la 'taxonomía numérica' tuvo su origen en el siguiente problema: dado un conjunto de especies u organismos, hallar la jerarquía taxonómica que los organice en especie, género, familia y clases. El principal propósito de esta jerarquía es plantear relaciones de evolución entre individuos.

Por ejemplo, suponga que un biólogo descubre un nuevo conjunto de gusanos. Luego de describir cada gusano por medio de algún conjunto de atributos (o "caracteres"), el biólogo ingresa este conjunto de instancias a un sistema de clustering. La jerarquía resultante debe definir clases de gusanos similares, así como también mostrar cómo esas clases se relacionan una con la otra. Esta información puede conducir a predicciones sobre cómo pueden comportarse gusanos relacionados entre sí y también a teorías acerca de cómo los atributos de los gusanos han evolucionado a través del tiempo.

El énfasis en esta disciplina es encontrar un método práctico en lugar de preocuparse acerca de las implicancias teóricas de una técnica particular. Por ejemplo, los biólogos no están muy interesados en validar o medir la eficiencia de una técnica dada. En su lugar, su medida de éxito es subjetiva: si un método produce una taxonomía útil (una que lleve a nuevas características internas o tenga interesantes implicancias de evolución), entonces es bueno. Este camino ha conducido a otros investigadores a establecer que "la visión, de los biólogos, acerca del clustering es considerada una aproximación radicalmente empírica" [Aldenderfer y Blashfield, 1984, p. 21].

Puesto que estos métodos están buscando explícitamente una taxonomía, las clases halladas deberían ser separadas y organizadas en una jerarquía. Una lista sencilla de clases o un conjunto de clases solapadas no es tan útil. Sin embargo, en contraste con métodos de aprendizaje automático, las clases creadas a través de técnicas de taxonomía numérica son usualmente por extensión, y medidas similares, más que funciones de evaluación, son usadas como la base para el clustering.

Aunque un algoritmo de propósito general, robusto, es apreciado en cualquier campo, éstas no son características relevantes para taxonomía numérica. En este paradigma, es razonable utilizar diferentes algoritmos para diferentes conjuntos de datos. También, para cualquier problema dado de clustering, el

investigador sólo está interesado en un conjunto finito de instancias (que es generalmente pequeño). Por lo tanto, la habilidad de un método de clustering para nuevas instancias en un proceso incremental no es muy importante.

Este estudio enfatiza métodos de clustering biológico; de cualquier modo, muchos de estos mismos caminos pueden ser vistos cuando son utilizadas técnicas de clustering en otras ciencias. En particular, el mismo énfasis en métodos prácticos, medidas de similitud y validación subjetiva aparecen en clustering para ecología y psicología.

2.1.2.3. CLUSTERING EN ESTADÍSTICAS

El estadístico tiene una visión mucho más formal del problema de clustering. En esta aproximación, los investigadores están interesados en una cuidadosa definición de clustering y en explorar implicaciones teóricas de métodos de clustering. Aunque este paradigma ha tenido algún éxito, la naturaleza heurística del clustering puede ser un obstáculo para el tipo de análisis riguroso preferido por los estadísticos. Asimismo, la evaluación objetiva del resultado de una técnica de clustering (validación) ha sido dificultosa.

Para el estadístico, un punto de partida es la comparación del análisis de cluster con otros bien establecidos métodos estadísticos tales como análisis de factores, análisis de varianzas y análisis por discriminante. Por ejemplo, los estadísticos señalan que la elección de un conjunto de atributos que describa instancias es el problema general tratado por el análisis de factores. De cualquier modo, aparece el análisis de factores de rendimiento como un paso de pre-proceso que tiene un efecto de detrimento en el clustering.[2] Similarmente, la práctica multivariante estándar de normalizar variables puede causar problemas: la normalización puede oscurecer diferencias que pueden ser esenciales para el clustering [Everitt, 1980].

Los estadísticos han analizado también y comparado los algoritmos y funciones de evaluación de métodos de clustering por ellos mismos. Si bien este esfuerzo ha demostrado que algunos métodos y medidas de similitud son redundantes

[2] Existe un debate considerable sobre este tema. Véase Everitt (1979) o Aldenderfer y Blashfield (1984) para más discusión.

[Anderberg, 1973], no ha sido capaz de establecer ningún simple método de clustering como el mejor. La dificultad es que, al contrario de la mayoría de los métodos estadísticos, el clustering es *heurístico*. Puesto que los algoritmos utilizan 'reglas de aproximación' que no garantizan la producción de soluciones correctas, son difíciles de analizar y comparar.

Aunque uno no puede medir el fin subjetivo de encontrar un 'interesante' conjunto de clases, los estadísticos están interesados en evaluar cuantitativamente aspectos de una solución. A diferencia de la perspectiva de los biólogos, la 'solución' de los estadísticos no necesita ser una jerarquía de clases: para algunos dominios, una lista sencilla de clases es más apropiada; para otros, pueden ser preferibles clases solapadas o probabilísticas.

2.1.2.4 CLUSTERING COMO TEORÍA DE LA DECISIÓN

Una abstracción del problema de clustering ha sido estudiada por unos pocos investigadores en el campo de la teoría de la decisión [Jaynes, 1986; Cheeseman, Kelly, Self, Stutz, Taylor y Freeman, 1988]. En esta visión, el objetivo es predecir correctamente las probabilidades de que una nueva instancia *x* sea un miembro de una clase ω_i: $P(\omega_i | x)$.

Esta expresión puede escribirse usando el teorema de Bayes:

$$P(\omega_i | x) = \frac{P(x|\omega_i) P(\omega_i)}{P(x)} \qquad 2.1.2.4.1$$

La probabilidad de cada clase, $P(\omega_i)$ es conocida usualmente, puede ser calculada como el número de componentes de ω_i dividido por el número total de instancias. Adicionalmente, puesto que $P(x)$ es la probabilidad de que *x* sea independiente de la clase, puede ser ignorada, al comparar dos clases diferentes, ω_1 y ω_2, el denominador es el mismo y puede ser eliminado. De hecho, el único término no conocido en el lado derecho es la función de densidad probabilística condicional de la clase, $P(x|\omega_i)$. Esta es la probabilidad de *x* dada ω_i, o el valor de *x* que sería predecido por la clase ω_i. Estas funciones deben ser estimadas; por ejemplo, se puede asumir una distribución normal y buscar una buena estimación de los parámetros μ y σ que caracterizan esta distribución.

Usar esta base para un método de clustering garantiza que el método maximizará la probabilidad de caracterizar correctamente el dato. Por lo tanto, si un sistema es confiable a la teoría, es “probablemente” óptimo y existe poca necesidad de una validación empírica. Por esta razón, los investigadores de este paradigma ponen menos énfasis en los algoritmos usados para implementar una clasificación Bayesiana. Sin embargo, usualmente es dificultoso crear un sistema que sea fiel a la teoría y deben efectuarse a lo largo del camino un número de suposiciones. Muchas soluciones basadas en la teoría de la decisión enfrentan los problemas usuales de la búsqueda heurística, pero a un nivel algorítmico más bajo. La principal ventaja de la estructura Bayesiana es que proveyendo una subyacente teoría de decisiones, el investigador puede fácilmente explicitar las implicaciones teóricas y suposiciones necesarias para cualquier solución de la tarea de clustering.

2.1.3. MEDIDAS DE SIMILITUD Y FUNCIONES DE EVALUACIÓN

Si bien no siempre está explícito, cualquier técnica de clustering produce un conjunto de clases, en el cual los componentes de una clase dada son similares entre sí de alguna manera. Utilizando la terminología de ‘búsqueda’, uno busca clases en las cuales la similitud entre instancias dentro de una clase dada es mayor que aquella entre instancias provenientes de diferentes clases. Desde esta perspectiva, la técnica de clustering puede ser caracterizada por lo que se define como ‘similitud’.

Para algunas técnicas este concepto está explícito: existe una *medida de similitud* o una *métrica de distancia* que mide cuantitativamente la distancia entre dos instancias (OTUs). Para otras, el objetivo es maximizar alguna *función de evaluación* o *criterio*. Una función semejante mide la ‘bondad’ de un conjunto de clases; usualmente esto está basado en la similitud entre clases, más que entre dos instancias.

Clustering depende de una medida de similitud; de otra manera, los valores resultantes de una medida de similitud dependen de los atributos utilizados para describir una instancia. Comienzo esta sección describiendo algunas discusiones y problemas referentes a atributos. Luego, presento un conjunto de medidas de similitud seguidas por funciones de evaluación. Esto se describe

independientemente de los algoritmos de clustering que las utilizan, de manera tal que puedan ser comparadas directamente. Esto también clarifica el concepto de que generalmente hay más de una elección de medida o de función de evaluación disponible para el investigador.

2.1.3.1 ATRIBUTOS: ELECCIONES Y REPRESENTACIONES

Para el biólogo o el científico social quienes se aproximan a las técnicas de clustering para utilizarlas como una herramienta, existe un potencial número infinito de atributos disponibles para describir una instancia. Antes que usar a ciegas tanta información como pueda ser hallada, los científicos pueden elegir sólo aquellos atributos que son 'relevantes' a la tarea. Como se mencionó anteriormente, análisis de factores es un método estadístico bien definido que es usado a veces para crear un número más pequeño de atributos más 'apropiados' del conjunto de atributos disponibles.

Desafortunadamente, esto es algo así como una lógica circular. A pesar de su uso difundido, Everitt (1979) argumenta en contra de usar análisis de factores a cualquier otro método que elimina atributos antes de utilizar clustering. El propósito de clustering es descubrir un conjunto de clases desconocidas. Mientras busca ésto, establecerá cuáles atributos son 'relevantes', pero para decidir esto de antemano se debería tender hacia el uso de un proceso de clustering. El análisis de factores puede tener el efecto perjudicial de esconder aquellos atributos que puedan ser decisivos para encontrar una jerarquía de clases. El análisis de factores asume una única clase conocida; por lo tanto, Everitt sugiere que puede ser utilizado luego de clustering pero nunca de antemano. Los investigadores también a veces colocan pesos a los atributos antes del clustering. Esto tiene el mismo efecto que usar análisis de factores y es vulnerable a la misma crítica.

Cualquier método para medir similitud depende en cierto grado de la representación utilizada para los atributos que describen una instancia. Anderberg (1973) señala que existen dos maneras de caracterizar atributos: la escala de medida utilizada para el atributo y el número de posibles valores que

un atributo puede tomar. En este trabajo describiré cuatro tipos principales de atributos: continuo, ordinal, simbólico y binario.[3]

Los atributos *continuos* tienen un rango infinito y son medidos a lo largo de una escala continua. Ejemplos de este tipo de atributos son las medidas reales estimadas de grandores (v.g: altura, peso, temperatura, etc.). Un atributo *ordinal* tiene un rango finito con un ordenamiento en los posibles valores de atributos. Por ejemplo, número de aletas o cualquier atributo continuo que ha sido redondeado, como la edad al año más próximo. Un atributo *simbólico* también tiene un rango finito pero no existe orden entre sus valores. Ejemplos de esto pueden ser forma, lugar de nacimiento o tipo de bote de vela. Finalmente, un atributo *binario* tiene sólo dos valores posibles. A menudo, éstos son atributos de presencia-o-ausencia y/o tales como tiene-firmeza o está-hambriento.

La medida de similitud o función de evaluación utilizada dependerá de los tipos de atributos usados para describir instancias. De hecho, las instancias pueden ser descriptas por una combinación de diferentes tipos de atributos. Desafortunadamente, uno de los tantos problemas no resueltos en el análisis de cluster (desde un punto de vista estadístico) es que no existe una 'buena' manera para combinar diferentes tipos de atributos. Esto es, a pesar de algunos intentos, no existen medidas de similitud teóricamente sólidas que puedan ser aplicadas a diferentes tipos de atributos, especialmente si se combinan atributos binarios y continuos. Por esta razón, las medidas descriptas más abajo están organizadas de acuerdo a si ellas son apropiadas para atributos continuos, ordinales, simbólicos o binarios.

2.1.3.2 MEDIDAS PARA ATRIBUTOS CONTINUOS U ORDINALES

Comenzaré considerando medidas de similitud para atributos continuos: medidas comparando dos instancias que son descriptas por un conjunto de atributos continuos u ordinales. Sean i y j las instancias, cada una descripta por K atributos, e.g., $i = \{ x_1, x_2, .., x_K \}$. Una de las medidas de similitud más obvias entre estos dos puntos es usar una *métrica de distancia.* Cada atributo define

[3] Nótese que mi terminología es diferente de la de Anderberg, reflejando mi preferencia por aprendizaje automático.

una dimensión y cada instancia puede ser representada en este espacio *K*-dimensional. La distancia Euclideana entre dos puntos es:

$$D_{ij} = [\sum_{k=1}^{K} (x_{ik} - x_{jk})^{2}]^{1/2} \qquad 2.1.3.2.1$$

Una métrica más simple y relacionada es la distancia de Manhattan:

$$D_{ij} = \sum_{k=1}^{K} |x_{ik} - x_{jk}| \qquad 2.1.3.2.2$$

La distancia Euclídea es probablemente la medida de uso más apropiada para similitud; se la utiliza también como la base para algunas funciones de evaluación. Aunque una distancia de Manhattan puede parecer menos intuitiva, es computacionalmente más económica y puede ser apropiada cuando son utilizados atributos ordinales.

Ambas medidas son sensibles a transformaciones lineales de los datos de entrada. Por ejemplo, si algunos atributos son convertidos de, digamos, pulgadas a millas, entonces estos datos normalizados tendrán medidas de similitud completamente diferentes.[4] Como señalan Duda y Hart (1973), esto puede o no ser un problema, depende si tales transformaciones son naturales a su dominio.

El uso de la medida de correlación de Pearson es una manera de lograr invariantes respecto a transformaciones lineales. El intento original de esta medición es correlacionar pares de atributos para análisis de factores. En orden de correlacionar instancias, los investigadores simplemente revirtieron la ecuación sintácticamente. Esto es, imagine el conjunto de datos como una *K* x *N* matriz de valores (si existen *K* atributos y *N* instancias), el uso corriente de correlación es medir similitud entre las columnas (los atributos), por tanto la ecuación inversa debería medir similitud entre las filas (las instancias). Esta correlación inversa está referida como un análisis de factor tipo Q [Anderberg, 1973, p. 113]. Ella establece que la correlación (distancia) entre dos instancias es:

[4] Nótese que este 'atributo normalizado' está relacionado con el peso del atributo descrito anteriormente. El peso del atributo es una práctica controvertida sólo si la medida de similitud utilizada es sensible a transformaciones lineales de los datos.

$$Dij = \frac{\sum_{k}^{K}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{[\sum_{k}^{K}(x_{ik} - \bar{x}_i)^2 \sum_{k}^{K}(x_{jk} - \bar{x}_j)^2]^{1/2}} \qquad 2.1.3.2.3$$

donde $\bar{x}_i = 1/K \sum_k^K x_{ik}$ es el valor promedio del atributo para una instancia dada, *i*.

Esta aproximación ha sido usada con algún éxito entre los investigadores en psicología [Aldenderfer y Blashfield, 1973, pp. 22-23]. Sin embargo, esta medida está muy desacreditada (especialmente en otros campos) porque no existe justificación para la inversión sintáctica. El significado de la ecuación se pierde: por ejemplo, puesto que $\bar{\bar{x}}$ es un promedio de atributos diferentes, podría estar promediando 'manzanas y naranjas' y podría no tener la semántica esperada para ese término.

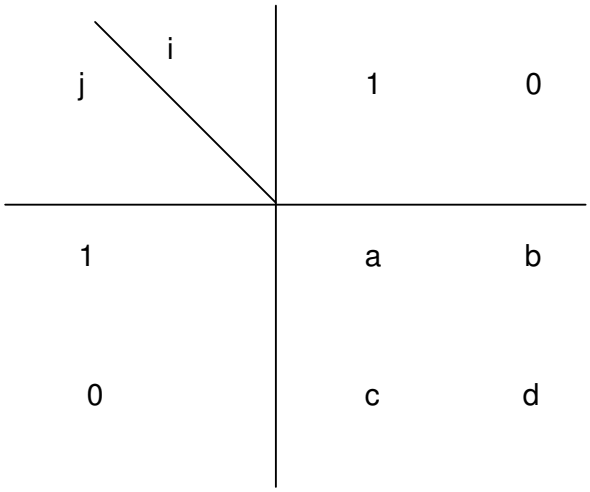
2.1.3.3 MEDIDAS PARA ATRIBUTOS BINARIOS O SIMBÓLICOS

Ni la correlación ni la distancia Euclídea pueden ser aplicados a un atributo con valores binarios o simbólicos. Una característica de un atributo semejante es que, dados dos valores, la expresión $x_i - x_j$ no tiene ningún significado. Una medida de similitud para atributos simbólicos está enfrentada con una simple comparación: o los dos valores son iguales, o ellos son diferentes. Sobre un conjunto de atributos, la manera más simple de comparar dos instancias es hallar el porcentaje de atributos coincidentes:

$$Sij = \frac{\textit{número de atributos coincidentes}}{\textit{número total de atributos}} \qquad 2.1.3.3.1$$

En inteligencia artificial, ésta es una 'coincidencia parcial': un resultado uno indica que todos los atributos coinciden, mientras que un cero expresa que ninguno de los atributos coinciden.

Tabla 2.1.3.3.1. A 2 x 2 tabla de asociación



Dado que los atributos binarios son muy comunes, los investigadores han tratado, en general, este caso por separado. Si uno mira dos instancias, *i* y *j*, existen cuatro relaciones posibles para cada atributo binario; éstas son mostradas en la tabla de asociación 2 x 2 como Tabla 2.1.3.3.1. Si éstas son totalizadas sobre todos los atributos, *a* y *d* representan el número de atributos coincidentes, mientras que *c* y *b* son no coincidentes. Por lo tanto, la simple medida de coincidencia descripta antes puede ser expresada como:

$$S_{ij} = \frac{a+d}{K} \qquad 2.1.3.3.2$$

donde $K = a + b + c + d$, o el número de atributos. Una medida de distancia puede ser definida como $b + c$: a mayor no coincidencia, mayor es la distancia entre las instancias. Esto se conoce como la *distancia de Hamming* [Hamming, 1980]. La distinción entre *a*, el número de coincidentes positivos, y *d*, el número de coincidentes negativos, es hecha porque los atributos binarios pueden expresar la presencia o ausencia de alguna característica observable. Este es a menudo el caso en biología; en un dominio semejante, podrá ser más apropiado usar una medida que no cuente características 'no' coincidentes:

$$S_{ij} = \frac{a}{a+b+c} \qquad 2.1.3.3.3$$

Esta medida de similitud es conocida como *coeficiente de coincidencia de Jaccard* [Romesburg, 1984, p. 143].

El coeficiente de Jaccard y la simple medida de coincidencia son las medidas de similitud más comúnmente usadas para atributos binarios. Sin embargo, existe un gran número de medidas de similitud relacionadas que pueden ser definidas en los términos de la Tabla 2.1.3.3.1. La mayoría de estas otras medidas están estrechamente relacionadas al coeficiente de Jaccard o a la simple medida de coincidencia; véase Romesburg (1984) para la descripción de 12 diferentes medidas para atributos binarios. Como un ejemplo, la similitud puede ser descripta como la probabilidad de que las instancias *i* y *j* coincidan en una variable menos la probabilidad de que aquellas no coincidan:

$$S_{ij}=\frac{(a+d)-(c+b)}{K} \qquad 2.1.3.3.4$$

Si bien estas medidas están definidas expresamente para atributos binarios, ellas pueden ser fácilmente adaptadas para atributos simbólicos. Esto puede ser llevado a cabo, ya sea convirtiendo un atributo simbólico en un conjunto de atributos binarios o convirtiendo la medida por sí sola. Convertir el atributo puede ser realizado usando los n posibles valores para crear n atributos binarios de presencia-o-ausencia. (Por supuesto, el uso de este tipo de atributos sugiere que el coeficiente de Jaccard es una medida apropiada). Alternativamente, una medida definida por atributos binarios puede ser usada con atributos simbólicos, si uno permite que $a + d$ sea el número de coincidentes, $b + c$ el número de no coincidentes y d el número de veces que el atributo correspondiente está ausente o no sea aplicable al par de instancias dado.[5]

Si bien no es difícil convertir atributos simbólicos a binarios, la única medida que puede comparar instancias con ambos atributos, simbólico y continuo, es una simple combinación de medidas existentes. El coeficiente de Gower [Sokal y Sneath, 1973] es una suma sobre todos los atributos de una de tres medidas: si el atributo es binario, es utilizado el coeficiente de Jaccard; si el atributo es simbólico, se utiliza la medida de coincidencia simple y si el atributo es continuo (u ordinal) se utiliza una métrica de distancia de Manhattan normalizada. A pesar de esta aparente generalidad, Romesburg (1984) señala que la medida raramente ha sido utilizada en la práctica. Esto puede deberse a que las

[5] Anderberg (1973) presenta esta conversión, tanto como un gran número de otras maneras de convertir medidas y atributos de un tipo a otro.

propiedades matemáticas del coeficiente de Gower son desconocidas; por supuesto, esta crítica puede ser hecha contra otras medidas muy usadas.

2.1.3.4 FUNCIONES DE EVALUACIÓN

Las funciones de evaluación se distinguen de las medidas de similitud en que ellas comparan conjuntos de clases, más que pares de instancias. Esta diferencia puede ser trivial; algunas funciones de evaluación son simples extensiones de medidas de similitud. Sin embargo, el énfasis en clases más que en instancias es importante para el aprendizaje automático. Desde esta perspectiva, de la manera en que aprende el sistema, la función de evaluación controla la búsqueda de clases útiles, evaluando la categoría de un conjunto de conceptos con respecto a los datos. En contraste con las medidas de similitud, las funciones de evaluación a menudo se refieren a un tipo particular de definición de conceptos. Como presento diferentes funciones de evaluación, remarcaré sus relaciones con varias medidas de similitud, como así también sus implicaciones para representaciones de conceptos.

2.1.3.4.1. DISTANCIA PROMEDIO

La manera más directa de evaluar clases es evaluar todos los componentes utilizando una función de similitud. Por ejemplo, un método común es obtener el promedio de las distancias de cada objeto a la clase media; cuanto menor es esta distancia promedio, mejor es la clase. Aunque esto podría ser definido con cualquier métrica, la función más común se basa en la distancia Euclídea, sobre todas las clases la función de evaluación es:

$$\textit{traza}\ (\mathbf{W}) = \sum_{j}^{J}\sum_{k}^{K}\frac{1}{N_j}\sum_{i}^{N_j}(x_i - \bar{x}_{jk})^2 \qquad 2.1.3.4.1.1$$

donde N_j es el número de componentes en clase j y $\bar{x}_{jk}$ es el promedio sobre todos los componentes de clase j para atributo k. Esta expresión se conoce como *traza (***W**), pues es la suma a lo largo de la diagonal del grupo interno de la matriz de covarianza. (También conocida como una *matriz de puntos*, esta es la matriz de todas las covarianzas posibles entre K atributos). Nótese que para una única clase y atributo, esta expresión corresponde a la varianza para ese atributo.

De hecho, esta función sugiere una representación de concepto que consiste en una lista de medias y varianzas para cada atributo, puesto que esta información es necesaria para calcular la *traza (***W***).*

Esta función de evaluación es una que pertenece a un conjunto basado en la matriz de identidad, **T** = **W** + **B**, donde **T**, **B** y **W** son el total, entre grupo y matrices de puntos de grupo interno, respectivamente.[6] En general, estas funciones intentan tanto minimizar **W** (una medida de las diferencias de grupo interno) como maximizar **W** (diferencias de entre grupo). Para comparar matrices, ellas deben ser convertidas a un escalar: uno puede utilizar ya sea el determinante de la matriz, o (más económicamente) la 'traza' de la matriz.

Las tres funciones más comunes son minimizar la *traza* (**W**) (definida anteriormente), maximizar la *traza* ($\mathbf{W}^{-1}\,\mathbf{B}$) y minimizar el determinante de **W**.

Es importante notar que la *traza* (**W**) tiene el mismo problema que la distancia Euclideana; es sensible a la normalización de los datos y a la transformación lineal de atributos. También prefiere clusters que formen hiperesferas en el hiperespacio de atributos. Sin embargo, las otras funciones propuestas son computacionalmente mucho más caras, especialmente por el hecho de que ellas necesitan computar la inversa de la matriz **W**. Anderberg (1973, p. 175) cita la evidencia de que no existen buenas razones para seleccionar una función sobre otra y por lo tanto sugiere considerar sólo la *traza* **(W**), la función más simple.

Estas funciones de distancia promedio son definidas para trabajar con instancias descriptas por atributos continuos. Si bien uno puede definir una 'distancia' entre dos instancias con atributos simbólicos, ciertamente no puede haber un significado para cada atributo simbólico. Si los datos incluyen atributos simbólicos, el investigador debe utilizar alguna otra función de evaluación.

2.1.3.4.2. CORRELACION DE ATRIBUTOS

Hanson y Bauer (1989) describen una función de evaluación para atributos simbólicos basada en la correlación entre atributos. Cada clase es definida por todos los pares de tablas de contingencia posibles (como en la Tabla 2.1.3.3.1). Para evaluar la clase, la distribución sobre cada tabla de contingencia, *Dij*, es

promediada para todos los pares de atributos; cuanto mayor es esta correlación de promedios, mejor es la clase. La 'matriz de concurrencia' (para usar la terminología de los autores) se define como:

$$Dij=\frac{\sum_m \sum_n (x_{mn} \log x_{mn})}{(\sum_m \sum_n x_{mn})(\log \sum_m \sum_n x_{mn})} \qquad 2.1.3.4.2.1$$

donde x_{mn} es un conteo en la tabla de contingencia para atributos *i* y *j.* Recordemos que cada posición en una tabla de contingencia cuenta el número de veces que el valor atributo *m* para el atributo *y* co-ocurre con valor *n* para el atributo *j.* El valor promedio de *Dij* es una medida de la cohesión de la clase interna. Los autores rotulan esto como W_c y establecen que la función de evaluación completa es para maximizar W_c / O_c, donde O_c es definida de una manera similar, excepto que la cohesión es medida por medio de las clases. Una vez más, el objetivo general es maximizar similitudes de grupo interno y diferencias entre grupo.

Esta medida de 'correlación', *Dij*, está más estrechamente relacionada a la prueba de independencia chi-cuadrado que al coeficiente de correlación de Pearson para atributos continuos. Por esta razón, al menos, su método es distinto del análisis de factor tipo Q. Sin embargo, Anderberg (1973) muestra que existen peligros en la interpretación de resultados con cualquier medida basada en la correlación de atributos. En particular, la medida de distribución puede no ser consistente a través de diferentes atributos; si *Dij* = .52 y *Dkl* =.41, uno *no puede* concluir que *i* y *j* están más estrechamente relacionados que *k* y *l.*

Otro problema con este método particular es que sólo se ajusta bien a dominios con relativamente pocos atributos binarios. Aunque no existen problemas teóricos con otros tipos de datos, los costos de almacenamiento y, en un grado menor, los costos computacionales, resultan muy rigurosos. Con *K* atributos, cada uno con valores *M*, cada concepto necesita *K (K - 1)* / 2 tablas, cada una de las cuales tiene M^2 entradas.

La ventaja de un método basado en la correlación es que puede fácilmente hallar conceptos que dependen de alguna relación entre dos atributos.[7] Debido a

[6] Véase Hand (1981) para una discusión más detallada de esta identidad, así como también otras referencias para la utilización de estas funciones.

[7] Por supuesto, como se usa aquí, la 'correlación' sólo se refiere a un par de atributos. Correlaciones en Tres-dimensiones o N-dimensiones no son exploradas.

que esta relación está explícitamente representada en las tablas de correlación, el concepto puede ser fácilmente descubierto y representado por el sistema de Hanson y Bauer. Los sistemas que no usan alguna forma de correlación tienen a menudo dificultad con este tipo de clase.

2.1.3.4.3. FUNCIONES BASADAS EN TEORIA DE INFORMACION

Gluck y Corter (1985) presentan una función de evaluación de la teoría de la información', una *categoría de utilidad*, para atributos simbólicos. Esta función se basa en la probabilidad de un valor de atributo, $P(x_{kv})$. Esta probabilidad puede ser expresada como el número de veces que el atributo k ha tenido el valor v, dividido por el número total de instancias. (Nótese que esta probabilidad está estrechamente relacionada con la medida de coincidencia simple). Las medidas de la categoría de utilidad de la información se obtienen particionando instancias en clases.

Para un atributo dado k, es:

$$\textit{Category Utility}(k)=\frac{\sum_{j=1}^{J}[P(C_j)\sum_{v=1}^{V}P(x_{kv}|C_j)^2]-\sum_{v=1}^{V}P(x_{kv})^2}{J} \qquad 2.1.3.4.3.1$$

donde V es el número de valores atributo para el atributo k y J es el número de clases.

$P(x_{kv}|C_j)$ es la probabilidad del valor atributo condicionado por la clase C_j, significando que sólo aquellas instancias en clase Cj son consideradas.

En contraste, $P(x_{kv})$ es aquella probabilidad sin ninguna información de clase; es la información en la clase padre.[8] Si bien la categoría de utilidad se basa en la medida de coincidencia simple, la substracción del término final, permite a la función medir ganancia de información de padres a hijos. Esta ganancia es

[8] Gluck y Corter (1985) definieron categoría de utilidad para dos clases; aquí, he mostrado la generalización de Fisher (1987a) para J clases. El modelo de información teórica también usa logaritmos en lugar de los términos cuadráticos ($P(x)log(P(x))$ en lugar de $P(x)^2$). De cualquier modo, el autor sostiene que esta diferencia no afectará el comportamiento del sistema de clustering.

luego dividida por el número de hijos, por tanto, aquellas particiones de tamaño diferente pueden ser comparadas.

Ambas, la categoría de utilidad y la función de evaluación de Hanson y Bauer, trabajan sólo con atributos simbólicos; debido a que ellas iteran alrededor de todos los posibles valores de cada atributo, ellas no pueden ser aplicadas a atributos continuos. Las clases se definen como un conjunto de probabilidades para cada par posible atributo-valor. Gennari, Langley y Fisher (1989) usan la categoría de utilidad como la base para una medida relacionada para atributos continuos normalmente distribuídos. Debido a que asumimos una distribución normal, esta medida se basa en la desviación estándar, σ_k, para un atributo dado *k*.

La función de evaluación utilizada por el sistema CLASSIT [Gennari et al.] es:

$$\frac{\sum_{j=1}^{J} P(C_j) / \sigma_{jk} - 1 / \sigma_{pk}}{J} \qquad 2.1.3.4.3.2$$

donde σ_{jk} es la desviación estándar dentro de una clase *j* dada y σ_{pk} es la desviación estándar sin información de ninguna clase.

Como uno podría suponer, esta medida es similar a minimizar la *traza* (**W**). Por ejemplo, las clases son definidas de la misma manera, un conjunto de medias y desviaciones estándar, para cada atributo. De hecho, la única diferencia importante es la substracción de $1/\sigma_{pk..}$ Sin embargo, esto es exactamente lo que distingue la categoría de utilidad de Gluck y Corter de otras funciones de evaluación. Adicionalmente, existe alguna fuerte evidencia de que sustrayendo esta información de contexto se reduce la sensitividad de la medida a normalizar o transformaciones lineales de atributos, un problema definido para la función *traza* (**W**).[9]

2.1.3.4.4. EVALUACION DE LA CLASE BAYESIANA

Si bien un sistema Bayesiano de clustering compara clases, usualmente no tiene el mismo tipo de función de evaluación como aquellos descriptos hasta aquí. En lugar de evaluar clases con respecto a todas las instancias, la ecuación básica de la teoría de la decisión (presentada en la sección 2.1.2.4) compara una única instancia con un conjunto de clases. La dificultad con esta ecuación es que para calcular las probabilidades de clase condicional, $P(x|\omega_i)$, se necesita una estimación de los parámetros de clase que definen cada ω_i. Duda y Hart (1973) concluyen que en general no existe una manera analíticamente simple para hallar esta estimación y que los costos computacionales para una solución exacta crecen exponencialmente con el número de instancias.

Sin embargo, existe un número de técnicas de estimación que han sido empíricamente exitosas. Fried y Holyoke, 1984, utilizan un algoritmo simple basado en la medida de similitud de la distancia Euclídea para hallar estimaciones de parámetros iniciales de clases. Utilizando estas estimaciones, ellas pueden entonces determinar $P(x|\omega_i)$. Anderson usa una *probabilidad*

[9] Este efecto fue observado por el autor mientras experimentaba con el sistema CLASSIT.

conjunta [Anderson, 1988], un parámetro definido por el usuario, para definir la probabilidad prioritaria de cada clase y la probabilidad de que una instancia sea componente de una nueva clase. Esto le permite obtener una partición inicial de instancias en clases. Entonces, para una nueva instancia, puede calcular $P(x|\omega_j)$ basado en los componentes de cada ω_j.[10]

Desafortunadamente, es difícil comparar estos sistemas de "funciones de evaluación" con otros. Parte de la dificultad es que la tarea es descripta como parámetros de clase estimados (o actualizados) más bien que evaluar la categoría de las definiciones de clases. Por ejemplo, Cheeseman et al. (1988) describe el algoritmo de actualización para su sistema Bayesiano pero no da una ecuación para este proceso de evaluación ni describe cómo se compara con otras funciones de evaluación de clustering.

2.1.4. ALGORITMOS PARA CLUSTERING

Cualquier técnica de clustering puede ser descripta como un algoritmo que usa algún tipo de medida de similitud o función de evaluación para buscar un conjunto de clases. Hasta aquí, he descrito estas medidas fuera de contexto; esta sección mostrará cómo ellas son usadas por algoritmos. Mi objetivo en presentar estos componentes por separado es mostrar que cualquier combinación de algoritmo y medida define una técnica de clustering. En cierto grado, se puede elegir una parte del sector 'a' y una parte del sector 'b' para crear un método de clustering. Por supuesto, muy pocos investigadores han tratado de hacer esto y no cualquier tuerca encajará en cualquier tornillo. Para cada algoritmo señalaré la medida de similitud original propuesta así como otras que podrían ser usadas.

Si bien existen varias maneras de organizar algoritmos, la más original se basa en su aproximación al problema de clustering. He elegido dividir algoritmos en tres grupos: algoritmos *aglomerativos*, algoritmos de *optimización iterativa* y algoritmos *incrementales*. La aproximación aglomerativa es la más antigua,

[10] Debido a que Anderson trabaja con atributos simbólicos, su función de evaluación está relacionada con la medida de coincidencia simple, excepto que la nueva instancia sea comparada con el conjunto de todas las instancias componentes. El trabajo de Anderson es también interesante pues su algoritmo es *incremental* (véase sección 2.1.4.3)

habiendo sido propuesto por los trabajadores en biología y ecología. Con el advenimiento de la computadora, los métodos de optimización iterativa se volvieron populares como una aproximación heurística más eficiente al clustering. Finalmente, los algoritmos incrementales fueron inspirados por la formación de concepto humana y fueron creados por investigadores en aprendizaje automático.

Además de describir algunos de los más importantes algoritmos, consideraré dos características de cada método. Primero, si bien cualquier algoritmo debe producir algún conjunto de clases como salida, la forma y organización de las clases es dictada por los objetivos del investigador. Por ejemplo, el investigador puede preferir una única lista de clases o puede necesitar una jerarquía de clases específica a general. Asimismo, el investigador puede preferir que cada instancia sea asignada a una única clase o a más de una clase o aún a todas las clases probabilísticamente. Segundo, algoritmos diferentes tienen costos de memoria y computacionales muy diferentes. La computadora no puede ser tratada como una máquina infinitamente potente. Especialmente desde el punto de vista de aprendizaje automático, es importante que el algoritmo y la medida de similitud sea tan barata como sea posible.

2.1.4.1 MÉTODOS AGLOMERATIVOS

Históricamente, los primeros algoritmos para clustering fueron métodos aglomerativos. Desde el momento que ellos fueron desarrollados por biólogos, produjeron una jerarquía (una taxonomía) de clases, desde la clase más general (incluyendo todas las instancias) a las clases más específicas (cubriendo sólo una instancia). Si bien éstos son aún los algoritmos más largamente usados, son caros en requerimientos tanto de espacio como de tiempo.

Un método aglomerativo empieza con cada instancia como una clase separada y repetidamente combina estas más pequeñas clases, clases específicas para formar clases más grandes y más generales. Este proceso construye una jerarquía de clases, finalizando cuando todas las instancias han sido aglomeradas en una clase de primer nivel.[11] Para determinar cuáles instancias

[11] Lo contrario a esta aproximación se conoce como un algoritmo *divisivo*. Esto comienza por asumir que cada instancia está en la misma clase de nivel más alto, entonces divide

'aglomeran', éstos algoritmos requieren una *matriz de similitud* que muestre cuan cerca, de acuerdo a algunas medidas de similitud, cada instancia está de otra instancia. Dada esta matriz, un algoritmo general aglomerativo puede describirse como sigue:

1. **Computar y almacenar la matriz de similitud.**
2. **Hallar el menor (y mejor) valor en la matriz y su par de instancia asociada.**
3. **Combinar estas dos instancias (o clases) en una clase más grande.**
4. **Crear una nueva matriz de similitud que incluya la nueva clase. (Esto puede requerir volver a calcular algunos valores).**
5. **Si existe sólo una clase, retornar a la jerarquía producida y detenerse. O si no, ir al paso 2.**

Esta descripción no clarifica exactamente cómo llevar a cabo el paso 4. Para ingresar una nueva clase en una matriz de similitud, debe decidirse cómo usar la medida de similitud para comparar una clase con una instancia o con otras clases. La solución más simple (o al menos más económica) es definir la similitud entre dos clases, *A* y *B*, como la similitud de las dos instancias más cercanas *a* y *b*, donde a $\in$ *A* y *b* $\in$ *B*. Esto se llama el "vecino más próximo" o algoritmo de "ligamiento simple" pues dos clases se combinan por el único ligamiento más corto entre ellas. Este algoritmo tiende a producir largas cadenas de clusters; dependiendo del dominio, esto puede ser una desventaja.

Otros algoritmos aglomerativos utilizan maneras diferentes para medir la 'similitud' entre dos clases. Por ejemplo, en lugar de un ligamiento simple, se puede encontrar la distancia de "ligamiento promedio" entre clases. Esto requiere calcular la clase media para cada atributo cada vez que una clase es creada o expandida y entonces medir la distancia a otros grupos o instancias desde esta media. Este método evita crear cadenas largas y, en su lugar, prefiere clusters que forman hiperesferas en el espacio instanciado.

repetidamente esta clase en un número de hijos hasta que cada (muy específica) clase tiene sólo una instancia. Si bien unos pocos algoritmos semejantes han sido propuestos (MacNaughton-Smith et al. 1964, Fisher 1984), ellos han sido raramente usados.

Desafortunadamente, este método es relativamente caro: al requerir volver a calcular parte de la matriz de similitud, tiene un costo computacional de $O(n^3)$, donde n es el número de instancias. El algoritmo del vecino más próximo puede implementarse de manera que no se tenga que volver a calcular la matriz de similitud pero el costo computacional es aún $O(n^2)$ y requiere una matriz de similitud ordenada. Dependiendo de la medida de similitud y del número de atributos por instancia, calcular y almacenar la matriz de similitud solamente puede ser prohibitivamente caro. De hecho, para dominios con un gran número de atributos e instancias, los métodos aglomerativos no son convenientes.

Otro problema significativo con algoritmos aglomerativos es que ellos producen sólo jerarquías *binarias*. En lugar de esta estructura, el investigador está frecuentemente interesado en hallar algún número 'óptimo' de clases. Aunque existen algunos métodos para cortar o allanar la jerarquía binaria [Aldenderfer y Blashfield, 1984], éstos no garantizan encontrar los mejores o más apropiados conjuntos de clases. De hecho, Everitt (1979) señala que tales técnicas pueden ser erróneas; concluye en que encontrar el número óptimo de clases de la jerarquía binaria es un problema abierto.

2.1.4.2 OPTIMIZACIÓN ITERATIVA

Los algoritmos de *optimización iterativa* fueron creados en respuesta a lo costoso de los métodos aglomerativos y a la dificultad de hallar el número correcto de clases. En lugar de tratar de encontrar este número ideal, se puede asumir que el número de clases k está dado por el sistema de clustering. La optimización iterativa busca estas clases k de instancias por iteración reasignando instancias a diferentes clases para mejorar el valor resultante de alguna función de evaluación. Si bien estos algoritmos no producen jerarquía de clases, son más eficientes que los métodos aglomerativos y al transferir instancias de clase a clase ellos pueden recuperarse de una 'mala' decisión inicial.[12]

[12] Duda y Hart (1973) dan la expresión exacta para el número de maneras de particionar n instancias en c clases; una aproximación es $c^n/c!$.

En general, se puede ver el problema de clustering como una búsqueda sobre el enorme espacio de posibles particiones de las instancias en clases. Un simple método examinaría cada posible partición y encontraría aquella con el mejor resultado de acuerdo a una función de evaluación. Desafortunadamente, esto es computacionalmente imposible aún con un relativamente pequeño número de instancias; por ejemplo, existen 5.28×10^{28} maneras de particionar 50 instancias en cuatro clases.[12] Por lo tanto, en lugar de una completa búsqueda a través de este espacio, los métodos de optimización iterativa usan técnicas de hill-climbing para implementar iterativamente el resultado de evaluación hasta que se alcanza un óptimo. Como con cualquier método de hill-climbing, el punto inicial para la búsqueda puede ser crítico y el algoritmo puede converger en un óptimo local en lugar del óptimo global.

Teóricamente, podemos usar cualquier función de evaluación como el criterio a optimizar en cada iteración. Sin embargo, para mantener la eficiencia del sistema completo de clustering, el investigador debería elegir una función de evaluación relativamente simple, una tal que aquél sistema pueda calcular económicamente tal como considera cada reasignación.

Por ejemplo, uno de los más simples y más populares es el algoritmo *k*:-means:

1. **Usar las primeras instancias *k* como puntos de orígen.**
2. **Asignar cada una de las restantes instancias a la clase representada por el punto de origen más cercano (distancia Euclideana).**
3. **Volver a calcular nuevos puntos de orígen tantos como los centroides (el promedio de los valores atributo) de cada clase.**
4. **Iterar entre los pasos 2 y 3 hasta que no se realice ningún reasignamiento.**

Si bien el número de iteraciones requeridas antes de detenerse es desconocida, Anderberg (1973) da una prueba de que tales algoritmos convergirán eventualmente y en la práctica esto es usualmente un número razonablemente

pequeño (menor que 10). Cuando se usa la distancia Euclídea, Hand (1981) muestra que este algoritmo es equivalente a optimizar la función de evaluación *traza* (**W**).

Se puede hacer un número de modificaciones a este algoritmo. Primero, dado que el punto inicial puede ser crítico a un investigador de hill-climbing, se pueden utilizar diferentes métodos para elegirlo. Por ejemplo, las instancias iniciales *k* pueden ser elegidas al azar o ellas pueden ser elegidas tal que todos los orígenes estén separados al menos en alguna mínima distancia. El algoritmo completo puede repetirse con diferentes selecciones de orígenes tal que el investigador puede comparar posibilidades [Duda y Hart 1973]. De hecho, ellos sugieren incluso utilizar aún un método aglomerativo para hallar la partición inicial, aunque esto parece caro. Anderberg (1973) también describe un número de técnicas de selección de orígenes.

Una segunda modificación puede hacerse calculando nuevos centroides de clases siempre que una instancia sea reasignada a una clase. En este caso, el algoritmo puede converger mucho más temprano; por ejemplo, el algoritmo *k*-means de MacQueen's (1967) [Sokal y Sneath, 1973] utiliza sólo dos pasadas a través de las instancias. En la primera pasada, los centroides son modificados a medida que sucede cada reasignación; durante la segunda pasada los centroides permanecen fijos.

Finalmente, se puede expandir el algoritmo de manera que apunte a dos problemas de nivel superior. Primero, porque no siempre se puede ser capaz de especificar el número de clases, *k, a priori,* se puede tratar el algoritmo *k*-means con diferentes valores de *k* permitiendo una estimación del 'mejor' *k*. Segundo, el investigador puede necesitar una jerarquía de clases más que la sencilla lista que los algoritmos de optmimización iterativa usualmente producen. Para este fin, se puede ejecutar simplemente el algoritmo recursivamente en cada uno de los *k* grupos identificados en esta primera ejecución.

Si bien estas extensiones parecen ser caras, soluciones 'bruta-forza', Michalski y Stepp (1983b) las han incorporado en algoritmos *k* -means en su programa CLUSTER/2.[13] Este sistema también incluye un paso que ayuda a evitar la

[13] Cluster/2 no es usualmente identificado como usuario del algoritmo de clustering de optimización iterativa. Esta caracterización se clarificó sólo después de estudiar estos métodos más antiguos.

óptima local: cuando el algoritmo *k* -means converge, CLUSTER/2 elige nuevos puntos de orígen en el extremo de cada clase, en lugar de hacerlo en los centroides. Entonces reinicia al algoritmo *k* -means con estos nuevos orígenes, y si la partición resultante permanece invariante, el sistema retorna aquellas clases como la solución. Finalmente, CLUSTER/2 difiere de la mayoría de las técnicas de optimización iterativa en que utiliza dominios con atributos simbólicos más que con atributos continuos. Como la función de evaluación *traza* (**W**) sólo trabaja con dominios continuos, este sistema usa una especial función de evaluación de propósito específico definida por un conjunto de parámetros (de uso específico). Desafortunadamente, la optimización iterativa no ha sido tratada con otras funciones de evaluación principales para atributos simbólicos.

También se usa una aproximación de optimización iterativa, para el sistema AUTOCLASS [Cheeseman et al.,1988], no particiona estrictamente instancias en clases. Este sistema conduce a una búsqueda de hill-climbing para el mejor conjunto de clases, donde cada clase está definida por una media y una desviación estándar. Sin embargo, la clase componente es completamente probabilística, especificando la probabilidad de que $x \in c$, para cualquier objeto x y cualquier clase c. Estas clases 'borrosas' son especialmente apropiadas cuando una instancia es descripta razonablemente bien por dos (o más) clases que compiten.

Si bien los algoritmos de optimización iterativa obtuvieron algún éxito, su dominio de aplicación está algo restringido. Si es aceptable una aproximación de *k,* si las instancias son descriptas por atributos continuos y si una sencilla lista de clases es suficiente, entonces la optimización iterativa es una buena aproximación al problema de clustering. Si bien el sistema CLUSTER/2 alivia estas restricciones, lo hace sólo al costo computacional esencial.

2.1.4.3 MÉTODOS INCREMENTALES

En contraste con otras aproximaciones, los algoritmos incrementales ven el conjunto de instancias como una secuencia potencialmente infinita. A medida que cada instancia es procesada, el algoritmo hace una modificación incremental a su conjunto corriente de conceptos. En cualquier instante de

tiempo, los conceptos reflejan información ganada de todas las instancias encontradas hasta ese instante. Esta aproximación fue diseñada con el sistema humano de clustering en mente; parece inverosímil que un aprendiz humano necesitaría recibir primero algunos números de instancias y entonces dejar de recibir y ejecutar el cómputo de la tarea de clustering.

Aún si uno no sigue este camino, estos sistemas incrementales son útiles por un número de razones más pragmáticas. Ellos requieren menos tiempo de cómputo que otros algoritmos, y por lo tanto pueden procesar bases de datos más grandes. Estos algoritmos evitan también el problema de seleccionar el número de clases. Finalmente, este tipo de algoritmo es casi esencial para cualquier aplicación donde las definiciones de clase son dinámicas. Schlimmer y Granger (1986) se refieren a esto como *tendencia conceptual* si nuevas instancias reflejan nuevos o diferentes conceptos, un algoritmo incremental puede ajustar sus definiciones de concepto fuera de tiempo.

Un algoritmo incremental general para sumar cada nueva instancia *x* a una jerarquía de clases puede describirse como sigue:

1. Incorpore x al nodo raíz.
2. Incorpore x a una clase de hijos existente o
b) Cree una nueva clase de hijos basada en x.
3. Repita (si desea) en cada clase de hijos.

Usualmente estos algoritmos producen una jerarquía de clases, pero el paso tres puede omitirse si se prefiere una simple lista de clases. A diferencia de los métodos aglomerativos, los algoritmos incrementales no producen jerarquías binarias: el factor de ramificación es variable y determinado por la frecuencia con que son creadas nuevas clases (en el paso 2b). Determinar cuándo crear una nueva clase resulta crítico para estos algoritmos, esta elección le permite a los algoritmos incrementales encontrar automaticamente un número apropiado de clases de los datos.

EPAM [Feigenbaum, 1963] es uno de los primeros sistemas en inteligencia artificial en aproximarse al problema del clustering. Este sistema aplica la

decisión *monotética* efectuada sobre el algoritmo básico incremental. El sistema asocia cada nivel en la jerarquía con un único atributo. Para determinar cuál acción ejecutar en el paso 2 (anterior), EPAM inspecciona el valor de ese atributo y crea una nueva disyunción si ese valor no coincide con cualquiera de los hijos existentes. De otra forma, el sistema ordena la instancia a la clase de hijos con un valor atributo coincidente.

En contraste, UNIMEM [Lebowitz, 1985, 1986] usa una estrategia *politética*. En cada nivel, el sistema inspecciona algún subconjunto de atributos y entonces utiliza estos valores en conjunción con una función de evaluación para elegir una clase. Este algoritmo también va detrás de una estricta partición para permitir *saltear* u ordenar instancias para más de una clase. Sin embargo, UNIMEM no permite la clasificación completamente probabilística sugerida por [Cheeseman et al., 1988].

Como se definió anteriormente, el algoritmo incremental es un hill-climber puro, puede quedar atrapado en la misma forma de óptima local como los métodos de optimización iterativa. El sistema Cobweb de [Fisher, 1987a, 1987b] agregó algunos operadores al algoritmo que fueron diseñados para atenuar este problema. Además, a las opciones a) y b) en el paso 2, el sistema considera combinar dos clases existentes o dividir una clase en sus hijos. Estos operadores le permiten al sistema salir de la óptima local, puesto que permiten una forma de retroceso a través del espacio de posibles jerarquías de concepto.

Como se describe anteriormente, el aprendizaje automático al aproximarse al clustering prefiere crear definiciones de clase por intensión. Los algoritmos incrementales modifican estas definiciones con cada nueva instancia, más que agregar la instancia a una lista de componentes. Esto sugiere que un sistema incremental no necesita almacenar cada instancia. Esta habilidad conduce a costos de memoria reducidos, especialmente si el número de instancias es muy grande. Formas de desarrollo para 'olvidar' instancias o evitar almacenarlas, en definitiva, es un tópico de investigación corriente. [Gennari et al., 1989].

2.1.5. EVALUANDO UN MÉTODO DE CLUSTERING

Con estos varios métodos de clustering para elegir, uno esperaría algunas maneras de postulados de comparación de los distintos métodos. En general, desearíamos tener una medida cuantitativa para ver qué resultado ha tenido un método dado. Desafortunadamente, debido a que los objetivos de clustering varían y están con frecuencia muy pobremente definidos, una única medida definitiva es imposible. En cierto grado, cualquier comparación cuantitativa de resultados es imposible porque el objetivo de clustering es subjetivo. Por ejemplo, algunos investigadores prefieren un método de clustering a otro simplemente porque produce un 'mejor', 'más intuitivo' o 'más placentero' conjunto de clases. Sin embargo, los investigadores con una trayectoria más formal han hallado un número de maneras para medir diferentes aspectos de una solución.

Una medida intuitiva utilizada en análisis discriminante y en aprendizaje automático es ver cómo a menudo un sistema clasifica instancias correctamente. En estos dominios la clase correcta es conocida de antemano, de manera tal que la jerarquía producida por algunos números de instancias de 'prueba' puede ser evaluada según lo bien que clasifique un número de instancias "prueba". Para los métodos de clustering, el problema es que la clase 'correcta' no es conocida. Mejor aún, esta medida puede ser utilizada para comparar un conjunto 'intuitivo' de clases con el nuevo conjunto producido a partir de los datos; esto usualmente *no* es una validación del proceso que creó nuevas clases.

Una medida que puede ser utilizada para comparar jerarquías es la "correlación cofenética". Este método compara la salida de algoritmos aglomerativos y es por lo tanto utilizado mayormente por biólogos. Esta medida utiliza la correlación para comparar la matriz de similitud original con una nueva matriz que es derivada de la jerarquía final. Una entrada (*i, j*) en la nueva matriz tiene un valor igual a la medida de similitud cuando las instancias *i* y *j* fueron juntadas en la jerarquía. Estas dos matrices pueden ser comparadas entonces calculando su correlación - una alta correlación entre entradas sugiere que el método clustering hizo un buen trabajo de captura de información en la matriz de similitud original. Desafortunadamente, Aldenderfer y Blashfield (1984) señalaron que esta medida no es estadísticamente conocida. El uso de la correlación implica que los

valores en las matrices son normalmente distribuidos. Dado que la matriz derivada depende de la matriz original y contiene mucho menos información, este no es usualmente el caso.

Como una alternativa, Aldenderfer y Blashfield sugieren que un camino mejor para comparar jerarquías de concepto es utilizar lo que ellos llaman método de validación 'Monte Carlo'. Existen tres pasos en este procedimiento. Primero, se genera un conjunto de datos al azar que es normalmente distribuido y basado en los promedios de los datos originales. En efecto, esto representa las instancias originales pero agrupadas como una única clase. Seguidamente, este conjunto de datos al azar es clasificado, resultando en una jerarquía "inicial-básica".

Finalmente, la jerarquía original es comparada a esta jerarquía inicial-básica (por ejemplo, comparando un análisis de varianza dentro de cada clase). Una gran diferencia entre jerarquías significa que el algoritmo ha hecho un buen trabajo en hallar clases a partir de los datos. Si bien el resultado derivado de este procedimiento tiene un significado absoluto pequeño, puede ser usado para comparar un conjunto de diferentes métodos, el método con la mayor diferencia es el 'mejor' para un conjunto de datos dado.

Más que comparar jerarquías, existe una manera más general de fijar el rendimiento de un método clustering. En lugar de evaluar si una clasificación es 'correcta', la idea es juzgar cuan 'útil' puede ser esa clasificación. Aunque esto pueda parecer difícil de hacer, una medida de la 'utilidad' es ver cuan bien las clases pueden predecir valores de atributos de una nueva instancia. Esta habilidad de predicción puede estar relacionada a la tarea de 'invocación' en psicología cognoscitiva y ha sido utilizada en aprendizaje automático [Fisher, 1987; Gennari et al., 1989]. La tarea de invocación dice que dado un conjunto de 'colas' (atributos) de una nueva instancia, el agente debería ser capaz de utilizar su memoria de instancias pasadas para invocar los atributos no especificados de la instancia.

Para usar la predicción como una tarea a desempeñar para evaluar un sistema de clustering, un atributo al azar es omitido de una instancia de prueba. Luego que el sistema clasifica esta instancia, la clase elegida es utilizada para predecir el valor del atributo omitido. Además de depender de la bondad de las clases, la exactitud de esta predicción depende de la instancia del conjunto de prueba (cuan típica es esa instancia) y el atributo omitido (cuan consistente es ese

atributo). Por lo tanto, una vez que este resultado es promediado sobre instancias y atributos, esta precisión predictiva promedio puede utilizarse como una medida general de la utilidad de las clases creadas por el sistema de clustering. Esta medida de rendimiento puede entonces ser utilizada para comparar diferentes métodos con los mismos datos o el mismo método con diferentes conjuntos de datos.

2.2. ESTADO DEL ARTE EN BASES DE DATOS

El diseño de Bases de Datos se realiza en tres fases: Diseño Conceptual de más alto nivel de abstracción, Diseño Lógico que permite convertir al Diseño Conceptual en requerimientos implantados en un sistema de computación y Diseño Físico donde se estructuran los datos y se determinan métodos de consulta.

Según los últimos avances teóricos el diseño lógico se puede desarrollar independientemente del administrador de la base de datos (DBMS Data Base Management System) [Batini,C., Ceri,E., y Navathe,S., 1992] y en los métodos del paradigma orientado a objetos los modelos de datos semánticos y las cadenas semánticas [Kim,W., Lochovsky,F., 1989], así mismo se utilizaron los modelos de entidad-relación ER [Chen,P.P.,1976] y [Bachman,C.V., 1974] y sus extensiones [Teorey, T.J., Fry, J.P., 1982], [Shan y Shixuan, 1984] y [Elmasri,R., Navathe,S., 1989] sobre todo para la discusión sobre tipos abstractos, reglas de integridad referencial y clases.

En esta discusión están involucrados la integración de las tecnologías de bases de datos, con la tecnología de los lenguajes de programación, con la tecnología de la inteligencia artificial y con los lenguajes lógicos.

De todas maneras hay que marcar una tendencia común a aumentar el poder expresivo de los modelos de datos y de los lenguajes de gestión de datos.

2.2.1. EXTENSIÓN DE LOS SISTEMAS RELACIONALES

Hay una tendencia a la extensión de los DBMG relacionales con varias funciones como la de representar directamente objetos complejos en el modelo relacional anidado (capas de software por niveles de abstracción), disparadores de

procesadores de interfaces (triggers), sistema de enlace automático cuando el sistema alcanza condiciones específicas relativas a los datos. [Roth,M.A. et al, 1988], [Shek,H.J., Scholl,M.H. 1986] y [Stonebraker,M. et al, 1990] (sistema Postgres).

2.2.1.1. BASES DE DATOS DINAMICAS

Teniendo en cuenta los trabajos que he presentado en los últimos años con colaboradores [Perichinsky,G., 1994], y el panorama desarrollado en este trabajo, he demostrado que ha surgido un nuevo enfoque en la Investigación en Bases de Datos [de Miguel,A., 1993] [de Miguel,A., 2000], que se apoya sobre una Base de Datos en la cual los datos se almacenan una sola vez, con independencia de su tratamiento, en sistemas orientados hacia los datos y se ha estabilizado conceptualmente, en los Modelos Relacionales estructural y semánticamente y el program embedding SQL [ISO 9075] que son las formas estandarizadas de los próximos años. El modelo trata a los dominios en forma independiente, por lo tanto la estructuración es más natural por la forma de agregación de las tuplas y una semántica n-aria de atributos. Con el predicado de operaciones conjuntistas se forman tablas virtuales o visiones. La arquitectura que se experimenta para el gerenciador es la propuesta de tres niveles [ANSI/X3/SPARC].

2.2.1.2. CONCEPTOS

Esta propuesta consiste en el desarrollo teórico-conceptual y de implementación de un sistema de base de datos relacional estructurado sobre dominios dinámicos de atributos [Perichinsky,G., 1989-1998].

El nuevo enfoque se trata de un mayor nivel de abstracción para conseguir el máximo de independencia lógica posible [(ANSI).1988], que es en la cual el gerenciador tiene la capacidad de que las referencias a los datos almacenados, especialmente en las aplicaciones y en sus descripciones de los datos, estén aislados de los cambios y de los diferentes usos en el entorno de los datos, como pueden ser la forma de almacenar dichos datos, el modo de compartirlos con otras aplicaciones y cómo se reorganizan para mejorar el rendimiento del

sistema de base de datos. Este nuevo enfoque hace que los dominios de los atributos sean dinámicos y conformen realmente la base del tratamiento de la teoría de conjuntos, y que las tuplas se generen dinámicamente a través de visiones.

Se trata de un diseño que cambia la estructuración tradicional de agrupamiento estático de valores de atributos en registros por la creación de dominios de atributos, formando conjuntos de valores de los mismos. Las tuplas (virtuales) se forman mediante las relaciones que como las visiones no existen, las tablas están establecidas a partir de dominios [Date,C.J., 1992.Data on Databases].

No se trata de una estandarización pues este concepto frena los desarrollos futuros, esto exige cautela pues el dilema es, que la estandarización orienta a los diseñadores.

La arquitectura ANSI/X3/SPARC tiene una técnica de diseño por niveles o máquinas anidadas y el flujo de datos pasa por las distintas capas, que están separadas por interfaces, cuyo número marca de alguna manera la capacidad de independencia.

Un diccionario de datos permite la estructuración del conjunto de datos o metadatos.

El tradicional esquema conceptual envuelve esta capa con una interfaz con el diccionario en la metabase de datos.

Las estructuras externas e internas forman parte de capas separadas por las interfaces 4 y 5. Cada uno tiene un administrador en la interfaz 3 se puede tener un conjunto de menús, que se utilizará para el administrador de la Base.

La manipulación de la Base de Datos se hará con SQL embbeding con un motor C y C++.

Una operación se ejecuta mediante transformadores conceptual/externo, interno/conceptual y almacenamiento/interno que utilizan los metadatos mediante las interfaces (binding).

Como no se especifica la instrumentación, se permite que:

- **el sistema evolucione rápidamente**
- **la utilización sea óptima**
- **se logre independencia lógica**
- **posibles reestructuraciones.**

En forma similar que en el SQL, las tablas son, en sentido estricto y dinámico, "visiones" con leyes de formación que surgen de la lógica de la aplicación.
Por ello hablamos de visiones y decimos "dinámicas", pues se pueden agregar o eliminar columnas-dominio de una tabla virtual y por supuesto se pueden modificar y eliminar valores de atributos, tanto como aumentar el cardinal de un dominio. De esta manera todo objeto o entidad de una aplicación puede mejorar dinámicamente su calificación e identificación. Se alcanza así una gran independencia tanto física como lógica de los datos, y una dinámica en el crecimiento o expansión (hasta en comportamiento) [Perichinsky,G., 1994].

2.2.1.3. MODELIZACIÓN

El modelo surge del par M=<S,O> donde S son las reglas y O las operaciones sobre objetos permitidos.
Las instancias determinan la dinámica del modelo.
El debate actual es extender la Bases de Datos Relacionales hacia la orientación a objetos [Third Generation Database System Manifesto. Carey et al.1990] y los puristas del modelo orientado a objetos [The Object-Oriented System Manifesto. Atkinson et al.1990], después [Staugaard, 1998].
Ante esta alternativa es preferible pensar en el avance teórico que representa la orientación a objetos y por lo tanto aplicarla de acuerdo a los requerimientos del diseño; por ejemplo una capa de nivel externo que dé la apariencia de objetos, sobre un modelo relacional [de Miguel,A., 2000].
Verdaderamente se obtienen grandes ventajas con este modelo, ya que lo expuesto implica una reestructuración que no depende de los datos sino de las aplicaciones. Las diferentes aplicaciones pueden "ver" a los datos de acuerdo a sus requerimientos y modos.
Se simplifica la visión de los usuarios. Los dominios de los atributos son la base del modelo y de la agregación lógica de estos, mediante operaciones y formas algebraicas relacionales, surge la estructura.
Las expresiones del álgebra relacional sirven a los propósitos de mantenimiento, actualización y recuperación de la información de los dominios, a través del manejo de tuplas, y preservando su homogeneidad e integridad. Al operar sobre

dominios, conjuntos de valores de atributos, se tiene la formalización concreta de la teoría de conjuntos.

Los valores de los dominios son atómicos y los atributos se califican por aplicación; esto permite operar en formas normales.

La capa externa propuesta anteriormente para tener la apariencia de objetos sería una aplicación más desde el modelo conceptual.

Lo mismo ocurre con una herramienta CASE Inteligente, o una capa de nivel inteligente para aprendizaje automático.

La TERCERA generación de Bases de Datos tendrá más que ver con nuevas capas de nivel de contenido semántico para tener en cuenta las Bases de Datos Distribuidas, Inteligentes y con Orientación a Objetos en un entorno abierto, heterogéneo y distribuido [Informe Lagunita, Silberschatz,A., 1990].

Las FAMILIAS de modelos que surgen a partir del modelo de Chen de entidad-relación ER [Chen,P.P., 1976] y [Bachman,C.V., 1974] y sus extensiones [Teorey,T.J., Fry,J.P., 1982], [Shan y Shixuan, 1984] y [Elmasri,R., Navathe,S., 1989], [Luque Ruiz, I., Gómez-Nieto, M. A., et al. 2002] tratan los tipos de entidades y sus relaciones, a partir de la estructuración estática de los atributos, asociadas a un predicado lógico.

2.2.1.4. DINÁMICA

El contenido semántico de las relaciones se ha completado conceptualmente con la cardinalidad, la dependencia en existencia y en identificación y la abstracción de generalización.

La cardinalidad tiene que ver con la cantidad de instancias que tienen las entidades relacionadas algebraicamente, y sus correspondencias según los axiomas de Armstrong [Wiederhold,G. 1983].

2.2.2. SISTEMAS DE GESTIÓN DE BASES DE DATOS ORIENTADAS A OBJETOS

Estos sistemas integran el paradigma de programación orientada a objetos y la ingeniería de software. Aún cuando no hay un fundamento teórico consolidado

para los lenguajes y los modelos la tendencia es a la utilización en desarrollos industriales.

2.2.3. SISTEMAS DE GESTIÓN DE BASES DE DATOS DEDUCTIVAS

Estos sistemas integran las tecnologías de las bases de datos con la programación lógica. La principal característica es que incluyen mecanismos de inferencia, basados en reglas, que generan información adicional a partir de los datos que tienen almacenados, son trabajos de laboratorios con poca incidencia industrial y de aplicaciones importantes [Bertino,E. y Mondesi,E. 1992] y [Cacase,F. et al, 1990].

2.2.4. SISTEMAS DE GESTIÓN DE BASES DE DATOS INTELIGENTES

Estos sistemas extienden las tecnologías de bases de datos incorporando paradigmas y técnicas desarrolladas en el campo de la inteligencia artificial como CLASSIC [Borgida,A. et al, 1989] y ADKMS [Bertino,E. et al, 1992].

2.2.5. PANORAMA

En el modelo relacional existe una estandarización y una formalización a partir de desarrollos de los artículos de la ACM.13(6),377-387 [Codd,E., 1970]. No existen puntos de referencias en modelos orientados a objetos ni estándares.
La evolución del paradigma de la programación orientada a objetos deriva del lenguaje Simula [Dahl,O.J., Nygaard, K., 1966] y la aparición de lenguajes como el Smalltalk [Goldberg,A., Robson,D., 1983], desde el lenguaje C al C++ [Stroustrup,B., 1986], el Clos [Moon,D.A., 1989] y el Eiffel [Meyer,B., 1988], donde los programas están formados por objetos independientes formando clases y que se comunican entre sí mediante mensajes, similar a los lenguajes basados en el conocimiento con diferentes interpretaciones [Fikes,R., Kehler,T., 1985], pero las bases de datos requieren un modelo de datos propio y si no existe un estándar se agrupan en un **modelo básico o central** (core) [Kim,W., 1990].

El modelo central es poderoso pero en orientación a objetos no captura restricciones de integridad y por lo tanto de dominio y las relaciones semánticas demasiado importantes para muchos tipos de aplicaciones, tales como la unicidad de los valores de un atributo, valor nulo para un atributo y el rango de valores que un atributo puede asumir y las relaciones semánticas del tipo “parte de” entre pares y asociaciones de objetos. Esto surge porque son características de las bases de datos y no de los lenguajes de programación que se pueden detallar como:

Objetos y entidades del mundo real con un estado (valores de los **atributos**) y un comportamiento o **método** (procedimiento que opera sobre los estados y un identificador de objeto (**IDO**).

Objetos complejos formados por conjuntos de atributos que a su vez pueden ser conjuntos de objetos.

Encapsulamiento de los objetos definidos por interfaces que definen de que manera se pueden acceder a los objetos para operar sobre ellos.

Clases de objetos cada una de las cuales agrupan objetos que tienen los mismos atributos y métodos. Cada objeto es un **ejemplar o instancia** de una clase.

Herencia es la conformación de **especializaciones de una clase** de la cual heredan métodos y atributos que se denomina **subclase**, mientras que una **generalización** es una **superclase** de la cual se hereda o **metaclase**.

Sobrecarga (overloading), **suplantación o anulación** (overriding) y **ligaduras estáticas** (late binding) o **ligaduras dinámicas** (dynamics binding) son formas de asociaciones de objetos mediante esas funciones sobre los métodos y atributos.

Claves candidatas son las formadas por varios atributos y las **claves primarias** son las específicas de una relación.

El IDO no tiene nada que ver con el estado, es una identificación propia del objeto o de clases de objetos si son complejos o sus instancias, sistemas GemStone del Smalltalk [Breitl,R. et al,1989], O_2 [Deux,O. et al,1990], Iris [Fishman,D. et al,1989] y Orion [kim,W. et al,1989].

2.2.6. PRIMERAS CONCLUSIONES

Los modelos semánticos de datos, al igual que el modelo entidad-relación y el modelo funcional, representan un intento de capturar tanto conjuntos de relaciones semánticas entre entidades del mundo real como sea posible. Las relaciones de generalización/especialización, agregación y la de instancia-de son modeladas eficientemente. En modelo de datos orientado a objetos es menos expresivo que el modelo semántico de datos, pero estos carecen del concepto de método, es por eso que se está tratando de extender al modelo orientado a objetos con funcionalidades tales como visiones y objetos compuestos [Kim,W., 1990]. Genéricamente se puede decir que mientras uno ofrece mecanismos para abstracción estructural y para representar conocimiento el otro ofrece mecanismos para representar abstracción del comportamiento, por lo tanto, similar a los lenguajes de programación orientados a objetos, justamente porque se trata de extender el lenguaje a la técnica de base de datos [Bertino,E., Martino,L., 1991].

En el caso de modelos de datos en red y jerárquicos hay una semejanza con los modelos de datos orientados a objetos, hay anidamiento de datos aunque en realidad se refieren a atributos, objetos y clases, pero no todas las jerarquías permiten ciclos y en orientados a objetos sí y en forma más natural. Otra similitud es la identificación de objetos con tipos punteros lógicos pero en la integridad hay que tener cuidado con la integridad referencial.

Se desarrollan bases de datos extensibles para incluir nuevas funciones [Schwarz et al,1986] o en bibliotecas de módulos básicos [Batory,D. et al,1988], pero siempre requieren un motor en un lenguaje orientados a objetos, entre una DBMGOO y una DBMG extensible es que esta última está orientada al diseño

arquitectónico físico y la primera está orientada al diseño arquitectónico lógico, con los inconvenientes del lenguaje ya mencionados anteriormente.

En el modelo relacional los objetos complejos hay que modelarlos en forma indirecta, no hay herencia ni asociación de operaciones pues no existe el método hay que hacerlo en las aplicaciones una de las soluciones es el modelo anidado que trata de obviar las formas normales.

De todas formas en el estado actual del arte los desarrollos ya sea teóricos como empíricos aún se inclinan por los estándares relacionales, sobre todo porque el modelo navegacional en el modelo de datos orientados a objetos se requiere un retroceso a formas jerárquicas o en red para usar punteros físicos y lógicos, "parches (patchiness) para resolver forzadamente los enlaces (links)".
Otro aspecto diferencial es que el modelo de datos orientados a objetos no tiene una teoría matemática coherente que le sirva de base, así como tecnologías simples de control de autorizaciones, control de concurrencia, la recuperación y el control de integridad, que aunque los posea son muy complejos [Bancilhon,F., 1988].

2.2.7. MINERÍA DE DATOS INTELIGENTE (DATA MINING).

La mayoría de las aplicaciones de la Inteligencia Artificial a tareas de importancia práctica construyen un modelo de conocimiento utilizable por un experto humano. En algunos casos, la tarea que el experto realiza es una *clasificación*, es decir, asigna objetos a categorías o clases determinadas según sus propiedades [Quinlan 1993d]. En un modelo de clasificación, la conexión entre clases y propiedades puede definirse utilizando desde un simple diagrama de flujo hasta un manual de procedimientos complejo y desestructurado. Si restringimos nuestra discusión a modelos ejecutables, es decir, a aquellos que pueden ser representados con métodos computacionales, existiendo dos maneras muy diferentes en las que se puede construir un modelo. Por un lado, el modelo puede obtenerse a partir de entrevistas relevantes con uno o más expertos. Por otro lado, si se cuenta con clasificaciones almacenadas con anterioridad, éstas pueden ser examinadas para construir un modelo inductivo a partir de ellas, mediante una generalización de ejemplos específicos. Los

sistemas ID3 y C4.5 pertenecen a este segundo grupo [Blockeel y De Raedt, 1997].

2.2.7.1. Marco teórico

Al plantear el problema de aprendizaje de un modelo de datos a partir de ejemplos desde un marco teórico, encontramos el siguiente esquema [Blockeel y De Raedt, 1997]:

Dados:

- un conjunto C de clases,
- un conjunto E de ejemplos preclasificados

Encontrar:

Una hipótesis H (conjunto de cláusulas) tal que:

$$\forall e \in E: H \cap e = c \wedge H \cap e \neq c' \qquad 2.2.7.1.1.$$

Donde c es la clase del ejemplo e y $c' \in C-\{c\}$

Presentar los resultados obtenidos como:

- un árbol de decisión,
- un conjunto de reglas de decisión.

El sistema generará un árbol de decisión fruto de la naturaleza en sí de los algoritmos de la familia TDIDT. El árbol de inducción resultante será construido desde la raíz hacia las hojas *(top-down).* El modelo generado es muy útil para el usuario ya que permite una fácil visualización de los resultados. Además, se transforma al árbol en reglas de decisión que pueden ser utilizadas por otros programas de clasificación o ser transformadas en sentencias SQL para clasificar nuevos datos rápidamente.

2.2.7.1.1. Datos de Entrada

Antes de analizar la familia TDIDT se debe tener en cuenta que no todas las tareas de clasificación son apropiadas para este enfoque inductivo, a continuación se listan los requerimientos que deben cumplirse [Mitchell, 1997], [Quinlan, 1986; 1993b]:

- **Descripciones de atributo-valor** *(Attribute-value description.) Archivos planos.*
- **Clases predefinidas** del tipo *{valor_atributo$_1$, valor_atributo$_2$,..., valor_atributo$_n$, clase$_k$}*.

- **Clases discretas y disjuntas** dado la naturaleza de los árboles de decisión, las clases deben ser discretas o discretizarse en caso de ser continuas.
- **Datos suficientes** los patrones generados por generalización inductiva no serán válidos si no se los pueden distinguir de las casualidades.
- **Los datos de entrenamiento pueden contener errores:** según Mitchell, los métodos de aprendizaje utilizando árboles de decisión son robustos frente a los errores, tanto en los valores de las clases como en los valores de los atributos de los datos de entrenamiento [Mitchell 1997].
- **Los datos de entrenamiento pueden contener valores de atributos faltantes** en los métodos de la familia TDIDT. El tratamiento de valores faltantes varía de un algoritmo a otro ID3 y C4.5.
- **Modelos lógicos generados:** los programas sólo construyen clasificadores que pueden ser expresados como árboles de decisión o como un conjunto de reglas de producción.

2.2.7.1.2. Resultados Generados. Características de los árboles de decisión.

Los árboles de decisión representan una estructura de datos que organiza eficazmente a los descriptores. Se construye un árbol de forma tal que en cada nodo se realiza una prueba sobre el valor de los descriptores y de acuerdo con la respuesta se va descendiendo en las ramas, hasta llegar al final del camino donde se encuentra el valor del clasificador. Se puede analizar un árbol de decisión como una caja negra en función de cuyos parámetros (descriptores) se obtiene un cierto valor del clasificador.

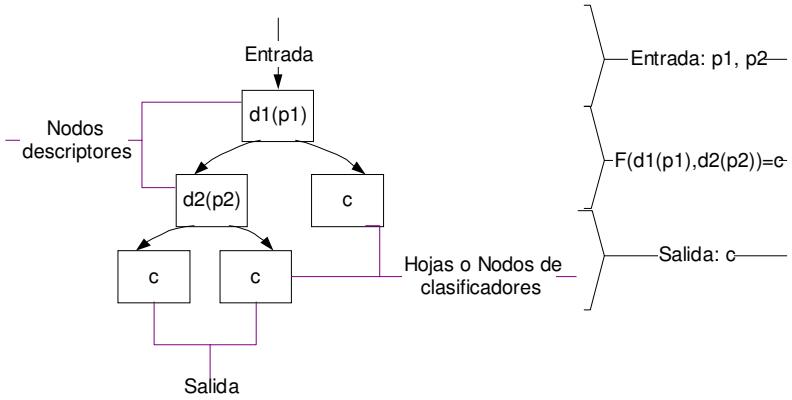


Figura 2.2.7.1.2.1.: Estructura de un árbol de decisión

Un árbol de decisión puede analizarse como una disyunción de conjunciones. Cada camino desde la raíz hasta las hojas representa una conjunción, y todos los caminos son alternativos, es decir, son disyunciones.

2.2.7.1.3. Características de las reglas de decisión

Las reglas de decisión o de producción son una alternativa a los árboles de decisión, y todo árbol de decisión puede llevarse a reglas de este tipo [Witten y Frank, 2000], [Korab, 1997], [Blurock, 1996].

Antecedente => Consecuente

Donde el antecedente es una conjunción entre distintas pruebas de valor sobre los valores de los atributos; y el consecuente es una clase para todos los casos que satisfagan al antecedente. Por ejemplo,

Si atributo$_1$=”valor a” y atributo$_2$= “valor y”, entonces Clase$_K$

Las reglas de decisión se presentan en orden, y deben interpretarse de esa manera. El orden determina cuáles reglas deben ejecutarse primero. Al clasificar un nuevo caso se avanza en la lista hasta llegar a un antecedente que sea satisfecho por el caso, entonces la clase del caso es la correspondiente al consecuente de dicha regla. El C4.5 en particular, agregar una última regla a la lista, esta no tiene antecedente, es la regla con la clase por defecto, es decir, si el caso no satisfizo ninguna de las reglas anteriores, entonces es de la clase indicada por la última regla que no tiene antecedente.

En el caso de las reglas de decisión, agregar una nueva regla implica simplemente añadirla a la lista de reglas sin necesidad de hacer cambios de estructura, mientras que agregar una nueva regla en un árbol implicaría rehacer la estructura del mismo.

2.2.7.1.4. Presentación de los resultados

Tanto el ID3 como el C4.5 generan un clasificador de la forma de un árbol de decisión, cuya estructura es [Quinlan 1993d]:

- Una *hoja*, indicando una clase, o
- Un *nodo de decisión* que especifica alguna prueba a ser realizada sobre un único atributo, con una rama y subárbol para cada valor posible de la prueba.

El árbol de decisión generado por el C4.5 cuenta con varias características particulares: cada hoja tiene asociados dos números, que indican el número de

casos de entrenamientos cubiertos por cada hoja y la cantidad de ellos clasificados erróneamente por la hoja. Es en cierta manera, un estimador del éxito del árbol sobre los casos de entrenamiento. El ID3, en cambio, no clasifica erróneamente a los datos de entrenamiento, con lo cual no son necesarios este tipo de indicadores. Es por ello, que este algoritmo, a diferencia del C4.5, corre el riesgo de caer en sobreajuste.

El propósito de construir modelos de clasificación no se limita al desarrollo de predictores precisos, también es esencial que el modelo construido sea comprensible para los seres humanos. Michie critica al ID3 al sostener que los resultados recientes demuestran que los programas construidos sobre la base de sistemas tales como el ID3 pueden ser considerados, de alguna manera, "súper-programas" y al mismo tiempo ser incomprensibles para las personas [Michie 1986, p.233].

Se han estudiado varias maneras de simplificar los árboles de decisión. Por ejemplo, en un sistema integrado, los árboles generados por el C4.5 como por el ID3 se transforman en un conjunto de *reglas de producción o decisión*, un formato que parece más comprensible que los árboles, cuando estos últimos son demasiado extensos o frondosos.

2.2.7.2. Descripción General de los Algoritmos

El algoritmo principal de los sistemas de la familia TDIDT, a la cual pertenecen el ID3 y su descendiente el C4.5, es el proceso de generación de un árbol de decisión inicial a partir de un conjunto de datos de entrenamiento. La idea original está basada en un trabajo de Hoveland y Hunt de los años 50, culminado en el libro *Experiments in Induction* [Hunt et al, 1966] que describe varios experimentos con varias implementaciones de sistemas de aprendizaje de conceptos *(concept learning systems - CLS).*

2.2.7.2.1. División de los datos

El método "divide y reinarás" realiza en cada paso una partición de los datos del nodo según una prueba realizada sobre el "mejor" atributo. Cualquier prueba que divida a T de una manera no trivial, tal que al menos dos subconjuntos distintos $\{T_i\}$ no estén vacíos, eventualmente resultará en una partición de subconjuntos de una única clase, aún cuando la mayoría de los subconjuntos contengan un

solo ejemplo. Sin embargo, el proceso de construcción del árbol no apunta meramente a encontrar cualquier partición de este tipo, sino a encontrar un árbol que revele una estructura del dominio y, por lo tanto, tenga poder predictivo. Para ello, es necesario un número importante de casos en cada hoja o, dicho de otra manera, la partición debe tener la menor cantidad de clases posibles. En el caso ideal, conviene elegir en cada paso la prueba que genere el árbol más pequeño.

Entonces, se está buscando un árbol de decisión compacto que sea consistente con los datos de entrenamiento. Se podrían explorar todos los árboles posibles y elegir el más simple. Desafortunadamente, un número exponencial de árboles debería ser analizado. El problema de encontrar el árbol de decisión más pequeño consistente con un conjunto de entrenamiento es de complejidad NP-completa.

La mayoría de los métodos de construcción de árboles de decisión, incluyendo el C4.5 y el ID3, no permiten volver a estados anteriores, es decir, son algoritmos golosos sin vuelta atrás. Una vez que se ha escogido una prueba para particionar el conjunto actual, típicamente basándose en la maximización de alguna medida local de progreso, la partición se concreta y las consecuencias de una elección alternativa no se exploran. Por este motivo, la elección debe ser bien realizada.

2.2.7.2.2. Elección del criterio de división

Para realizar la división de los datos en cada paso, Quinlan propone la utilización de los métodos de la Teoría de la Información. En un principio, el ID3 utilizaba la ganancia como criterio de división. Sin embargo, a partir de numerosas pruebas se descubrió que este criterio no era efectivo en todos los casos y se obtenían mejores resultados si se normalizaba el criterio en cada paso. Por lo tanto, comenzó a utilizarse la ganancia de información, con mayor éxito. El C4.5 también utiliza este último criterio para realizar la división de los casos. Quinlan afirma que en su opinión el criterio de proporción de ganancia es robusto y generalmente da resultados más consistentes que el criterio de ganancia [Quinlan 1988b].

La solución propuesta permite la utilización de ambos criterios. Se deben estudiar y comparar los resultados obtenidos con el ID3 y con el C4.5 utilizando la ganancia y la proporción de ganancia.

2.2.7.2.3. Criterio de Ganancia

La definición de ganancia en información puede calcularse como la disminución en entropía. Es decir:

$$I(S,at) = H(S) - H(S,at) \qquad \text{2.2.7.2.3.1.}$$

Supongamos que tenemos una prueba posible con *n* resultados que particionan al conjunto *T* de entrenamiento en los subconjuntos $T_1, T_2, \ldots, T_n$. Si la prueba se realiza sin explorar las divisiones subsiguientes de los subconjuntos T_i, la única información disponible para evaluar la partición es la distribución de clases en *T* y sus subconjuntos.

Consideremos una medida similar luego de que ***T*** ha sido particionado de acuerdo a los *n* resultados de la prueba ***X***. La información esperada (entropía) puede determinarse como la suma ponderada de los subconjuntos, de la siguiente manera

$$H(T,X) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times H(T_i) \qquad \text{2.2.7.2.3.2.}$$

La cantidad

$$I(T,X) = H(T) - H(T,X) \qquad \text{2.2.7.2.3.3.}$$

mide la información ganada al partir *T* de acuerdo a la prueba *X*. El criterio de ganancia, entonces, selecciona la prueba que maximice la ganancia de información. Es decir, antes de particionar los datos en cada nodo, se calcula la ganancia que resultaría de particionar el conjunto de datos según cada uno de los atributos posibles. Se realiza la partición que resulta en la mayor ganancia.

2.2.7.2.4. Criterio de Proporción de Ganancia

El criterio de ganancia tiene un defecto muy serio: presenta una tendencia muy fuerte a favorecer las pruebas con muchos resultados. Analicemos una prueba sobre un atributo que sea la clave primaria de un conjunto de datos, en la cual, obtendremos un único subconjunto para cada caso, y para cada subconjunto tendremos *I (T,X) = 0*, entonces la ganancia de información será máxima. Desde el punto de vista de la predicción, este tipo de división no es útil.

Esta tendencia inherente al criterio de ganancia puede corregirse mediante una suerte de normalización, en la cual se ajusta la ganancia aparente, atribuible a pruebas con muchos resultados. Consideremos el contenido de información de un mensaje correspondiente a los resultados de las pruebas. Por analogía a la definición de la *I(S)* tenemos:

$$I\_división(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \log_2\left(\frac{|T_i|}{|T|}\right) \qquad \text{2.2.7.2.4.1.}$$

Esto representa la información potencial generada al dividir *T* en *n* subconjuntos, mientras que la ganancia de información mide la información relevante a una clasificación que nace de la misma división. Entonces,

$$proporción\_de\_ganancia(X) = \frac{I(T,X)}{I\_división(X)} \qquad \text{2.2.7.2.4.2.}$$

expresa la proporción útil de información generada en la partición. Si la partición es casi trivial, la información de la división será pequeña y esta proporción se volverá inestable. Para evitar este fenómeno, el criterio de proporción de ganancia selecciona una prueba que maximice la expresión anterior, sujeto a la restricción de que la información de la división sea grande, al menos tan grande como la ganancia promedio sobre todas las pruebas realizadas.

2.2.7.3. ID3

El algoritmo ID3 fue diseñado por J. Ross Quinlan [Quinlan, 1993a, 1993b]. El ID3 toma objetos de una clase conocida y los describe en términos de una colección fija de propiedades o de atributos, y produce un árbol de decisión sobre estos atributos que clasifica correctamente todos los objetos [Quinlan, 1993b]. Hay ciertas cualidades que diferencian a este algoritmo de otros sistemas generales de inferencia. La primera se basa en la forma en que el esfuerzo requerido para realizar una tarea de inducción crece con la dificultad de la tarea. El ID3 fue diseñado específicamente para trabajar con masas de objetos, y el tiempo requerido para procesar los datos crece sólo linealmente con la dificultad, como producto de:

- la cantidad de objetos presentados como ejemplos,
- la cantidad de atributos dados para describir estos objetos, y
- la complejidad del concepto a ser desarrollado (medido por la cantidad de nodos en el árbol de decisión)

Esta linealidad se consigue a costo del poder descriptivo: los conceptos desarrollados por el ID3 sólo toman la forma de árboles de decisión basados en los atributos dados, y este "lenguaje" es mucho más restrictivo que la lógica de primer orden o la lógica multivaluada, en la cual otros sistemas expresan sus conceptos [Quinlan, 1993b].

El ID3 fue presentado como descendiente del CLS creado por Hunt. El ID3, como contrapartida de su antecesor, es un mecanismo mucho más simple para el descubrimiento de una colección de objetos pertenecientes a dos o más clases. Cada objeto debe estar descrito en términos de un conjunto fijo de atributos, cada uno de los cuales cuenta con su conjunto de posibles valores de atributos. Por ejemplo, el atributo humedad puede tener los valores {alta, baja}, y el atributo clima, {soleado, nublado, lluvioso}.

Una regla de clasificación en la forma de un árbol de decisión puede construirse para cualquier conjunto C de atributos de esa forma [Quinlan, 1993b]. Si C está vacío, entonces se lo asocia arbitrariamente a cualquiera de las clases. Si no, C contiene los representantes de varias clases; se selecciona un atributo y se particiona C en conjuntos disjuntos C_1, C_2,..., C_n, donde C_i contiene aquellos miembros de C que tienen el valor *i* para el atributo seleccionado. Cada una de estos subconjuntos se maneja con la misma estrategia. El resultado es un árbol en el cual cada hoja contiene un nombre de clase y cada nodo interior especifica un atributo para ser testeado con una rama correspondiente al valor del atributo.

2.2.7.3.1. Descripción del ID3

El objetivo del ID3 es crear una descripción eficiente de un conjunto de datos mediante la utilización de un árbol de decisión. Dados datos consistentes, es decir, sin contradicción entre ellos, el árbol resultante describirá el conjunto de entrada a la perfección. Además, el árbol puede ser utilizado para predecir los valores de nuevos datos, asumiendo siempre que el conjunto de datos sobre el cual se trabaja es representativo de la totalidad de los datos.

Dados:

- Un conjunto de datos
- Un conjunto de descriptores de cada dato
- Un clasificador/conjunto de clasificadores para cada objeto.

Se desea obtener:

- Un árbol de decisión simple basándose en la entropía, donde los nodos pueden ser:

1. Nodos intermedios: en donde se encuentran los descriptores escogidos según el criterio de entropía, que determinan cuál rama es la que debe tomarse.
2. Hojas: estos nodos determinan el valor del clasificador.

Este procedimiento de formación de reglas funcionará siempre dado que no existen dos objetos pertenecientes a distintas clases pero con idéntico valor para cada uno de sus atributos; si este caso llegara a presentarse, los atributos son inadecuados para el proceso de clasificación.

Hay dos conceptos importantes a tener en cuenta en el algoritmo ID3 [Blurock, 1996]: la entropía y el árbol de decisión. La entropía se utiliza para encontrar el parámetro más significativo en la caracterización de un clasificador. El árbol de decisión es un medio eficiente e intuitivo para organizar los descriptores que pueden ser utilizados con funciones predictivas.

2.2.7.3.2. Algoritmo ID3

A continuación se presenta el algoritmo del método ID3 para la construcción de árboles de decisión en función de un conjunto de datos previamente clasificados.

Función ID3
(R: conjunto de atributos no clasificadores,
C: atributo clasificador,
S: conjunto de entrenamiento) devuelve un árbol de decisión;

Comienzo
Si S está vacío,
    devolver un único nodo con Valor Falla;
Si todos los registros de S tienen el mismo valor para el atributo clasificador,
    Devolver un único nodo con dicho valor;
Si R está vacío, entonces
devolver un único nodo con el valor más frecuente del atributo clasificador en los registros de S [Nota: habrá errores, es decir, registros que no estarán bien clasificados en este caso];

Si R no está vacío, entonces
D ← atributo con mayor Ganancia(D,S) entre los atributos de R;
Sean {dj| j=1,2,…, m} los valores del atributo D;
Sean {Sj| j=1,2, .., m} los subconjuntos de S correspondientes a los valores de dj respectivamente;
Devolver un árbol con la raíz nombrada como D y con los arcos nombrados d1, d2, .., dm que van respectivamente a los árboles
ID3(R-{D}, C, S1), ID3(R-{D}, C, S2), .., ID3(R-{D}, C, Sm);
Fin

2.2.7.3.3. Poda de los árboles de decisión

La poda de los árboles de decisión se realiza con el objetivo de que éstos sean más comprensibles. Lo cual implica que tengan menos niveles y/o sean menos frondosos. La poda aplicada en el ID3 se realiza una vez que el árbol ha sido generado y es un mecanismo bastante simple: si de un nodo nacen muchas ramas, las cuales terminan todas en la misma clase, entonces se reemplaza dicho nodo por una hoja con la clase común. En caso contrario, se analizan todos los nodos hijos.

2.2.7.3.4. Pasaje a reglas de decisión

Para pasar a reglas de decisión, el ID3 recorre el árbol desde la raíz hasta las hojas y genera una regla por cada camino recorrido. El antecedente de cada regla estará compuesto por la conjunción de las pruebas de valor de cada nodo visitado, y la clase será la correspondiente a la hoja. El recorrido del árbol se basa en el recorrido de preordenado (de raíz a hojas, de izquierda a derecha). Al estar trabajando con árboles n-arios, este recorrido es único.

2.2.7.3.5. Atributos desconocidos

Es necesario que todos los casos presentados al ID3 estén descriptos por los mismos atributos. Esto limita la aplicación del algoritmo, ya que no siempre se cuenta con toda la información necesaria. Imaginemos una base de datos histórica en la que se fueron agregando atributos a medida que se lo consideró necesario, para los primeros casos de la misma no se conocerán los valores de los nuevos atributos. El ID3 puede trabajar con atributos desconocidos, los

considera como si fuesen un nuevo valor, por ello, se llega a la convención de que los valores desconocidos, deben expresarse con un "?" en los datos. El "?" constituye un nuevo valor posible para el atributo en cuestión.

Una vez calculadas las ganancias y proporciones de ganancia para todos los atributos disponibles, se debe elegir el atributo según el cual se divide a este conjunto de datos. Tanto en el caso de la ganancia como en el de la proporción de ganancia, el mejor atributo para la división es aquel que la maximiza. En este ejemplo, la división según el atributo Estado es la que mayor ganancia y proporción de ganancia ofrece. Esto significa que el nodo raíz del árbol será un nodo que evalúa el atributo Estado.

2.2.7.3.6. Transformación a reglas de decisión

Como se explicó en la sección 2.2.7.3.4. para pasar un árbol de decisión a reglas de decisión, el ID3 lo recorre en preorden y cada vez que llega a una hoja, escribe la regla que tiene como consecuente el valor de la misma, y como antecedente, la conjunción de las pruebas de valor especificados en todos los nodos recorridos desde la raíz para llegar a dicha hoja. Analicemos el pasaje del árbol a reglas de decisión.

Una limitación del ID3 es que puede aplicarse a cualquier conjunto de datos, siempre y cuando los atributos sean discretos. Este sistema no cuenta con la facilidad de trabajar con atributos continuos ya que analiza la entropía sobre cada uno de los valores de un atributo, por lo tanto, tomaría cada valor de un atributo continuo individualmente en el cálculo de la entropía, lo cual no es útil en muchos de los dominios. Cuando se trabaja con atributos continuos generalmente se piensa en rangos de valores y no en valores particulares.

Existen varias maneras de solucionar este problema del ID3, como la agrupación de valores presentada en [Gallion et al, 1993] o la discretización de los mismos explicada en [Blurock, 1996], [Quinlan, 1993d]. El C4.5 resolvió el problema de los atributos continuos mediante la discretización.

El proceso descrito para la construcción de árboles de decisión asume que las operaciones de cálculo, especialmente, las de evaluación de las frecuencias relativas (en las que se deben contar elementos) del conjunto C, pueden ser realizadas eficientemente, lo cual significa, en la práctica, que para que el proceso sea rápido, C debe residir en memoria. La solución aplicada por ID3,

para instancias en la memoria, es una solución iterativa, que crea sucesivos árboles de decisión de precisión cada vez mayor, hasta llegar al árbol de decisión óptimo. El método puede resumirse como [Quinlan, 1993b]:
Elegir un conjunto aleatorio de instancias (llamado *ventana).*
Repetir:
Formar una regla para explicar la ventana actual
Encontrar las excepciones a la regla en el resto de las instancias
Crear una nueva ventana a partir de la ventana actual y las excepciones a la regla generada a partir de ella
Hasta que no queden excepciones a la regla.

El proceso termina cuando se forma una regla que no tenga excepciones y sea correcta para todo C.

2.2.7.4. C4.5

El C4.5 se basa en el ID3, por lo tanto, la estructura principal de ambos métodos es la misma. El C4.5 construye un árbol de decisión mediante el algoritmo "divide y reinarás" y evalúa la información en cada caso utilizando los criterios de entropía y ganancia o proporción de ganancia, según sea el caso. A continuación, se explicarán las características particulares de este método que lo diferencian de su antecesor.

2.2.7.4.1. Algoritmo C4.5

El algoritmo del método C4.5 para la construcción de árboles de decisión a grandes rasgos muy similar al del ID3. Varía en la manera en que realiza las pruebas sobre los atributos, tal como se detalla en las secciones siguientes.
Función C4.5
(R: conjunto de atributos no clasificadores,
C: atributo clasificador,
S: conjunto de entrenamiento) devuelve un árbol de decisión;

Comienzo
Si S está vacío,
devolver un único nodo con Valor Falla;

Si todos los registros de S tienen el mismo valor para el atributo clasificador,
Devolver un único nodo con dicho valor;
Si R está vacío, entonces
devolver un único nodo con el valor más frecuente del atributo clasificador en los registros de S [Nota: habrá errores, es decir, registros que no estarán bien clasificados en este caso];
Si R no está vacío, entonces
D ← atributo con mayor Proporción de Ganancia (D,S) entre los atributos de R;
Sean {dj| j=1,2, .., m} los valores del atributo D;
Sean {Sj| j=1,2, .., m} los subconjuntos de S correspondientes a los valores de dj respectivamente;
Devolver un árbol con la raíz nombrada como D y con los arcos nombrados
d1, d2, .., dm que van respectivamente a los árboles
C4.5(R-{D}, C, S1), C4.5(R-{D}, C, S2), .., C4.5(R-{D}, C, Sm);
Fin

2.2.7.4.2. Características particulares del C4.5. Pruebas utilizadas.

En cada nodo, el sistema debe decidir cuál prueba escoge para dividir los datos. Los tres tipos de pruebas posibles propuestas por el C4.5 son [Quinlan, 1993d]:

1. La prueba "estándar" para los atributos discretos, con un resultado y una rama para cada valor posible del atributo.
2. Una prueba más compleja, basada en un atributo discreto, en donde los valores posibles son asignados a un número variable de grupos con un resultado posible para cada grupo, en lugar de para cada valor.
3. Si un atributo A tiene valores numéricos continuos, se realiza una prueba binaria con resultados $A \leq Z$ y $A > Z$, para lo cual debe determinarse el valor límite Z.

Todas estas pruebas se evalúan de la misma manera, mirando el resultado de la proporción de ganancia, o alternativamente, el de la ganancia, resultante de la división que producen. Ha sido útil agregar una restricción adicional: para cualquier división, al menos dos de los subconjuntos T_i deben contener un número razonable de casos. Esta restricción, que evita las subdivisiones casi triviales es tenida en cuenta solamente cuando el conjunto T es pequeño.

2.2.7.4.3. Pruebas sobre atributos continuos

Las pruebas para valores continuos trabajan con un valor límite arbitrario. El método utilizado para ello por el C4.5 es muy simple [Quinlan, 1993d] [Quinlan, 1996a]. Primero, los casos de entrenamiento *T* se ordenan según los valores del atributo *A* continuo que está siendo considerado. Existe un número finito de estos valores.

Sean $\{v_1, v_2, \ldots, v_m\}$ los valores que toma el atributo *A*. Cualquier valor límite entre v_i y v_{i+1} tendrá el mismo efecto al dividir los casos entre aquellos cuyo valor para *A* pertenece al subconjunto $\{v_1, v_2, \ldots, v_i\}$ y aquellos cuyo valor pertenece a $\{v_{i+1}, v_{i+2}, \ldots, v_m\}$. Entonces, existen sólo $m - 1$ divisiones posibles de según el valor de *A* y todas son examinadas. Al estar ordenados, las sucesivas pruebas para todos los valores, pueden realizarse en una única pasada.

Típicamente se elige el punto medio del intervalo como valor límite representativo, entonces el *i*ésimo valor límite sería:

$$\frac{v_i + v_{i+1}}{2} \qquad \text{2.2.7.4.2.1.}$$

C4.5 se diferencia de otros algoritmos en que elige el mayor valor de *A* en todo el conjunto de casos de entrenamiento que no excede el punto medio presentado, en lugar del punto medio en sí mismo, como valor límite; de esta manera se asegura que todos los valores límites que aparezcan en el árbol y/o las reglas ocurran al menos una vez en los datos.

El método utilizado para la binarización de atributos tiene una gran desventaja. Mientras que todas las operaciones de construcción de un árbol de decisión crecen linealmente con el número de casos de entrenamiento, el ordenamiento de *d* valores continuos crece proporcionalmente a $d \times \log(d)$. Entonces, el tiempo requerido para construir un árbol a partir de un gran conjunto de datos de entrenamiento, puede estar dominado por el ordenamiento de datos con valores continuos.

2.2.7.4.4. Atributos desconocidos

C4.5 asume que todos los resultados de pruebas desconocidos se distribuyen en forma probabilística según la frecuencia relativa de los valores conocidos. Un caso (posiblemente fraccional) con un valor desconocido se divide en fragmentos cuyos pesos son proporcionales a dichas frecuencias relativas, dando por resultado que un caso puede seguir múltiples caminos en el árbol.

Esto se aplica tanto cuando los casos de entrenamiento se dividen durante la construcción del árbol, como cuando el árbol se utiliza para clasificar casos.

2.2.7.4.5. Evaluación de las pruebas

La modificación del criterio de *ganancia* es bastante directa. La *ganancia* de una prueba mide la información sobre la pertenencia a una clase que puede esperarse como resultado de partir un conjunto de datos de entrenamiento, calculada al restar la información que se espera que sea necesaria para identificar la clase de un objeto después de la partición a la misma cantidad antes de la partición. Es evidente que una prueba no puede proveer información de pertenencia a una clase si no se conoce el valor de un atributo.

Sea *T* el conjunto de datos de entrenamiento y *X* una prueba basada en un atributo *A*, supongamos que el valor de *A* se conoce únicamente en una fracción *F* de casos en *T*. Sean *I(T)* e $I_X(T)$ calculadas según, si el atributo *at* divide el conjunto *S* en los subconjuntos S_i, i = 1, 2, …, n, entonces, la entropía total del sistema de subconjuntos será:

$$H(S,at) = \sum_{i=1}^{n} P(S_i) \cdot H(S_i) \qquad \text{2.2.7.4.5.1.}$$

excepto que sólo se tienen en cuenta los casos para los cuales el valor de *A* es conocido. La definición de ganancia puede corregirse a:

$$\begin{aligned} &Ganancia\,(X) = probabilid\,ad\_A\_sea\_conocido \times (I(T) - I_X(T)) \\ &+ probabilid\,ad\_A\_no\_sea\_conocido \times 0 = \\ &F \times (I(T) - I_X(T)) \end{aligned}$$

2.2.7.4.5.2.

o, en otras palabras, la ganancia aparente de mirar a los casos con valores conocidos, multiplicada por la fracción de dichos casos en el conjunto de entrenamiento.

El cálculo de la proporción de ganancia se realiza de la misma manera. La definición de información de la división puede modificarse de manera similar, considerando los casos con valores desconocidos como un grupo más, entonces, si una prueba tienen *n* resultados, su información de la división se calcula como la prueba dividido *n+1*subconjuntos.

$$I\_división\ (X) = -\sum_{i=1}^{n+1} \frac{|T_i|}{|T|+1} \times \log_2 \left(\frac{|T_i|}{|T|+1} \right)$$

2.2.7.4.5.3.

2.2.7.4.6. Partición del conjunto de entrenamiento

Una prueba puede seleccionar del conjunto de pruebas posibles, como antes, pero utilizando las versiones modificadas de ganancia e información de la división. Si la prueba *X* con resultados $O_1, O_2, ..., O_N$ es escogida y tiene algunos valores desconocidos para algunos de los datos de entrenamiento, el concepto de particionamiento debe ser generalizado, según un criterio probabilístico.

Cuando un caso *T* con un resultado conocido O_i es asignado al subconjunto T_i, esto significa que la probabilidad de que el caso pertenezca a T_i es 1 y de que pertenezca a todos los otros subconjuntos es 0. Cuando el resultado es desconocido, sólo se puede realizar una afirmación estadística más débil. Entonces, se asocia con cada caso del subconjunto T_i un *peso* representando la probabilidad de que el caso pertenezca a cada subconjunto. Si el resultado para el caso es conocido, entonces el peso es 1; si el caso tiene un resultado desconocido, entonces el peso es simplemente la probabilidad del resultado O_i en este punto. Cada subconjunto T_i es una colección de casos fraccionales posibles, tal que $|T_i|$ debe ser reinterpretada como la suma de los pesos fraccionales de los casos pertenecientes al subconjunto.

Los casos de entrenamiento en *T* pueden tener pesos no unitarios, ya que *T* puede ser el resultado de una partición previa. Entonces, en general, un caso de *T* con peso *p* cuyo resultado no se conoce, es asignado a cada subconjunto T_i con peso:

P x *probabilidad_de_resultado_*O_i 2.2.7.4.6.1.

La *probabilidad_de_resultado_*O_i se estima como la suma de los pesos de los casos en *T* con valores conocidos que tienen resultado O_i, sobre la suma de los pesos de los casos en *T* con resultado conocidos para la prueba.

2.2.7.4.7. Clasificación de un caso nuevo

Se toma un enfoque similar cuando el árbol de decisión es utilizado para clasificar un caso. Si en un nodo de decisión el atributo relevante no se conoce, de manera tal que el resultado de la prueba no puede determinarse, el sistema

explora todos los resultados posibles y combina aritméticamente las clasificaciones resultantes. Como para cada atributo pueden existir múltiples caminos desde la raíz del árbol hasta las hojas, una "clasificación" es una distribución de clases más que una única clase. Cuando la distribución de clases total para un caso nuevo ha sido establecida de esta manera, la clase con la probabilidad más alta, es asignada como "la" clase predicha.

La información de la división aún se determina a partir del conjunto de entrenamiento completo y es mayor, ya que existe una categoría extra para los valores desconocidos.

Cada hoja en el árbol de decisión resultante tiene asociados dos valores: *(N/E)*. *N* es la suma de los casos fraccionales que llegan a la hoja; y *E* es el número de casos cubiertos por la hoja, que no pertenecen a la clase de la misma.

2.2.7.4.8. Poda de los Árboles de Decisión

El método recursivo de particionamiento para construir los árboles de decisión descrito anteriormente, subdividirá el conjunto de entrenamiento hasta que la partición contenga casos de una única clase, o hasta que la prueba no ofrezca mejora alguna. Esto da como resultado, generalmente, un árbol muy complejo que sobreajusta los datos al inferir una estructura mayor que la requerida por los casos de entrenamiento [Mitchell, 2000b] [Quinlan, 1995]. Además, el árbol inicial generalmente es extremadamente complejo y tiene una proporción de errores superior a la de un árbol más simple. Mientras que el aumento en complejidad se comprende a simple vista, la mayor proporción de errores puede ser más difícil de visualizar.

Para entender este problema, supongamos que tenemos un conjunto de datos dos clases, donde una proporción $p \geq 0.5$ de los casos pertenecen a la clase mayoritaria. Si un clasificador asigna todos los casos con valores indeterminados a la clase mayoritaria, la proporción esperada de error es claramente $1 - p$. Si, en cambio, el clasificador asigna un caso a la clase mayoritaria con probabilidad p y a la otra clase con probabilidad $1 - p$, su proporción esperada de error es la suma de:

- la probabilidad de que un caso perteneciente a la clase mayoritaria sea asignado a la otra clase, $p \times (1 - p)$, y

- la probabilidad de que un caso perteneciente a la otra clase sea asignado a la clase mayoritaria, (1 – *p*) x *p*

que da como resultado 2 x *p* (1 – *p*). Como *p* es al menos 0.5, esto es generalmente superior a 1 – *p*, entonces el segundo clasificador tendrá una mayor proporción de errores. Un árbol de decisión complejo tiene una gran similitud con este segundo tipo de clasificador. Los casos no se relacionan a una clase, entonces, el árbol manda cada caso al azar a alguna de las hojas.

Un árbol de decisión no se simplifica borrando todo el árbol a favor de una rama, sino que se eliminan las partes del árbol que no contribuyen a la exactitud de la clasificación para los nuevos casos, produciendo un árbol menos complejo, y por lo tanto, más comprensible.

Existen, básicamente, dos maneras de modificar el método de particionamiento recursivo para producir árboles más simples: decidir no dividir más un conjunto de casos de entrenamiento, o remover retrospectivamente alguna parte de la estructura construida por el particionamiento recursivo.

El primer enfoque, conocido como pre-poda, tiene la ventaja de que no se pierde tiempo en construir una estructura que luego será simplificada en el árbol final. Los sistemas que lo aplican, generalmente buscan la mejor manera de partir el subconjunto y evalúan la partición desde el punto de vista estadístico mediante la teoría de la ganancia de información, reducción de errores, etc. Si esta evaluación es menor que un límite predeterminado, la división se descarta y el árbol para el subconjunto es simplemente la hoja más apropiada. Sin embargo, este tipo de método tiene la desventaja de que no es fácil detener un particionamiento en el momento adecuado, un límite muy alto puede terminar con la partición antes de que los beneficios de particiones subsiguientes parezcan evidentes, mientras que un límite demasiado bajo resulta en una simplificación demasiado leve.

El C4.5 utiliza el segundo enfoque, el método de divide y reinarás procesa los datos de entrenamiento libremente, y el árbol sobreajustado producido es podado después. Los procesos computacionales extras invertidos en la construcción de partes del árbol que luego serán podadas pueden ser sustanciales, pero el costo no supera los beneficios de explorar una mayor

cantidad de particiones posibles. El crecimiento y la poda de los árboles son más lentos, pero más confiables.

La poda de los árboles de decisión llevará, sin duda, a clasificar erróneamente una mayor cantidad de los casos de entrenamiento. Por lo tanto, las hojas de un árbol podado no contendrán necesariamente una única clase sino una distribución de clases, como se explicó con anterioridad. Asociado a cada hoja, habrá una distribución de clases especificando, para cada clase, la probabilidad de que un caso de entrenamiento en la hoja pertenezca a dicha clase.

Generalmente, la simplificación de los árboles de decisión se realiza descartando uno o más subárboles y reemplazándolos por hojas. Al igual que en la construcción de árboles, las clases asociadas con cada hoja se encuentran al examinar los casos de entrenamiento cubiertos por la hoja y eligiendo el caso más frecuente. Además de este método, el C4.5 permite reemplazar un subárbol por alguna de sus ramas.

Supongamos que fuera posible predecir la proporción de errores de un árbol y sus subárboles. Esto inmediatamente llevaría al siguiente método de poda: "Comenzar por las hojas y examinar cada subárbol. Si un reemplazo del subárbol por una hoja o por su rama más frecuentemente utilizada, lleva a una proporción de errores predicha *(predicted error rate)* menor, entonces podar el árbol de acuerdo a ello, recordando que las proporciones de errores predichas para todos los subárboles que lo contienen se verán afectadas". Como la proporción de errores predicha para un árbol disminuye si disminuyen las proporciones de errores predichas en cada una de sus ramas, este proceso generaría un árbol con una proporción de errores predicha mínima.

Ejemplos de esta familia son:

- *Poda según la complejidad del costo (Cost-complexity pruning).*
- *Poda de reducción de errores (Reduced-error pruning)* [Quinlan, 1987e].

2.2.7.4.9. Estimación de la Proporción de Errores para los Árboles de Decisión

Una vez podados, las hojas de los árboles de decisión generados por el C4.5 tendrán dos números asociados: *N* y *E*. *N* es la cantidad de casos de

entrenamiento cubiertos por la hoja, y *E* es la cantidad de errores predichos si un conjunto de *N* nuevos casos fuera clasificados por el árbol.
La suma de los errores predichos en las hojas, dividido el número de casos de entrenamiento, es un estimador inmediato del error de un árbol podado sobre nuevos casos.

2.2.7.5. SISTEMA INTEGRADOR

Para estudiar los algoritmos propuestos se desarrolló un sistema que integra el ID3 y el C4.5. El sistema recibe los datos de entrenamiento como entrada y permite que el usuario elija cuál algoritmo y con qué criterio de decisión (ganancia o proporción de ganancia) desea aplicar. Una vez generados el árbol y las reglas de decisión, el usuario puede evaluar los resultados sobre los datos de prueba. En el caso del ID3, esta evaluación se realiza a partir de las reglas de decisión cuya performance, es idéntica a la de los árboles. La evaluación de los resultados del C4.5, en cambio, se realiza por separado y se obtienen, por lo tanto, dos evaluaciones distintas, una para el árbol y otra para las reglas. Esto se debe a que, el modelo de clasificación generado con el C4.5 como árbol de decisión es distinto al generado como reglas de decisión.

2.3. ESTADO DEL ARTE EN TAXONOMIA

2.3.1. INTRODUCCIÓN

En forma preliminar se puede definir a la clasificación como al agrupamiento de objetos en clases, sobre la base de los atributos que poseen en común y/o sus relaciones, que pueden en sí mismas ser consideradas atributos.
Ante la diversidad se recurre a la clasificación como medio de evitar la confusión, en forma instintiva o consciente.
Desde la época de Carlos de Linneo hasta nuestros tiempos (hace 300 años) un punto débil en las Ciencias Biológicas ha sido la ausencia de significados cuantitativos en términos clasificatorios.

El enfoque de la clasificación o taxonomía numérica comprende un aspecto filosófico de la teoría de la clasificación o fenética y otro de técnicas numéricas, que son los pasos operativos para aplicar dicha teoría.

La cuestión puede ser puesta en otras categorías de clasificación, rango o nivel de una clasificación jerárquica, tales como Especie, Género, Familia, Orden, Clase, División, Reino, etc. En ningún caso la respuesta es total y en muchos casos no puede ser respondida. Hasta que las respuestas puedan ser dadas adecuadamente nuestros esquemas clasificatorios nunca pueden ser satisfactorios o naturales. Pueden ser un poco mejor que mnemotécnicas (artificio para mejorar la memoria por medios artificiales o codificación), simples esqueletos o estructuras de las cuales sostenemos, suspendemos, (enganchamos o enlazamos) algunos fragmentos del conocimiento.

¿Que es una clase y si una clase A difiere de una clase B tanto como una clase C difiere de una clase D?

Al final del siglo XIX, las doctrinas evolucionistas, de todos los sistemas de clasificación, no dieron respuesta a esas dificultades. La geología ha dado algunas respuestas aisladas. Pero describir la manera en la cual varios grupos de cosas existentes surgieron (o se originaron) es muy diferente a asignarle valores cuantitativos a esos grupos.

El rápido y explosivo desarrollo de la taxonomía numérica y el gran interés por este campo, produjo no solo nuevo material a ser estudiado sino también una estructura científica con su propia perspectiva. En tal sentido se han desarrollado una considerable cantidad de métodos numéricos en la taxonomía, muchos de ellos aplicados en computadoras. La influencia de estos métodos en otras disciplinas hizo que surgieran nuevos conceptos y técnicas para sistematizarla.

La taxonomía numérica, que es el agrupamiento mediante métodos numéricos de unidades taxonómicas basadas en el estado de sus caracteres (caracteres ≡ atributos: signos distintivos de las cosas), se ha potenciado y sostenido en su rápido desarrollo, por el desarrollo simultáneo de técnicas de cómputos.

El propósito de la Taxonomía Numérica es desarrollar métodos que sean objetivos, explícitos y repetibles para evaluar las relaciones taxonómicas y establecer el orden (taxa). Además, los métodos numéricos han abierto un amplio campo para la medición exacta de la razón de la evolución y del análisis

filogenético. El éxito de la aproximación intuitiva usada en el pasado para reconocer rápidamente con habilidad las observaciones, sin embargo inexactas, es totalmente similar en el detalle morfológico. Tal reconocimiento no es fácil con el continuo crecimiento de las bases de datos taxonómicas, ahora a menudo en forma de tablas, como ocurre con los caracteres en microbiología, química y fisiología.

'El uso de métodos numéricos con esos caracteres es una necesidad'.

2.3.2. EL OBJETIVO Y LOS PRINCIPIOS DE LA TAXONOMÍA NUMÉRICA

Los propósitos : Obtener teórica y prácticamente procesos clasificatorios y contrastar la visión convencional con los conceptos que sean susceptibles de evolución.

La **evidencia taxonómica** que implica la selección de objetos (organismos) de estudio, la selección y definición de caracteres taxonómicos y los criterios homológicos.

La estimación de la semejanza taxonómica entre (organismos) objetos y el agrupamiento de los mismos en taxones en función de sus semejanzas.

La evolución de la investigación sistemática y la aproximación al estudio detallado de la filogenética y el tratamiento con Taxonomía Numérica a poblaciones, conduce a la discusión de patrones fenotípicos y a la evolución estructural y semántica, que es la clave y la nomenclatura para la identificación.

En general se trata de mantener un simbolismo uniforme para los caracteres, unidades taxonómicas operacionales (OTU) y taxones (taxones ≡ taxa: grupos de OTUs).

2.3.3. DEFINICIONES

Sistemático: El estudio científico de las clases de tipos y diversidad de organismos y de cualquiera y todas las relaciones entre ellos.

Tomado en el sentido más amplio es el arreglo de organismos en taxones y su nominación y además las causas y orígenes de los arreglos.

Clasificación: Es el ordenamiento de los organismos en grupos (conjuntos, sets) en función de sus relaciones.

Relaciones pueden implicar relaciones filogenéticas o una relación puede indicar simplemente la semejanza o una completa similitud que surge de los caracteres de los organismos sin ninguna implicancia con sus antecesores. Ambas relaciones pueden ser distinguidas denominando a unas relaciones filogenéticas y a otras relaciones, como relaciones fenotípicas, definidas por el fenotipo a partir de sus caracteres y no de su filogenia.

En general es el resultado de un proceso donde la clasificación es el producto. No es una identificación y no es un arreglo convencional de clases en un espectro continuo sin distinguir divisiones.

Identificación: Es la ubicación o asignación, de un objeto adicional no identificado, en una clase una vez que se ha hecho al proceso de clasificación. En el sentido estricto convencional ubicar un objeto en una clase es una clasificación individual.

Taxonomía: Estudio teórico de la clasificación, incluyendo sus bases, principios, procedimientos y reglas. Es el producto del proceso taxonómico, la clasificación numérica [Simpson].

En contraste [Blackwelder] se entiende como el proceso diario de la práctica dedicada a encontrar especies de objetos.

Esta incluye manejo e identificación de especies, publicación de datos, estudio de la literatura y el análisis de la variación de las especies. Este procedimiento no tiene significado teórico importante y los grupos que se forman se denominan TAXON (plural TAXA) [H.J.LAM].

El término Taxon se aplica a un grupo de objetos considerado como unidad de cualquier rango en un sistema clasificatorio.

La **taxonomía numérica** es el agrupamiento de unidades taxonómicas por métodos numéricos en TAXONES (TAXA) en base a los estados de sus caracteres.

La información sobre entidades taxonómicas deben ser transformadas en cantidades numéricas.

Esto incluye las "tiradas" de las inferencias filogenéticas de estudios estadísticos u otros matemáticos, en toda su extensión.

Algunos autores tratan de extender la denominación por el desarrollo de los métodos numéricos de estimación de relaciones cladísticas, pero se mantiene el anunciado.

2.3.4. ESTADO OPERATIVO

Es el resumen que encierra la era neo-Adansoniana (Michel Adanson), que se supone origen de la taxonomía y las correcciones de Sneath [Peter Sneath 200 años después].

1. Cuanto mayor sea la cantidad de información contenida en los TAXONES (TAXA) de una clasificación y cuantos más caracteres contenga su base hará que se obtenga una mejor clasificación .
2. A priori, cada carácter tiene el mismo peso en la creación de TAXONES (TAXA) naturales.
3. La SIMILITUD COMPLETA entre cualquier par de entidades es una función de las SIMILITUDES INDIVIDUALES de los caracteres, que están siendo comparados, de las mismas.
4. Los distintos TAXONES pueden ser reconocidos pues las correlaciones de los caracteres difieren en los grupos de objetos en estudio.
5. Las inferencias filogenéticas pueden ser hechas desde las estructuras Taxonómicas de un grupo y desde las correlaciones de caracteres, si se dan como asumidas ciertas trayectorias y mecanismos de la evolución.
6. La Taxonomía es vista y practicada como una ciencia empírica.
7. Las clasificaciones están basadas en similitudes fenotípicas.

SECUENCIA

1. Los objetos y los caracteres son encontrados y registrados.
2. Se calculan las semejanzas (similitudes) entre objetos.
3. Las semejanzas (similitudes) definen los TAXONES (TAXA).
4. Se realizan generalizaciones sobre los TAXONES mediante inferencias fenotípicas.

GENERALIZACIÓN

1. Una generalización no puede ser realizada antes que los taxones sean reconocidos.
2. Los taxones no pueden ser reconocidos antes que la semejanza (similitud) entre objetos sea conocida.
3. La semejanza no puede ser estimada antes que los objetos y sus caracteres han sido examinados.
4. Los pasos pueden ser combinados en un procedimiento computacional pero el orden no puede ser cambiado pues se destruye el producto clasificatorio.

2.3.5. ESTIMACIÓN DE LA SEMEJANZA

Estimar la similitud es un paso importante y fundamental en la taxonomía numérica. Comienza con la recolección de información sobre caracteres del grupo taxonómico en estudio.

La información puede existir o debe ser descubierto y para que el método sea confiable deben ser tenidos en cuenta muchos caracteres.

Toda clase de caracteres son igualmente deseables: morfológicos, fisiológicos, etológicos y hasta distributivos. Uno debe tener cierta reserva solo de introducir un sesgo de caracteres o caracteres insignificantes.

A priori seguimos nuestra afirmación de que todos los caracteres tienen el mismo peso.

Hay una variedad de caminos a seguir para medir la similitud.

En general se la representa mediante coeficientes de similitud que pueden valer entre 0 y 1. El último si son iguales o acordes y el primero por ninguna desigualdad o disimilitud cualquiera sea. O por coeficientes de disimilitud o distancia usualmente en un rango entre cero y un valor positivo cualquiera indefinido, el primero para la identidad y el último para la máxima distancia o disparidad.

Los coeficientes de similitud son estructurados (tabulados) en forma matricial con un coeficiente por cada par de entidades taxonómicas.

Si una matriz simétrica se construye a partir de t entidades (en forma especular) el resultado es una matriz de t x t.

Las similitudes entre entidades taxonómicas pueden ser representadas geométricamente por puntos en el espacio. Un máximo de t-1 dimensiones son necesarias para una correcta representación de t puntos (t entidades) en el espacio. Las distancias entre puntos del espacio son observables como distancias taxonómicas.

2.3.6. CONSTRUCCIÓN DE TAXONES (TAXA)

La clasificación en taxonomía numérica está generalmente basada en una matriz de semejanza, en la cual los taxones son construidos mediante varias técnicas designadas para descubrir o exponer y resumir la estructura de la matriz.

Un bosquejo, una representación gráfica de la estructura de la matriz, puede ser obtenido mediante el sombreado de varios elementos diferenciados por un previo agrupamiento, tal que supuestamente las formas similares están cerca una de otras.

Si, como metodológicamente es preferible, las entidades son colocadas en la matriz sin un orden preestablecido, los agrupamientos no son visibles sin un arreglo de las mismas.

El procesamiento de datos para formar agrupamientos (clustering) es metodológicamente tan eficiente estando ordenados o no. Como los métodos imponen un uso numérico de evaluación son preferibles y se denominan análisis de agrupamiento (cluster analysis) y se aplican sobre la matriz de similitud (t x t) y es parecido a un contorno de un mapa topográfico de valles y picos.

Los agrupamientos (clusters) están generalmente basados en similitudes fenotípicas y no necesariamente deben tener connotaciones filogenéticas.

Las diferencias metodológicas se refieren a las reglas de formación de agrupamiento y el particionamiento de objetos en el espacio taxonómico.

La importancia de los métodos es que tienen en común una forma de delimitar los grupos taxonómicos que es objetiva dada una matriz de coeficientes de relaciones.

Las fronteras entre los grupos taxonómicos pueden ser visualizadas a partir de la línea de contorno del mapa de las líneas diagonales y sus intersecciones con las ramificaciones de los árboles comunes en los diagramas de relaciones empleados en taxonomía numérica.

Fronteras y divisiones en diagonal permiten un mayor o menor grado de refinamiento de los grupos taxonómicos.

Un tamaño amplio o pesado o una dispersión de los grupos (splitting) hacen que además debamos proporcionar un objetivo o un criterio de cohesión a los taxones (Taxa).

Esto depende fuertemente del método de agrupamiento (cluster) y de hecho de la estabilidad de la clasificación afecta de dos maneras:

- Es posible acumular mas información (en la constitución de nuevos caracteres). Si la evaluación inicial de la similitud ha sido hecha en base a muestras de un gran número de caracteres, se entiende que las similitudes relativas cambian muy poco al agregar caracteres posteriormente.
- Es posible incluir nuevas entidades taxonómicas en estudios posteriores. La utilización de criterios previos para niveles y número de divisiones diagonales pueden producir nuevos y diferentes taxones (taxa). Se ha llegado a la conclusión que no se producirán nuevos niveles de fronteras en los grupos. En general las categorías no tienen gran implicancia en el estudio.
- Los fenotipos son un intento de una terminología más natural.
- Un aspecto importante es la representación de taxones en diagramas de árbol indicando las relaciones fenotípicas.
- Las representaciones de relaciones taxonómicas en modelos de grafos de dos o tres dimensiones, hiperespaciales, conduce a patrones de estructuras taxonómicas de mucho interés.

2.3.7. IDENTIFICACIÓN DE ESPECÍMENES

Una vez que una clasificación ha sido establecida por algún método operacional, el siguiente y obvio paso, es el de encontrar claves apropiadas de identificación de objetos por computadora. Esto se logra obteniendo caracteres cuyo peso permita una identificación mediante algún criterio de diferenciación. Si estas estructuras de caracteres no es posible determinar hay que encontrar algunas estructuras alternativas. Sin embargo las claves cambian muy poco dada una lógica de construcción, solo una adaptación al procesamiento por computadora. Las formas interactivas son muy útiles en la construcción de claves no tan sofisticadas como las probabilísticas.
La evolución es a formas inteligentes como las redes neuronales.

2.3.8. PRINCIPIOS TAXONÓMICOS

La taxonomía tradicional inclusive la post Darwiniana evolucionó en conceptos y procedimientos.

Las nuevas formas sistemáticas [Huxley, 1940] [Sokal y Sneath, 1973] y los avances en la genética, la citología y la variación geográfica condujo a considerables progresos en el entendimiento de los mecanismos de la evolución de las especies y de las infraespecies, pero constituyó poco en la comprensión de la naturaleza y evolución de las mas altas categorías y de la estructura taxonómica en general.

Son algo más que simples generalizaciones descriptivas (se intentó hacer todas y no se hizo ninguna bien).

Las funciones que se intentaron son (1) clasificar (2) nominar (3) indicar el grado de semejanza y (4) mostrar relaciones condescendientes. En forma convencional estas funciones no pueden hacerse simultáneamente.

2.3.9. APROXIMACIONES EMPÍRICAS Y OPERACIONALES

Los esfuerzos hechos a partir de asumir las bases filogenéticas, en forma sistemática, fueron un corsé para las observaciones taxonómicas y sus bases clasificatorias, que en realidad sirvieron para la descripción de patrones variacionales que de hecho existen en la naturaleza y conceptos e ideas de todos los niveles.

El análisis empírico de datos taxonómicos condujo naturalmente a una aproximación operacional hacia la taxonomía.

Un contexto operacional implica que las sentencias y hipótesis sobre la naturaleza debe ser probado mediante observaciones y experimentos. Se deben establecer criterios para definir categorías y operaciones, para no caer en discusiones científicas sin sentido.

Se debe tener un grado de certeza respecto a que los datos, con los cuales trabajamos, están sujetos a operaciones lógicas definibles, estas operaciones, en las cuales estamos interesados, llevadas a cabo para responder preguntas sobre la materia, pueden ser comunicadas en forma no ambigua a otras personas inteligentes tanto como máquinas que permitan manejar una lógica y hacer cómputos. Estos principios son un gran avance en la taxonomía a modo de guía.

La idea fuerza es que las mediciones y operaciones, siempre y cuando no estén mal definidos sus conceptos o sean vagos, sirven para evolucionar teóricamente.
El hecho es no realizar categóricamente conceptos no operacionales. La taxonomía operacional y empírica, procedimientos operacionales usados durante la clasificación y observaciones empíricas de datos taxonómicos, no son congruentes completos.
La taxonomía numérica como es generalmente aplicada es tanto operacional como empírica.
Es deseable que sea solo operacional.

2.3.10. SISTEMA NATURAL

En la metodología sistemática la clasificación natural es la basada en la naturaleza de las cosas, en oposición a sistemas artificiales o arbitrarios (correlaciones entre caracteres o filogenéticas).
La lógica Aristotélica es una metodología sistemática basada en axiomas, usada para intentar descubrir la esencia de los grupos taxonómicos, que da lugar a propiedades. Estos sistemas lógicos son conocidos como sistemas de entidades analizadas. En contraposición existen sistemas de entidades no analizadas pues sus propiedades no pueden ser inferidas por sus definiciones en si mismos.
Los géneros y especies son grupos taxonómicos de base lógica.
Solo un carácter puede diagnosticar un grupo taxonómico siendo taxones (taxa) de cualquier rango.

Son formas monotípicas o monotéticas.
Las ideas rectoras de grupos monotéticos o monotípicos es que están formados por sucesivas divisiones lógicas que tienen un conjunto característico único de atributos, el cual es condición necesaria y suficiente para definir un miembro.
Todos los miembros del grupo tienen el mismo conjunto característico, por eso son monotípicos.

En los grupos monotéticos naturales se corren riesgos de mal clasificación desde el punto de vista fenético, pues el estado de un carácter puede producir que un organismo quede totalmente descolocado del taxon.
Las formas politéticas o politípicas son agrupamientos donde los miembros son colocados juntos cuando el estado de los caracteres o atributos son compartidos no todos y un solo estado no es suficiente o esencial para un miembro o un grupo, salvo los atributos que hacen que pertenezcan al grupo.

2.3.11. DEFINICIÓN FORMAL DE BECKNER

Una clase esta comúnmente referida a un conjunto de propiedades, por principio, necesarias y suficientes para cada miembro de la clase. Es posible definir un grupo $\underline{K}$ en función de un conjunto G de propiedades f_1, f_2,...,f_n de manera diferente. Supongamos que tenemos un agregado de elementos, que aún no denominamos clase, tal que:

1. Cada uno posee un gran número (inespecífico) de propiedades de *G*.
2. Cada *f* de *G* es poseído por cada elemento en el agregado y
3. Ningún *f* de *G* es poseído por cada elemento en el agregado.

En término de (3), ningún *f* es necesario por miembros de ese agregado y nada debe ser dicho para garantizar o desechar la posibilidad de tener algún *f* de *G* para ser un miembro del agregado.
Una clase es politética si cumple las dos primeras condiciones y es politética completa si cumple también la (3).
Los grupos taxonómicos son clases politípicas, pero los conceptos politípicos no implican taxonomía.
En general los taxones no son politípicos completos porque siempre se encuentra estados de caracteres que son comunes (caracter de estado alelo).
Existen alelos idénticos en todos los miembros de un género.
En la genética existe un lugar geométrico de los genes de una población que se mantiene invariante con la evolución, limitado por mecanismos de reemplazo.
Si las *f* individuales no son compartidos por los miembros entonces no se producen clases politéticas, de allí la importancia de la sentencia (2).

El número de objetos no es una constante de una clase sino sus atributos, es decir que el número no implica clase y si agregamos un objeto no cambia la clase sino sus propiedades.
Un cambio de propiedades cambia de clase y un objeto pertenece a otra clase.
El axioma (3) no solo dispone que los objetos idénticos no pueden ser distribuidos en clases diferentes, sino que también establece un criterio para las clases unitarias.
El problema de clasificar es tener un objetivo clasificatorio, pues una clasificación puede ser irrelevante.
¿Cuál es el elemento de juicio que se utiliza para elegir una clasificación entre varios posibles de un conjunto de objetos?
La respuesta es:

'Comprensión profunda de sus propiedades, semejanzas, diferencias e interrelaciones'.

Una clasificación es superior cuando más leyes científicas y contribuyan más a la formulación de hipótesis explicativas.

Principio organizador del conocimiento: más estable, más robusta y más predictiva.
Estable: no hay modificación drástica por incorporar nueva información.
Robusta: no es alterada por la incorporación de nuevos objetos.
Predictiva: todo nuevo objeto tiene las mismas propiedades de las entidades del grupo.

ASPECTOS FILOSÓFICOS

Hay cuatro esquelas que son doctrinarias en la clasificación: Esencialismo, Cladismo, Evolucionismo y Feneticismo.

Esencialismo. (como ya se expuso Aristotélico).

1. La esencia de los objetos (naturaleza o forma) existe y puede ser descubierta y discriminada.
2. La tarea de la clasificación es descubrir y discriminar las esencias.
3. Las esencias tienen nombre y descripción es decir se las define.
4. Los objetos reflejan una serie de tipos básicos y formas inmutables.
5. Todos los objetos de un taxon tienen la misma naturaleza esencial y el mismo tipo básico.
6. Toda variación dentro de un mismo taxon es desechable a los fines clasificatorios, por ser el producto de una desviación de los arquetipos básicos.

También se la denomina escuela tipológica. Esto implica que la clasificación se "descubre", no se "construye".

Cladismo. (Ramas de un árbol genealógico)

Se basa en la filogénia o historia evolutiva (no evolucionismo).

1. Todo taxon debe ser monofilético, es decir los objetos del grupo tienen entre si el mismo antecesor reciente, más reciente que los de otro taxon de igual rango.
2. Se eligen caracteres en los cuales pueda determinarse el estado primitivo (en el estado presente es el más reciente antecesor). Se pueden estudiar el grupo y grupos afines.
3. Se debe establecer la secuencia de las ramificaciones del árbol genealógico y la posición relativa en el tiempo de esas ramificaciones. Se usan caracteres elegidos y fósiles.
4. Los grupos se forman por monofiletismo y antecesor común de cercanía temporal. Estados evolucionados recientes evitan tener distorsiones a nivel temporal velocidades de evolución diferentes de los caracteres.
5. La categoría taxonómica conforme a la jerarquía linneana (Carlos de Linneo) asignada a cada grupo está en relación directa con su desprendimiento en el tiempo de otra línea evolutiva (ramificación). Cuanto más antiguo más alto el rango jerárquico.

Evolucionismo. Enfoque ecléctico entre genealogía y similitud en nivel jerárquico o rango.

1. Filogénia.
2. Cantidad de modificaciones evolutivas desde el antecesor. Determinar el grado de diversificación y el tamaño de los taxones a formar. Es decir el monofiletismo y la similitud entre ellos.
3. Homogeneidad interna, respecto a la similitud, en cada infra-taxon.
4. Asociación del taxon y el entorno.
5. El rango adjudicado a un taxon debe estar asociado a un cardinal considerable salvo que la distorsión en similitud sea muy grande.
6. Debe existir equivalencia entre jerarquías entre taxones afines.

Los autores le dan distintos pesos a cada ítem.

Feneticismo. Calificación de objetos por sus atributos.

1. Las clasificaciones deben efectuarse teniendo en cuenta un gran número de caracteres descriptivos y de su ciclo de vida.
2. Todos los caracteres tienen la misma importancia en la formación de taxon.
3. La similitud total (global) entre dos objetos es la suma de las similitudes en cada uno de los caracteres utilizados en la clasificación.
4. Los grupos de taxones a formar se reconocen por una correlación de caracteres diferentes.
5. La clasificación es una ciencia empírica, libre de genealogías y solo es válida la experiencia.
6. La clasificación debe basarse solo en la fenética. Incluyendo por lo tanto cualquier tipo de carácter.
7. El número de taxones encontrados es arbitrario aunque coherente con los resultados obtenidos.

Una vez obtenido un criterio de delimitación debe seguir aplicándose el mismo criterio al taxon.

Hay opiniones que expresan que el origen de la Fenética se debe a Adanson a pesar de que no tiene en cuenta la filogenética, aunque una vez elaborada una clasificación se pueden hacer estudios relativos a la misma.

Empirismo, Operacionismo y Feneticismo

El empirismo [John Locke, siglo XVII] epistemologicamente indica que todo conocimiento depende de la experiencia y toda teoría debe verificarse experimentalmente [Bacon, F. 1605-1620]. Al aplicarlo en la clasificación no filogenética el empirismo lleva al feneticismo y salvo casos no verificables, puede aplicarse el operacionismo fundado por P.W. Bridgman [The logic of Modern Physics 1927] [Sokal y Sneath, 1973] [Hempel, C.G. 1996].
Consiste en la descripción de operaciones que conducen a establecerlo. Un símbolo adquiere interpretación operacional y operativa si se establece una correspondencia entre él y los resultados de una operación real o posible, arbitrada para medirlo [Bunge, M. 1969]
En el Feneticismo la preponderante es el operacionismo aunque no se descarta alguna excepción [Sokal y Sneath, 1973]
Siempre debe existir una enumeración de pasos a seguir para completar las operaciones de observación y medición.

Relaciones Taxonómicas

1. Relaciones Fenéticas o de similitud.

- Parecido o Semejanza entre objetos (sin tener encuentra su genealogía). Se basan en los atributos observados y se expresan, estas relaciones, como proporciones de las similitudes y diferencias entre objetos [Sneath, 1978].

1. Relaciones cronísticas o temporales.

- Indican el grado de cercanía en el tiempo de dos o más objetos.

1. Relaciones de parentesco o filogenéticas.

- Se basan en la genealogía e indican el grado por el cual dos o más objetos están relacionados por un antecesor común y el grado de cambio evolutivo que ha ocurrido desde el mismo.

1. Relaciones espaciales o geográficas.

- Denotan el grado de situación espacial relativa en dos o más objetos.
- No hay única respuesta salvo si las relaciones son congruentes en el espacio, el tiempo, la similitud y el parentesco.
- Solo se puede decir que hay preponderancia de alguna relación respecto a otras.

DESCRIPCIÓN DEL PROBLEMA

Nuestro ideal no llega a las estrellas / es sereno, sencillo;/...

Federico García Lorca

CXIII

3. DESCRIPCIÓN DEL PROBLEMA

Al conceptualizar la problemática, surge un programa de investigación científica (PIC) como sucesión de teorías emparentadas semántica y sintáticamente, que se van generando en distintas disciplinas por observaciones intrigantes, que se captan históricamente y llaman la atención pues se comportan en forma desconcertante o funcionan de una manera diferente a la esperada, constituyendo familias de fenómenos intrigantes.

En el conjunto de los problemas uno de los problemas fundamentales no resuelto es la clasificación de objetos en familias o clusters mediante un método automático de formación de regiones, el cual es un proceso de agrupamiento de objetos en clases teniendo en cuenta sus relaciones y atributos comunes, a los efectos de realizar estudios de sus características y las propiedades estructurales y de su comportamiento relativo.

La originalidad de la solución surge primero de la analogía entre la representación de entidades de bases de datos dinámicas y las unidades operacionales taxonómicas (OTU´s o taxones) que aparecen en las matrices de datos y similitud, en el proceso clasificatorio de agrupamiento de la taxonomía numérica al evaluar la afinidad y similitud, aplicando luego, en forma original, la espectroscopia para representar a los objetos mediante sus espectros característicos y obtener sus familias o clusters al aplicar los principios de superposición y de interferencia a las regiones encontradas en forma automática utilizando invariantes surgidas del proceso de clasificación basado en las propiedades de los OTU´s bajo estudio.

Las regiones encontradas surgen de un proceso de cómputo automático a partir del cálculo de invariantes tales como el centro de masa o gravedad, rango, centroide y radio del teorema de Tchebycheff y de la desigualdad de Bienaymé-Tchebycheff, tratadas en este tema en forma original.

Las bases de datos relacionales dinámicas surgen en la evolución de las bases de datos y aplicando el método de contrastación como parte del método científico de investigación, estructurando el par <atributo,valor> en dominios integrados dinámicamente y hacia las bases de datos del paradigma de la programación orientada a objetos y en analogía original con los caracteres de los

OTU´s basados en el estado de sus caracteres, como resultante de estudios teóricos de los procesos clasificatorios de la taxonomía que incluyen sus principios, procedimientos y reglas.

La otra originalidad que aparece es que la taxonomía convencional encuentra las regiones "a ojo" cuasi arbitrariamente, para resolver problemas empíricos que su metodología no resuelve, ya que en la así llamada evidencia taxonómica de similitud y afinidad de los objetos, las estructuras de esas formas fenotípicas de los dendrogramas/fenogramas, agrupamientos, no surgen de los procesos de tecnología informática (IT) o taxonomía computacional automática. Pero al tratar con espectroscopia a los objetos usando las invariantes en el hiperespacio taxonómico u otras formas no fenotípicas aunque los llamen dendrogramas no lo son, y tomando además conceptos de entropía de la teoría de la información se logra resolver en forma automática y no arbitraria el problema.

No aparecen clusters solapados ni problemas de isotropía y homogeneidad al resolver los problemas de dilatación del espacio y conservación de la masa.

Finalmente se utiliza Data Mining Inteligente, para verificar la fortaleza del método encontrado, utilizando las familias de árboles de inducción y la ganancia en información que puede calcularse como la disminución en entropía. El sistema generará un árbol de decisión fruto de la naturaleza en sí de los algoritmos de la familia TDIDT, que al podarse permite calcular el error del ejemplar clasificado.

SOLUCIÓN PROPUESTA

"La noción de un Diseño Inteligente está vertiginosamente reemplazando la evolución Darwinista como el principio central de la ciencia biológica. …¡viene una revolución! "

Jonathan Wells, Biólogo Postdoctoral y Científico principal del Instituto Discovery, Seattle.

CXVI

4. SOLUCIÓN PROPUESTA

ESPECTROS DE EVIDENCIA TAXONOMICA EN BASES DE DATOS DINAMICAS

En este capítulo se presentan todos los aspectos de la solución propuesta. Para ello se describen las características generales de un Nuevo Criterio para resolver el problema de la construcción de familias de objetos, para lo cual se constituye un programa de investigación científica (PIC) como sucesión de teorías emparentadas de manera semántica y sintáctica, que se generan en distintas disciplinas por observaciones y fenómenos intrigantes, dado el problema describir el marco teórico. La Taxonomía Numérica (comienzo del Capítulo) tiene por objetivo agrupar unidades taxonómicas operacionales en clusters (OTU´s o taxones o taxa), usando el análisis de estructura por medio de métodos numéricos. Estos clusters, que constituyen familias, son el propósito de esta tesis y de mis proyectos, y surgen del análisis estructural, por su característica fenotípica (Secciones 4.2., 4.2.1-4.2.6.). La Entidades formadas por dominios dinámicos de atributos, pueden cambiar de acuerdo, a los requerimientos taxonómicos: Clasificación de objetos para formar familias o clusters. Los objetos Taxonómicos son representados mediante la aplicación de la semántica del modelo de Base de Datos Relacional Dinámica (Secciones de 4.1.). Exhibiendo las relaciones en lo que se refiere a las calidades de similitud de los OTU´s, al emplear herramientas como, distancias Euclídeas y técnica de vecinos más cercanos. Así la evidencia taxonómica recoge, para cuantificar, la similitud de cada par de OTU´s (método pair-group) obtenido de la matriz de datos básica y de esta manera el concepto, importante y principal, de espectro de OTU´s es introducido, por el principio de interferencia y superposición, tomando como base el estado de sus caracteres (Sección 4.2.6.). El concepto de los espectros de familias surge, si el principio de superposición se aplica a los espectros de los OTU´s, y los grupos se delimitan a través del teorema de Tchebycheff y del máximo de la inecuación de Bienaymé-Tchebycheff, que determina Invariantes (el centroide, varianza y radio) (Sección 4.2.7.), junto con la normalización del rango y el principio de entropía máxima (Sección 4.2.8.) y el algoritmo capaz de generar las familias (Sección 4.2.9.). Un nuevo criterio taxonómico es así establecido por la Tesis, surgiendo una Aplicación que ha funcionado. Así se logra un mayor acercamiento a la Taxonomía Computacional y se presenta como explicación científica, ya que además, ha sido empleado con referencia a Minería de Datos (Data Mining), cuando se aplican técnicas de Machine Learning, en particular a los algoritmos de C4.5, creado por Quinlan, el grado de eficacia logrado por los algoritmos de la familia de TDIDT cuando genera modelos válidos de datos en los problemas de clasificación con la Ganancia de Información a través del Principio de la Entropía Máxima.

La Taxonomía Numérica, ha sido una disciplina definida como la evaluación numérica de la afinidad y similitud entre unidades taxonómicas y el agrupamiento de esas unidades en taxones, basado en sus caracteres. Permite agrupar, a través de métodos numéricos llamados análisis de clusters, unidades taxonómicas operacionales (OTU´s) en taxa (grupos de taxones u OTU´s). Los clusters constituyen familias mediante un análisis estructural basado en su característica fenotípica que muestra la relación, en grado de similitud, entre dos OTU´s o grupos de OTU´s [Sokal y Sneath, 1973] [Crisci y López Armengol, 1983] [Stepp, 1987] [Gennari, 1989 a, b] [Hanson y Bauer, 1989] [Jiménez Rey, Grossi, 1996].

Las OTU´s toman valores de los dominios dinámicos de atributos que forman entidades que van cambiando de acuerdo a las necesidades taxonómicas: clasificar para formar familias o clusters.

Aplicando la semántica del modelo de datos de Bases de Datos Relacionales Dinámicas se representan los objetos taxonómicos.

Las familias de OTU´s que se asocian por su grado de similitud, se obtienen mediante la distancia Euclídea y la aplicación de técnicas de *"vecinos próximos"*, lográndose la fuente de la **evidencia taxonómica** al cuantificar, a partir del coeficiente de similitud, la semejanza para cada par de OTU´s de la matriz básica de datos.

Considerando la clasificación como el proceso de agrupamiento de objetos en clases teniendo en cuenta sus atributos comunes (caracteres, propiedades, etc.) y relaciones, la taxonomía es el estudio teórico de la clasificación incluyendo sus principios, procedimientos y reglas.

La Taxonomía Numérica obtiene, usando operaciones matemáticas, la afinidad entre unidades taxonómicas basada en sus caracteres (atributos, propiedades, relaciones, etc.)

Los objetos a ser clasificados se denominan unidades taxonómicas operacionales. Los objetos se clasifican según un proceso basado en las propiedades de los mismos, a los que llamamos (OTU´s). La diferencia entre ellas es la fuente de la evidencia taxonómica.

Un carácter podría ser definido como cualquier propiedad que caracteriza la OTU en estudio. El conjunto de valores posibles de este carácter se llama estado.

Para estimar la similitud taxonómica usada para agrupar las OTU's tratamos de expresar esta semejanza en una manera cuantitativa. Usamos el coeficiente de similitud para cuantificar esta semejanza, es decir, para obtener la semejanza para cada par de OTU's en la matriz básica de datos.

Conceptualmente aplicaremos los conceptos expresados en las “Bases de Datos Dinámicas y su Relación con la Taxonomía” desarrollados en los últimos años [Perichinsky, G. et Al., 1989-2007].

En esta aproximación conceptual en forma similar que en el SQL, las tablas son, en sentido estricto y dinámico, "visiones" con leyes de formación que surgen de la lógica de la aplicación.

Por ello hablamos de visiones y decimos "dinámicas", pues se pueden agregar o eliminar columnas-dominio de una tabla virtual y por supuesto se pueden modificar y eliminar valores de atributos, tanto como aumentar el cardinal de un dominio.

De esta manera todo objeto o entidad de una aplicación puede mejorar dinámicamente su calificación e identificación. Se alcanza así una gran independencia tanto física como lógica de los datos, y una dinámica en el crecimiento o expansión.

La matriz de datos queda representada estructuralmente mediante dominios de atributos (caracteres) dinámicamente integrados.

La aportación original en esta tesis es proponer el concepto de espectro de los estados de los caracteres de los pares de OTU´s respecto al total, el espectro de familias, por el principio de superposición al procesar los espectros de los pares de OTU´s y la obtención de Invariantes (centroide, varianza y radio).Se logra un algoritmo más eficiente por un mejor tratamiento matricial. Complementariamente se desarrolla una teoría de bases de datos dinámicas para dar soporte al proceso de clasificación basada en espectros de evidencia taxonómica.

4.1. BASES DE DATOS DINAMICAS

Teniendo en cuenta los trabajos que he presentado en los últimos años con colaboradores [Perichinsky,G., 1994], y el panorama desarrollado en este trabajo, he demostrado que ha surgido un nuevo enfoque en la Investigación en Bases de

Datos [de Miguel,A., 1993] [de Miguel,A., 2000], que se apoya sobre una Base de Datos en la cual los datos se almacenan una sola vez, con independencia de su tratamiento, en sistemas orientados hacia los datos y se ha estabilizado conceptualmente, en los Modelos Relacionales estructural y semánticamente y el program embedding SQL [ISO 9075] que son las formas estandarizadas de los próximos años. El modelo trata a los dominios en forma independiente, por lo tanto la estructuración es más natural por la forma de agregación de las tuplas y una semántica n-aria de atributos. Con el predicado de operaciones conjuntistas se forman tablas virtuales o visiones. La arquitectura que se experimenta para el gerenciador es la propuesta de tres niveles [ANSI/X3/SPARC].

4.1.1. CONCEPTOS

Esta propuesta consiste en el desarrollo teórico-conceptual y de implementación de un sistema de base de datos relacional estructurado sobre dominios dinámicos de atributos [Perichinsky,G., 1989-2007].

El nuevo enfoque se trata de un mayor nivel de abstracción para conseguir el máximo de independencia lógica posible [(ANSI).1988], que es en la cual el gerenciador tiene la capacidad de que las referencias a los datos almacenados, especialmente en las aplicaciones y en sus descripciones de los datos, estén aislados de los cambios y de los diferentes usos en el entorno de los datos, como pueden ser la forma de almacenar dichos datos, el modo de compartirlos con otras aplicaciones y cómo se reorganizan para mejorar el rendimiento del sistema de base de datos. Este nuevo enfoque hace que los dominios de los atributos sean dinámicos y conformen realmente la base del tratamiento de la teoría de conjuntos, y que las tuplas se generen dinámicamente a través de visiones [Date, 1981-1992].

Se trata de un diseño que cambia la estructuración tradicional de agrupamiento estático de valores de atributos en registros por la creación de dominios de atributos, formando conjuntos de valores de los mismos. Las tuplas (virtuales) se forman mediante las relaciones que como las visiones no existen, las tablas están establecidas a partir de dominios [Date,C.J., 1992.Data on Databases].

No se trata de una estandarización pues este concepto frena los desarrollos futuros, esto exige cautela pues el dilema es, que la estandarización orienta a los diseñadores [Codd, 1970-1990].

La arquitectura ANSI/X3/SPARC tiene una técnica de diseño por niveles o máquinas anidadas y el flujo de datos pasa por las distintas capas, que están separadas por interfaces, cuyo número marca de alguna manera la capacidad de independencia.

Un diccionario de datos permite la estructuración del conjunto de datos o metadatos.

El tradicional esquema conceptual envuelve esta capa con una interfaz con el diccionario en la metabase de datos.

Las estructuras externas e internas forman parte de capas separadas por las interfaces 4 y 5. Cada uno tiene un administrador en la interfaz 3 se puede tener un conjunto de menús, que se utilizará para el administrador de la Base.

La manipulación de la Base de Datos se hará con SQL embbeding con un motor C y C++ [Staugaard, 1998].

Una operación se ejecuta mediante transformadores conceptual/externo, interno/conceptual y almacenamiento/interno que utilizan los metadatos mediante las interfaces (binding).

Como no se especifica la instrumentación, se permite que:

- **el sistema evolucione rápidamente**
- **la utilización sea óptima**
- **se logre independencia lógica**
- **posibles reestructuraciones.**

En forma similar que en el SQL, las tablas son, en sentido estricto y dinámico, "visiones" con leyes de formación que surgen de la lógica de la aplicación.

Por ello hablamos de visiones y decimos "dinámicas", pues se pueden agregar o eliminar columnas-dominio de una tabla virtual y por supuesto se pueden modificar y eliminar valores de atributos, tanto como aumentar el cardinal de un dominio. De esta manera todo objeto o entidad de una aplicación puede mejorar dinámicamente su calificación e identificación. Se alcanza así una gran independencia tanto física como lógica de los datos, y una dinámica en el crecimiento o expansión (hasta en comportamiento) [Perichinsky,G., 1994].

4.1.2. MODELIZACIÓN

El modelo surge del par M=<S,O> donde S son las reglas y O las operaciones sobre objetos permitidos.

Las instancias determinan la dinámica del modelo.

El debate actual es extender la Bases de Datos Relacionales hacia la orientación a objetos [Third Generation Database System Manifesto. Carey et al.1990] y los puristas del modelo orientado a objetos [The Object-Oriented System Manifesto. Atkinson et al.1990], después [Staugaard, 1998.

Ante esta alternativa es preferible pensar en el avance teórico que representa la orientación a objetos y por lo tanto aplicarla de acuerdo a los requerimientos del diseño; por ejemplo una capa de nivel externo que dé la apariencia de objetos, sobre un modelo relacional [de Miguel,A., 2000].

Verdaderamente se obtienen grandes ventajas con este modelo, ya que lo expuesto implica una reestructuración que no depende de los datos sino de las aplicaciones. Las diferentes aplicaciones pueden "ver" a los datos de acuerdo a sus requerimientos y modos.

Se simplifica la visión de los usuarios. Los dominios de los atributos son la base del modelo y de la agregación lógica de estos, mediante operaciones y formas algebraicas relacionales, surge la estructura.

Las expresiones del álgebra relacional sirven a los propósitos de mantenimiento, actualización y recuperación de la información de los dominios, a través del manejo de tuplas, y preservando su homogeneidad e integridad. Al operar sobre dominios, conjuntos de valores de atributos, se tiene la formalización concreta de la teoría de conjuntos.

Los valores de los dominios son atómicos y los atributos se califican por aplicación; esto permite operar en formas normales.

La capa externa propuesta anteriormente para tener la apariencia de objetos sería una aplicación más desde el modelo conceptual.

Lo mismo ocurre con una herramienta CASE Inteligente, o una capa de nivel inteligente para aprendizaje automático.

La TERCERA generación de Bases de Datos tendrá más que ver con nuevas capas de nivel de contenido semántico para tener en cuenta las Bases de Datos

Distribuidas, Inteligentes y con Orientación a Objetos en un entorno abierto, heterogéneo y distribuido [Informe Lagunita, Silberschatz,A., 1990].
Las FAMILIAS de modelos que surgen a partir del modelo de Chen de entidad-relación ER [Chen,P.P., 1976] y [Bachman,C.V., 1974] y sus extensiones [Teorey,T.J., Fry,J.P., 1982], [Shan y Shixuan, 1984] y [Elmasri,R., Navathe,S., 1989], [Luque Ruiz, I., Gómez-Nieto, M. A., et al. 2002] tratan los tipos de entidades y sus relaciones, a partir de la estructuración estática de los atributos, asociadas a un predicado lógico.

4.1.3. DINÁMICA

El contenido semántico de las relaciones se ha completado conceptualmente con la cardinalidad, la dependencia en existencia y en identificación y la abstracción de generalización.
La cardinalidad tiene que ver con la cantidad de instancias que tienen las entidades relacionadas algebraicamente, y sus correspondencias según los axiomas de Armstrong [Wiederhold,G. 1983].

4.1.4. GENERALIZACIÓN

Las jerarquías de entidades hace que existan subtipos (especialización) y supertipos (generalización), que se corresponden con la noción de "es-un" o "es-un-tipo-de", originado en la inteligencia artificial extraído de las redes semánticas [Quinlan, 1986], pueden cubrir parcial o totalmente. En nuestro modelo por operaciones del álgebra relacional tales como solapamiento, disyunción, parcial y total [Davis, 1990].
Esto constituye una verdadera red de generalización, donde aparece la herencia simple y múltiple, que queda a criterio de las aplicaciones, pudiéndose producir conflictos en otros modelos por complejidad operativa (join, proyección y selección) recordar la dependencia respecto al lenguaje.

Dentro del marco desarrollado y las propuestas estandarizadas se tienen seis proyectos [ISO/IEC]:

1. El Marco Conceptual para el sistema basado en DRI (IRDS).
2. Interfaces de Servicios (SI).
3. Exportación/Importación.
4. Interfaces del Lenguaje de Sentencias.
5. Interfaces de panel del sistema basado en DRI.
6. Soporte del Sistema para el SQL.

La arquitectura del sistema basado en el DRI se basa en el concepto de descripción recursiva que presenta cuatro niveles de datos y tres pares de niveles asociados: nivel de definición de estructuras de atributos, nivel de definición del diccionario y nivel de aplicaciones en forma dinámica.

- Las interfaces con los demás componentes lógicos son:
- Repositorios de herramientas CASE.
- Generadores de Informes o Pantallas.
- Lenguajes de cuarta generación 4GL.
- Programas fuentes de metadatos para el DRI.
- Interfaces amigables para usuarios.
- Métodos de Acceso del Sistema Operativo (manejo de archivo).

El paso siguiente es por lo tanto la construcción de los procesadores de interfaces a partir de un prototipo ya probado.

Análisis de seguridad, de integridad y de confidencialidad.

La aplicación de estos conceptos en los expresados en la nueva concepción para las bases de datos es directa.

4.1.5. CONTRASTACIÓN

La contrastación es parte del método científico [Codd,E.F.,1985] y [Perichinsky,G. et al , 1996a].

Parte estructural

En un modelo relacional una relación sobre dominios $D_1,D_2,...,D_n$ consiste en una cabecera y un cuerpo. La cabecera consiste en un conjunto fijo de atributos $A_1,A_2,...,A_n$ tal que cada atributo A_i corresponde a un dominio D_i.El cuerpo consiste en un conjunto de tuplas que varían con el tiempo donde cada tupla consiste en un conjunto de pares < atributo,valor >=< A_i,v_i > donde v_i es el valor del dominio único D_i asociado con el atributo A_i.

El grado de la relación es "n" y la cantidad de tuplas su cardinal.Tendrá nombre y será real o será una expresión del álgebra relacional y será virtual en términos de otra real.

Una clave en realidad es la parte rectora, subconjunto de atributos, y se la denomina candidato por ser factible como identificador de tupla. Para eliminar su trivialidad deben cumplir estas propiedades:

- Identificación unívoca: No existir dos dominios, filas, con valores iguales de sus atributos.

- No redundancia: no se puede eliminar ningún atributo de la clave sin destruir la propiedad anterior. Esto implica que la clave primaria es mínima. La clave primaria se la elige del conjunto de claves candidatos (no puede ser nula, ni ninguna de sus componentes).

Parte manipulativa

Cada operador del álgebra relacional toma una o dos relaciones y sus operandos, y produce como resultado una nueva relación.

Codd definió originalmente ocho operadores en 2(dos) grupos de 4(cuatro) cada uno.

- El conjunto tradicional de operaciones de unión, intersección, diferencia y producto cartesiano.

- El conjunto de operaciones relacionales de selección, proyección, junta(join) y división.

Integridad

La integridad asegura que la base de datos satisfaga un conjunto de restricciones predefinidas.
En general una restricción nace cuando una relación incluye una referencia a otra:
Diccionario-de-atributos - Dominio - Dominio-Inverso - Lista.
Un valor nulo, indica que ese ítem no aporta información relevante; y dos referencias que son las direcciones de los nodos raíz del dominio de valores del atributo y del dominio inverso o conjunto de listas invertidas correspondientes a cada valor de un atributo. Las referencias internas se disponen ordenadas.

4.1.6. ANÁLISIS DE REQUERIMIENTOS

Codd propone 12 reglas básicas, 9 estructurales, 18 de manipulación y 3 de integridad.

1. Reglas básicas

REGLA 0 - BASICAS.
Cualquier sistema debe tener algún soporte relacional con capacidad para manejar datos y por lo tanto soportar reglas de información y garantizar reglas de acceso.

REGLA1 - Regla de Información.
Toda información en una base de datos está representada en el nivel lógico por medio de tablas, columnas y dominios nominados como strings de caracteres en el catálogo del sistema.

La regla de Información no solo se debe cumplir para la productividad del USUARIO, sino también para definir paquetes de software en la interfaces con la base de datos.

Estos paquetes recuperan información y agregan información en el catálogo y hacen que se mantenga la Integridad.

En el nuevo enfoque se mantiene un catálogo para la integridad pero sus objetos son virtuales.

REGLA 2 - Regla de acceso.

Cada valor atómico debe ser accedido lógicamente por nombre y valor de atributo.

Esta regla, que no es trivial en el modelo estático, es la base del nuevo enfoque y depende de la aplicación y la ubicación estructural del valor.

REGLA 3 - Valores no existentes.

Para mantener la integridad de la base de datos, debe ser posible especificar que los NULOS no son permitidos en columnas con claves primarias u otras columnas que el Administrador considere apropiado una restricción de Integridad (v.g.columnas de claves foráneas).

Puede mantenerse una indicación acerca de si el elemento dato es esencial u opcional en la aplicación. Puede resultar conveniente la creación de aplicaciones que tengan elementos críticos faltantes.

Una de las causas de la existencia de datos indefinidos es que los valores de datos no estén disponibles cuando se crea el dominio.

En el cambio de enfoque la problemática surge de la aplicación y los dominios son dinámicos, por lo tanto el carácter sensible y extensible es propio, y forma parte del carácter del atributo.

No son situaciones a tratar en forma especial.

REGLA 4 Regla de acceso.

La Descripción de la Base de Datos está representada en el nivel lógico de la misma manera que los datos, por lo que los usuarios pueden el modelo de los datos y el catálogo.

En el modelo dinámico simplificamos las visiones y hasta la operación de los datos puede ser individual por dominio y esquema.

REGLA 5 Regla del lenguaje.

Debe existir un lenguaje que soporte los siguientes items:

- **Definición de datos**
- **Definición de Vistas**
- **Manipulación Interactiva y por Programa de los datos**
- **Restricciones de Integridad**
- **Autorización (seguridad)**
- **Transacciones límites (begin, commit y rollback)**

REGLA 6 Regla de modificaciones teóricas.

Se denominan modificaciones teóricas a las que se realizan sobre las vistas mediante un algoritmo independiente del tiempo y determina una serie de cambios en la base, también son modificables por el sistema (Insertar, Borrar o Modificar).

En el nuevo enfoque por medio de operaciones y formas algebraicas sobre los dominios según requerimientos y modos, tanto en las aplicaciones como en el sistema, se tratan las vistas.

REGLA 7 Regla de capacidad de acceso.

La capacidad de manejo de una relación base o una derivada, o un simple operador, se aplica para recuperar, insertar, modificar o borrar datos. Esto permite al sistema determinar que camino de acceso tomar para elegir el código más eficiente.

Dinámicamente la capacidad de manejo de una aplicación es más simple. Hay un único camino de acceso que está dado por el análisis de la expresión de conjuntos que nos permite armar la estructura de los atributos.

REGLA 8 Independencia física de los datos.

Ante cualquier cambio en el almacenamiento o en los métodos de acceso la lógica de los programas de aplicación no cambia.

La base de datos tiene un límite claro entre los aspectos lógicos y semánticos y los aspectos físicos y de performance.
Estos se estudian e implementan separadamente sin que afecte la lógica de los programas de aplicación.

REGLA 9 Independencia lógica de los datos.
Los programas de aplicación y las actividades terminales permanecen intactas lógicamente cuando se realizan cambios de información de cualquier tipo en la base de datos.
Esta regla permite que el diseño lógico cambie dinámicamente.

Matriz de afinidad y pertenencia.

En nuestro enfoque pensamos al modelo de abstracción de datos, como una matriz integral dinámica donde cada columna es un dominio, cada elemento es un puntero de indirección para eliminar redundancia, cada atributo está tipificado y cada fila puede tener elementos nulos.
Como esta matriz es transparente se pueden agregar o eliminar columnas a esa matriz virtual, dinámicamente los valores de los dominios se pueden modificar o agregar si es que ya no existen, pues es el conjunto de punteros de una columna lo que define su dominio.De esta manera todo objeto o entidad puede mejorar su calificación e identificación.Se alcanza así una gran independencia tanto física como lógica de datos y una dinámica en el crecimiento o expansión.

REGLA 10 Independencia de la Integridad.
Las restricciones de Integridad son almacenadas en el catálogo.
Dinámicamente la Integridad se puede manejar a partir de la aplicación (regla de entidad) y por medio de ella crear los dominios, o bien a partir de los dominios crear las relaciones (regla de relación).

REGLA 11 Distribución Independiente.
La base de datos maneja datos distribuidos al introducirlos por primera vez o cuando los datos están distribuidos (v.g.: SQL).

REGLA 12 Regla de no subversión.

Si se tiene un lenguaje de bajo nivel de un registro por vez, no se deben subvertir las reglas de integridad y de restricciones expresadas en el lenguaje de alto nivel.

Tanto en la regla 11 como en la regla 12, el modelo dinámico tiene un tratamiento independiente de los datos de dominios y por lo tanto no importa si se usa o no un sublenguaje de datos.

2. Reglas estructurales

E.1. Grados de Clasificación de las relaciones.

E.2. Tablas bases representando almacenamiento de datos.

E.3. Tablas interrogantes o tabla resultado y que debe ser salvada para una operación posterior.

E.4. Vistas de tablas o tablas virtuales representadas enteramente por comandos relacionales y no por datos almacenados.

E.5. Tablas instantáneas, que son evaluadas y almacenadas en la base de datos y en el catálogo con su especificación.

E.6. Atributos representados por columnas de las tablas.

E.7. Dominio o conjunto de valores de atributos representados en las columnas de las tablas.

E.8. Regla de parte o clave primaria, formado por atributo(s) candidato(s), de una tabla base, cuyo(s) valor(es) de columna(s) identifica(n) unívocamente una fila.

E.9. Clave foránea es una Regla de Parte o candidato que está sobre el mismo dominio de la Clave Primaria de otra Relación. Sirve para hacer un soporte de integridad sin introducir vínculos entre usuarios.

En el modelo dinámico se representan las tablas bases físicamente por medio de una estructura de datos dinámica bipartita.

Las tablas interrogantes existen lógicamente y se forman por medio de operaciones algebraicas sobre los dominios formando tablas virtuales.

Las tablas instantáneas están representadas por el catálogo de atributos u operaciones relacionales virtualizadas.

La Regla de Parte se define de acuerdo a la aplicación.

Un atributo candidato para una aplicación puede no serlo para otra.

3. Reglas de manipulación.

Operaciones: M1 a M18.

En el modelo dinámico muchas reglas de manipulación de Codd pierden sentido, pues las reglas suficientes son Unión, Intersección y División. Ellas sirven a los propósitos de mantenimiento, actualización y Recuperación de la información sobre los dominios, a través del manejo de tuplas. Al operar sobre los dominios, se tiene la formalización concreta de la teoría de conjuntos y los axiomas de Armstrong [Wiederhold,G., 1983].

4. Reglas de integridad

I.1. Integridad de la Entidad.
Ningún atributo que forma parte de la clave primaria de una base de datos puede aceptar valores nulos.

I.2. Integridad Referencial.
Dada la relación R1 que tiene una clave primaria multiatributo.Si el atributo A de la clave primaria multiatributo está definido sobre el dominio primario D, entonces, para cada valor k de A en R1 debe existir una relación R2 en la base de datos con clave primaria k definida sobre D, tal que k ocurra como un valor en R2.

I.3. Integridad definida por el usuario.

4.1.7. NOTAS DE IMPLEMENTACION

4.1.7.1. CONCEPTOS

La propuesta consiste en el desarrollo teórico-conceptual y de implementación de un sistema de base de datos relacional estructurado sobre dominios dinámicos de atributos.

Se trata de una inversión conceptual respecto de la concepción de las bases de datos existentes, que se basan en las nociones de tablas, tuplas y claves con representación física total o parcial en un almacenamiento real.
Las tuplas (virtuales) se forman mediante las relaciones establecidas a partir de dominios [ver Date en la bibliografía].

Las Bases de Datos existentes se basan en archivos cuyos registros se identifican mediante claves, interrelacionándolos y obteniendo visiones de los datos contenidos con la formación de tuplas pertenecientes a relaciones.

El desarrollo técnico y conceptual de las bases de datos demuestra una fuerte dependencia del concepto de serialización de registros formando archivos. Se observa, en el modelo relacional, que al estructurar se invierte la ley natural de operar sobre los datos. Lo natural es tener conjuntos de datos, información, valores de atributos que deben ser procesados siguiendo procedimientos que los hacen variables en el tiempo. Estos procedimientos lógica y/o matemáticamente representan la ley que hace que los valores de los atributos se relacionen formando tuplas, observables, como entrada o salida de las aplicaciones. Por lo tanto, las tuplas surgen de las aplicaciones a partir de los dominios. Por ello surge la propuesta de operar dinámicamente sobre los dominios.

El modelo es estructuralmente libre: no existen claves o reglas de parte sino calificación de los atributos, y se opera sobre ellos con relaciones a través de pares atributo-valor.

Mediante la calificación de los atributos se fija la forma o modo de relacionarlos entre sí, las conexiones débiles o fuertes entre unos y otros según la aplicación, estableciendo el nivel conceptual del modelo.

Contrariamente a lo expuesto en los modelos relacionales, las tablas se forman a partir de los dominios pero "no" conforman un tipo de registro conceptual, pues no hay algo "real" físicamente almacenado: el modelo es dinámico.

En forma similar que en el SQL [ver párrafos **4.1.** de este capítulo], las tablas son, en sentido estricto y dinámico, "visiones" con leyes de formación que surgen de la lógica de la aplicación.

Podemos asemejar al modelo, estructuralmente, a una gran matriz dinámica donde cada columna es un dominio, y cada fila puede o no contener valores de

todos los dominios. Por ello se habla de visiones y se dice "dinámicas", pues se pueden agregar o eliminar columnas-dominio en la “matriz” y por supuesto se pueden modificar y eliminar valores de atributos, tanto como aumentar el cardinal de un dominio. De esta manera todo objeto o entidad de una aplicación puede mejorar dinámicamente su calificación e identificación. Se alcanza así una gran independencia tanto física como lógica de los datos, y una dinámica en el crecimiento o expansión.

Verdaderamente se obtienen grandes ventajas con este modelo, ya que lo expuesto implica una reestructuración que no depende de los datos sino de las aplicaciones. Las diferentes aplicaciones pueden "ver" a los datos de acuerdo a sus requerimientos y modos.

Se simplifica la visión de los usuarios. Los dominios de los atributos son la base del modelo y de la agregación lógica de ellos mediante operaciones y formas algebraicas relacionales surge la estructura.

Las expresiones del álgebra relacional sirven a los propósitos de mantenimiento, actualización y recuperación de la información de los dominios, a través del manejo de tuplas y preservando su homogeneidad e integridad. Al operar sobre dominios, conjuntos de valores de atributos, se tiene la formalización concreta de la teoría de conjuntos.

Los valores de los dominios son atómicos y los atributos se califican por aplicación; esto permite operar en formas normales.

En las bases de datos existentes se trata de poner restricciones sobre las relaciones para lograr el mínimo de cardinalidad, mediante normalizaciones, sin perder composición en las tuplas.

La dependencia funcional (FD Forma Normal de Boyce-Codd), la dependencia multivaluada (MVD) y la dependencia por composición o juntura (JD) generan cinco formas normales [Date, 1986], que al tratar libremente los dominios, no afectan a las visiones en el modelo propuesto por la atomicidad de los valores, pero sí en las bases de datos existentes al ser la clave un identificador de tuplas, objetos de entidades instanciadas.

El riesgo, sobre todo en la implementación, es el de tener redundancia. Asimismo la libertad de dominios y la dinámica del modelo propuesto puede tener un costo

de performance. De todas maneras se pone un gran énfasis en la claridad conceptual que aportamos y en el comienzo de una seria investigación, que permitirá un debate en torno a lo expuesto y a su implementación, sobre todo en lo que hace a la búsqueda de optimización de la performance. De todas maneras, parafraseando a Antonio Machado: "Caminante no hay camino, se hace camino al andar".

Esta inversión conceptual hace que los dominios de los atributos sean dinámicos y conformen realmente la base del tratamiento de la teoría de conjuntos, y que las tuplas se generen dinámicamente a través de relaciones.
Es una "Base de Datos Relacional Estructurada Sobre Dominios Dinámicos de Atributos" [Perichinsky,G. et al, 1989 a, b, 1990, 1991, 1992].
El diseño de esta base se caracteriza fundamentalmente por desechar el concepto de registro como medio de almacenamiento de tablas de relaciones. Su estructura consiste de dominios de atributos únicos que, integrados dinámicamente, conforman las relaciones que indica el usuario.
Esta nueva filosofía de diseño de bases de datos presenta varias ventajas, entre las cuales cuentan el logro de una total independencia entre los datos (tanto entre los datos entre sí como entre ellos y las relaciones a que pertenezcan) y la naturalidad y absoluta libertad en la formulación de las consultas. Sin embargo se debe dejar en claro que en contraposición con las mejoras logradas respecto a otros manejadores, el propuesto adolece de una degradación de la performance proporcional a las mismas, los conceptos anteriormente expuestos y en adelante permiten aportar una posible solución a la aludida degradación de performance del modelo [Perichinsky,G. et al, 1990].

El modelo se conceptualiza mediante dominios independientes dinámicamente integrables por medio de referencias internas; es decir, cada atributo de cada relación descripta por el usuario pertenece a un dominio de valores independiente y se "relacionará" con los restantes mediante un número que indica la pertenencia a la relación. Consecuentemente, la clave del manejo eficiente de los datos está en la operatoria del diccionario de atributos y en las estructuras de soporte, tanto de éste como de los dominios e índices de valor.
El diccionario de atributos especifica el tipo de cada atributo, cómo acceder a sus valores según las referencias internas, cómo acceder a las listas invertidas para

realizar operaciones relacionales y cómo manejar las estructuras de soporte correspondientes.

Cada atributo posee una tabla de listas invertidas para, entrando por un valor, recuperar todas las referencias internas correspondientes y así efectuar las operaciones relacionales que desencadene el criterio de selección especificado por el usuario al invocar al manejador. Una vez evaluado el criterio, con las referencias internas resultantes se entra a las tablas de cada dominio que involucre la relación corriente y se recuperan los valores asociados para cumplir con ellos el objetivo de la invocación.

Nótese que al aludir a las invocaciones del usuario al manejador, se habla de relaciones asociadas en modo liberal, puesto que es en éstos momentos cuando las relaciones quedan definidas, corporizando así el concepto de Dominios Dinámicamente Integrables.

En cuanto a las estructuras de soporte, tanto del diccionario de atributos como de las tablas de valores y de listas invertidas, con el objeto de lograr la máxima performance en el tiempo de respuesta y luego de comparar estadísticas sobre distintas estructuras optativas (árboles binarios, árboles B y organizaciones de acceso directo e indexado) encontramos que las más adecuadas para tal propósito son las de árboles B+. Estas estructuras, al tener mayores factores de ramificación resultan en árboles de menor altura, lo que obviamente reduce la cantidad de accesos para recuperación de información. Con respecto al tamaño de los nodos, en general resulta conveniente establecerlo en concordancia con el tamaño de la unidad de intercambio de los sistemas operativos (páginas, sectores, bloques, etc.). Asimismo cabe señalar que en el caso de que sea posible manejar estructuras íntegras en memoria volátil (por características arquitectónicas del equipo de soporte o simplemente por el volumen de datos involucrado) las estructuras más apropiadas en estos casos son la de árboles binarios, ya que los accesos no juegan un papel importante y en cambio sí lo juega el ahorro de espacio.

4.1.7.2. ESTRUCTURAS DE DATOS DEL MANEJADOR

La descripción formal de cualquier implementación de una estructura de datos es similar a la descripción formal de los lenguajes de programación. En la definición

de cualquier lenguaje, usualmente se comienza por presentar una sintaxis de programas válidos en forma de gramática y luego se establecen restricciones de validez (v.g.: uso de reglas para nombres simbólicos) que no pueden expresar la gramática. Análogamente, una implementación de una estructura de datos será válida si satisface una gramática sintáctica y también obedece a ciertas restricciones. Por ejemplo, para que una estructura de datos sea un árbol binario balanceado en carga, deberá cumplir con las reglas gramaticales para los árboles binarios y al mismo tiempo satisfacer las restricciones de balanceo.

Gramática para objetos de datos

Para definir una secuencia de números reales, se puede usar la producción BNF:

<s> :: = [real, <s>] | **nada**

Así una secuencia de números reales será:

nada, [real, nada] , [real, [real, nada]],...

Análogamente, se pueden definir secuencias de enteros, strings, etc.
Así resulta una voluminosa colección de reglas de producción para todas las secuencias de datos distintas posibles. Se podría tratar de evitar la superabundancia de definiendo una secuencia de datos **"en general"** (abstracción):

<s> :: = [<d> , <s>] | **nada**

<d> :: = real | **entero** | **string** | **...**

Sin embargo este par de reglas de producción genera secuencias indeseables como:

[real, [entero , nada]],...

en lugar de secuencias homogéneas.

Para superar estas anomalías, se puede definir la sintaxis de los objetos de datos del modelo utilizando una gramática W (también llamada gramática de dos niveles o de van Wijngaarden) [van Wijngaarden, 1976].

Sin utilizar efectivamente todas las capacidades de las gramáticas W, presentamos la sintaxis usando las producciones BNF equivalentes junto con reglas de reemplazo uniforme.

Una gramática W genera un lenguaje en dos pasos. En el primero, se usa una colección de reglas generalizadas para crear reglas de producción más específicas. En el segundo, se usan reglas de producción generadas en el primer paso para definir las estructuras de datos efectivas.

Seguidamente especificamos una gramática W para generar tipos de datos convencionales:

4.1.7.2.1. METAPRODUCCIONES

- **M1. D ::** entero; real; carácter; string; booleano;...;
  - **D[N..N];** "arreglo"
  - **Registro;** (Registro); "registros"
  - **[D];** "referencia"
  - **s-D;** "secuencia"
  - **ag-D-Hoja;** "árbol general"
  - **Diccionario;** "estructuras de diccionario"
- **M2. Diccionario ::** Clave[N..N]; s-Clave; "búsqueda secuencial"
  - **ab-Clave-Hoja;** "árbol binario"
  - **am-N-Clave-Hoja;** "árbol multipropósito"
- **M3. Registro ::** D; D, Registro. "definición de registro"
- **M4. Hoja ::** nada; D.
- **M5. N ::** Dígito; Dígito N.
- **M6. Dígito ::** 0; 1; 2; 3; 4; 5; 6; 7; 8; 9.
- **M7. Clave ::** entero; real; carácter; string; (Clave, Registro). "clave de búsqueda → regla de parte"

4.1.7.2.2. HIPERREGLAS

- **HR1.** estructura de datos: D.
- **HR2.** s-D: [D, s-D]; nada.
- **HR3.** ab-D-Hoja: [D, ab-D-Hoja, ab-D-Hoja]; Hoja.
- **HR4.** am-N-D-Hoja: [entero, D[1..N], am-N-D-Hoja[0..N]]; Hoja.
- **HR5.** ag-D-Hoja: [D, s-ag-D-Hoja]; Hoja.

4.1.7.2.3. ESPECIFICACIÓN DE DATOS

- **Atributo** → (Clave, Registro)
- **Clave** → string (1)
- **Registro** → carácter(2), entero (3), entero (4), (Dominio, Dominio-Inverso) (5)

Restricciones:

- (1) Nombre del Atributo.
- (2) Tipo: E, entero; R, real; C, carácter; S, string; B, booleano; etc.
- (3) Longitud: si tipo es E, cantidad de dígitos; si es R, cantidad de dígitos de la parte entera; si es S, cantidad de caracteres; si es C o B, nada; etc.
- (4) Decimales: si tipo es R, cantidad de dígitos de la mantisa, sino nada.
- (5) Referencias a estructuras de soporte definidas en Organización de Datos.
  - **Elemento-Dominio** → (Clave, Registro)
  - **Clave** → Referencia-Interna
  - **Registro** → D

Restricciones:

- **D** tipo atómico **(M1)** coincidente con el tipo de Atributo al que corresponde.
- **Elemento-Dominio-Inverso** → (Clave, Registro)
- **Clave** → D (1)
- **Registro** → entero (2), Referencia-Interna[1..N] (3), Lista (4)

Restricciones:

- (1) Tipo atómico coincidente con el del atributo asociado.
- (2) Cantidad de referencias (longitud de la lista de inversión)
- (3) Arreglo con las primeras referencias.
- (4) Referencia al resto de las referencias, organizado en una estructura descripta en Organización de Datos.

Las referencias se desdoblan en (3) y (4) por razones de performance, puesto que aquellos atributos que sean naturalmente clave de alguna relación tendrán pocas instanciaciones dentro del dominio y no necesitarán la implementación de (4).

- **Referencia-Interna→** entero

Restricciones:

Las referencias internas han de manejarse naturalmente en una variable global del sistema con la previsión de umbrales que, para evitar el crecimiento desmedido de valores, provoquen al ser sobrepasados solicitudes de reorganización.

4.1.7.2.4. Organización de datos

- **Diccionario-de-atributos→** am-2n-D-nada: [entero, Atributo[1..2n], am-2n-D-nada[0..2n]]; nada
- **Dominio** → am-2n-D-nada: [entero, Elemento-Dominio[1..2n], am-2n-D-nada[0..2n]]; nada
- **Dominio-Inverso** → am-2n-D-nada: [entero, Elemento-Dominio-Inverso[1..2n], am-2n-D-nada[0..2n]]; nada
- **Lista→** am-2n-D-nada: [entero, Referencia-Interna[1..2n], am-2n-D-nada[0..2n] [...]; nada

Restricciones:

El entero inicial en cada bloque de cada árbol indica la cantidad de claves efectivas en el nodo, manteniéndose siempre entre n y 2n. Las claves efectivas estarán almacenadas contiguamente en el arreglo de claves comenzando en la primera posición. El número ***n*** tiene un valor propio en cada estructura.

Como ejemplificación de lo expuesto hasta aquí, consideremos un atributo "Persona", cuyos valores sean el apellido y nombre/s de una persona física.

Así el objeto correspondiente de Atributo sería

Persona, S, 30, nada, r1, r2

donde S indica que los valores del atributo son de tipo string; 30, que la longitud máxima de los strings será de treinta caracteres; el valor nulo, que ese ítem no aporta información relevante; y las dos referencias siguientes, son las direcciones de los nodos raíz del dominio de valores del atributo y del dominio inverso o conjunto de listas invertidas correspondientes a cada valor de un atributo.

Continuando con el ejemplo, algunos objetos de Elemento-Dominio podrían ser

- 1, Errecaborde Ignacio
- 2, Maturano María Elena
- 3, Fernández Fernando
- 4, Atanasov Anastasia
- 5, Fernández Fernando

y sus correspondientes objetos de Elemento-Dominio-Inverso,

- Errecaborde Ignacio, 1, (1, nada, ..., nada), nada
- Maturano María Elena, 1, (2, nada, ..., nada), nada
- Fernández Fernando, 2, (3, 5, ..., nada), nada
- Atanasov Anastasia, 1, (4, nada, ..., nada), nada

Las referencias internas se disponen ordenadas de menor a mayor en el arreglo y si lo desbordaran, seguirían ordenadas en un árbol B+ cuya raíz indicaría la última referencia de los objetos.
Tanto los objetos de Elemento-Dominio como los de Elemento-Dominio-Inverso se organizan en sendas estructuras de árbol B+. Para simplificar, consideremos los objetos de Elemento-Dominio; así, si el parámetro n del árbol es 1 (elegido aquí convenientemente pequeño por propósitos obvios) y los objetos fueron creados en el orden de sus referencias internas, el árbol sería

- raíz: 2, ((2, Maturano María Elena), (4, Atanasov Anastasia)),(ra, rb, rc)
- ra: 1, ((1, Errecaborde Ignacio), nada),(nada, nada, nada)

- rb: 1, ((3, Fernández Fernando), nada),(nada, nada, nada)
- rc: 1, ((5, Fernández Fernando), nada),(nada, nada, nada)

4.1.7.2.5. IMPLEMENTACIÓN

En el estado de las investigaciones, con este trabajo se consiguió fijar conceptualizaciones sobre el tratamiento de la información en una base de datos, con gran parte de la rigurosidad del formalismo en materia de implementación y performance. De todos modos las ideas asociadas corresponderán a avances más refinados.

4.1.7.2.5.1. ALMACENAMIENTO DE LA INFORMACIÓN

El agrupamiento de la información en el modelo se fundamenta en el concepto de "dominios" de atributos. Así es que la estructura de la base se conforma de:

⇒ Un **Diccionario de Atributos**, en el cual se encuentra toda la información necesaria para la identificación y manejo de los atributos. Este diccionario no es más que un archivo cada uno de cuyos registros contiene las características (longitud, tipo, etc.) de un atributo y las referencias a dos tablas (lógicas) organizadas **(1),** una por Referencias Internas para la identificación de las tuplas y **(2),** la otra por Valores de Atributo (lista invertida).

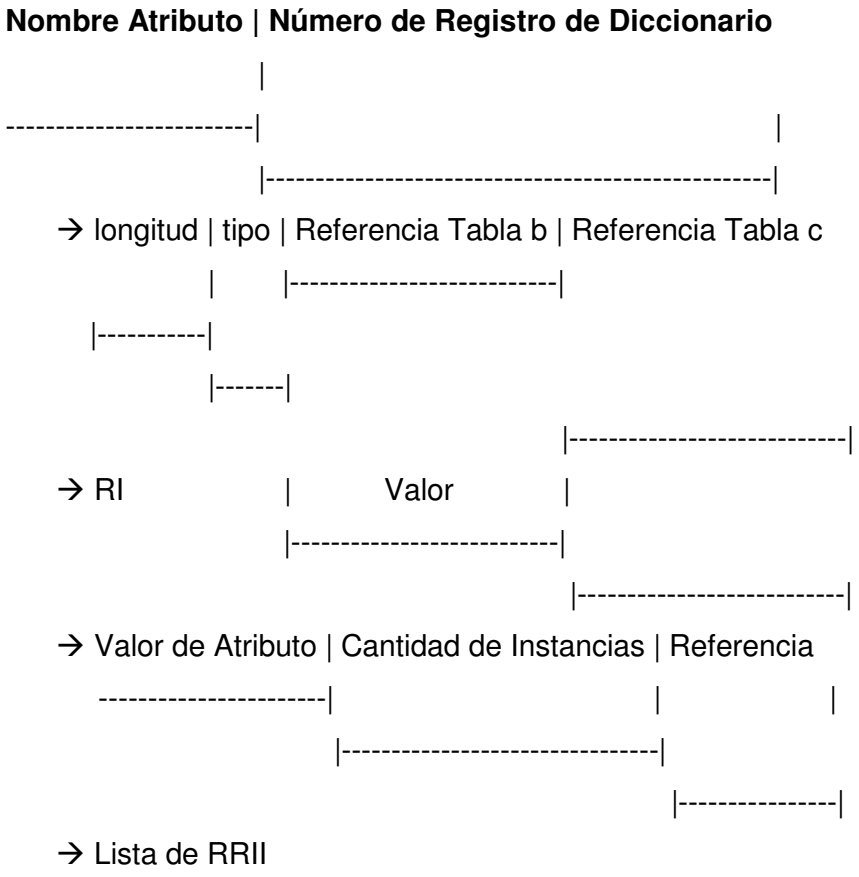
⇒ El diccionario se puede acceder por una tabla Nombre de Atributo|Número de Registro del Diccionario, implementada como árbol tipo B+ o con acceso multinivel, a efectos de optimizar el tiempo de acceso.

⇒ Una tabla **Referencia Interna | Valor** para cada uno de los atributos, referenciada por el diccionario y cuyo objeto es permitir la identificación de las aplicaciones, ya que los valores de atributos correspondientes a una instancia particular de una aplicación (tupla) estarán asociados con la misma Referencia Interna (en adelante RI).

⇒ Tablas **Valor de Atributo | Cantidad de Instancias | Lista de RRII**, correspondientes a cada atributo, que clasifican la información de la base

según los distintos valores de cada dominio, facilitando considerablemente las consultas complejas.

Respecto de las dos tablas asociadas a cada atributo, su implementación se puede efectuar utilizando archivos de bloques de acceso directo, estructurados como árboles tipo B+ o en niveles múltiples y dinámicos de indirección.



→ apuntan a una instancia

→ apuntan a una estructura

4.1.7.2.6. MANIPULACIÓN DE LA INFORMACIÓN

El manejador de la base soporta seis comandos básicos: dos de ellos orientados al manejo de atributos y los cuatro restantes orientados al manejo de aplicaciones sobre conjuntos de atributos.

Los comandos activan rutinas del Sistema Operativo que trabajan en áreas de comunicación específicamente reservadas y en las que se almacenan las estructuras de control. No preocupa la sintaxis de los comandos, sino que se trabaja partiendo de la idea de confeccionar un program embbeding de un lenguaje de datos o programación (SQL estándar) y la estructura propuesta.

4.1.7.2.6.1. COMANDOS ORIENTADOS AL MANEJO DE ATRIBUTOS

Permiten la incorporación y eliminación de un atributo a la base.

El comando de incorporación o alta de atributo consiste en, dada la información nombre de atributo, tipo y longitud, y luego de corroborar la no existencia de otro con igual nombre, la generación del correspondiente registro del diccionario, junto con las referencias a la tabla de valores por RI y a la lista invertida (obviamente nulas).

Se supone que la tabla de valores por RI está estructurada en forma de árbol tipo B+, por lo que en adelante se denominan "árbol de valores". En cuanto a la otra tabla se sigue llamando "lista invertida".

El segundo comando de baja de atributo dado el nombre del mismo, consiste en el acceso al registro del diccionario que le corresponde y una vez recuperadas las referencias al árbol de valores y a la lista invertida, en la eliminación de dichas estructuras (vía comando al sistema operativo, por ejemplo, en el caso de que dichas estructuras estén implementadas en sendos archivos). Una vez suprimidas ambas tablas sólo resta la supresión del registro del diccionario para cumplir totalmente el objetivo.

4.1.7.2.6.2. COMANDOS ORIENTADOS AL MANEJO DE TUPLAS

Permiten efectuar altas, bajas, modificaciones y consultas de instancias de un conjunto de atributos.

Dada una lista de pares Nombre de Atributo-Valor, en la que el conjunto de valores constituye una tupla de la aplicación deseada, el alta de esta última consiste en la generación de una referencia interna (RI) que vincule a los valores y en la actualización del árbol por un lado, y de la lista invertida por el otro.

A los efectos de mantener la consistencia e integridad, uno de los atributos intervinientes debe estar calificado como único, es decir, se impone que su valor sea irrepetible en la aplicación. Así es que con el valor de este atributo se accede a la lista invertida que le corresponde; si existe una entrada para dicho valor se obtiene la RI con que se lo localiza dentro de su dominio, sino se genera una, sea tomándola de una lista de RRII libres, o bien generando realmente una en caso de que la lista esté vacía. Y así se continúa con los restantes atributos hasta confirmar que la instancia que se desea ingresar no existe ya y se completa el comando.

El comando de altas fuerza a recorrer todas las estructuras asociadas a todos los atributos involucrados y a realizarlas definitivamente cuando se concluyen dichos recorridos. Debido a esto, y a los efectos de optimizar la performance, se mantiene una cadena de referencias para no tener que acceder en forma total o parcial a las estructuras reiteradamente. Otra alternativa probada es intersecar, a medida que se recorren los atributos y siempre comenzando por el calificado "único", los conjuntos de RRII de cada uno y efectuar el alta al comprobar que la intersección final resulta vacía.

En cuanto a los comandos restantes de esta categoría, todos ellos coinciden en la forma de recuperar a las tuplas que habrán de ser eliminadas, modificadas o simplemente consultadas. Es por ello que necesitan partir, en su ejecución, de un criterio de búsqueda común, cuya aplicación puede resultar en un conjunto vacío, unitario o multitudinario de tuplas. El criterio de búsqueda que se define como sigue:

criterio := condición | **criterio-operador lógico-condición**

condición := nombre de atributo-operador-constante |

nombre de atributo-operador-nombre de atributo

operador lógico := o | y

operador := = | /= | > | >= | < | <=

donde | implica disyunción y **-** concatenación.

Para obtener el resultado del criterio se procede sencillamente:

I. se agrupan las condiciones que involucren a los mismos atributos;

II. se accede a la lista invertida correspondiente a cada atributo interviniente y se recuperan todas sus RRII (se pueden almacenar en memoria real o, en caso de ser necesario en una zona de resultados intermedios en memoria auxiliar);

III. se realizan las operaciones lógicas que correspondan con las listas de RRII obtenidas.

De la cumplimentación de los tres pasos precedentes se obtiene el conjunto de RRII a valores de atributos que cumplen con el criterio general. A partir de estas RRII y con la restante información de los comandos se conforman las tuplas sobre las que el comando se completa.

Para el caso de las bajas, acompaña al criterio de búsqueda una lista de nombres de atributos y el comando se completa accediendo a los árboles correspondientes a cada uno de ellos, eliminando las hojas correspondientes a las RRII y suprimiendo de las listas invertidas las mismas RRII.

En cuanto a las modificaciones, obviamente sólo se podrán efectuar luego de conformar las tuplas que desean modificarse, por lo que el comando aludido se constituye de dos fases. Estas fases, que dependen en su implementación de las características u orientación del administrador en general, corresponden a la baja de los valores iniciales de los atributos en primer lugar y al alta de los valores que sustituyen a aquellos en segundo término. Desde luego todo esto se realiza sobre RRII ya existentes. En conclusión, los datos que debe aportar el usuario para desarrollar este comando podrían ser el criterio de búsqueda para individualizar una tupla, que incluya a todos los nombres de atributos cuyos valores se modificarán; junto con la lista de valores que se harán corresponder a los aludidos atributos. Para el caso de que el comando actúe sobre un conjunto de varias tuplas, las modificaciones no pueden ser individuales sino responder a un criterio general, que de alguna manera uniforme el procedimiento.

Por último, sólo resta agregar que en el comando de consultas, el más simple de todos, acompaña a la especificación del criterio de búsqueda únicamente la lista de nombres de atributos que se desean consultar y que concuerden con el criterio especificado.

4.1.7.2.6.3. COMANDOS

Cabe consignar que en la descripción general de la acción de los comandos se alude a una lista de RRII cuya instrumentación no se aclara en el desarrollo de "Almacenamiento de la Información". Respecto a ello se sugiere que se implemente en un archivo en bloques con estructura, por ejemplo de pila y que entre los procedimientos o rutinas utilitarias del administrador figure una de mantenimiento de dicha lista, cuyo objetivo sea, una vez alcanzado cierto límite en el valor de las RRII generadas, recolectar todas aquellas que no tengan asociado valor alguno de atributos y conformar así la pila en el archivo. Así toda vez que se requiera una referencia se tomará una del último registro del archivo [ver **"contrastación"** párrafo **2.2.10.** del estado del arte de las bases de datos].

Actualización

update { secuencia-atributo-valor, criterio }

Borrado

delete { secuencia, criterio }

Inserciones

insert { secuencia-atributo-valor }

Consultas

select { secuencia, criterio, { ,orden } }

Donde:

- **Secuencia-atributo-valor** es una lista de pares de la forma **<atributo,valor>**.

- **Secuencia** es una lista de atributos.
- **Criterio** es una selección de tuplas.
- **Orden** es una lista de atributos.

4.1.7.2.7. FUNCIONES

En el modelo se distinguen dos tipos de funciones u operaciones distintas: las de manejo del diccionario y las de manejo de datos.
Las operaciones de manejo del diccionario se pueden asociar a las operaciones básicas de "construcción" en las bases de datos tradicionales [Codd, Date, Wiederhold, de Miguel, etc.], cuales son las de manipulación de esquemas. En este caso consisten en la construcción, actualización y recuperación de objetos Atributo en una organización en árbol B+; o sea que se traducen en altas, bajas, modificaciones y consultas en una estructura tal. Son operaciones simples que constituyen la base o corazón de las restantes, más complejas.
Las operaciones de manejo de datos son más complicadas y naturalmente constituyen el espíritu paradigmático del modelo. Entre éstas distinguimos a las altas de elementos de dominios, que pueden ser ligadas a una relación existente o no; a las bajas de elementos de dominios, que pueden combinarse con las anteriores para constituir modificaciones a instancias de relaciones preexistentes y a la recuperación de elementos de dominios.
Seguidamente especificamos estas últimas funciones, considerando a las altas, bajas, modificaciones y consultas sobre árboles B+ como primitivas.

4.1.7.2.7.1. ALTAS INDEPENDIENTES

Entradas: s-(string, D) donde el string es un nombre de atributo y D un tipo atómico concordante con el del atributo nominado. Por ejemplo, si consideramos los atributos "Persona", ya definido y otro "DNI", cuyos valores son números de documentos, una secuencia de entrada podría ser
(Persona, "Paredes Joaquín"), (DNI, 17.520.232)

Proceso: consiste en la generación de una nueva referencia interna para vincular los objetos que se incorporan a sendos dominios; efectivizar dicha incorporación

con la mencionada referencia y finalmente buscar en los dominios inversos los valores ingresados y si se encontraren, agregar a sus listas de referencias la recién generada y sino incorporar el nuevo valor junto con su primer referencia a la estructura.

Salidas: código de resolución, para el caso de que hubiese ingresado algún nombre de atributo incorrecto o para asegurar el éxito de la operación.
Observaciones: como sería muy costoso detectar la duplicación de instancias de una relación, queda como responsabilidad del usuario evitar dicha contingencia, obligándose a un manejo muy cuidadoso de los datos a ingresar o a una consulta previa.

4.1.7.2.7.2. ALTAS RELACIONADAS

Entradas: s-(string, D), s-(string, D) donde la primer secuencia es análoga a las de las altas independientes y la segunda constituye un criterio de selección de las instancias a ampliar. Por ejemplo, si consideramos existente la relación "Persona" con "DNI" especificada anteriormente y otro atributo "Dirección" cuyos valores representan un domicilio, una posible entrada sería;
((Dirección, "Av. 32 nro. 1356")), ((Persona, "Paredes Joaquín"), (DNI, 17.520.232))
Desde luego sería posible que la segunda secuencia tuviera uno solo de los pares especificados, pues en la práctica, cuanto más simple el criterio tanto mejor.

Proceso: esta operación puede identificarse con la de modificación de esquemas en una base tradicional, lo que en tal marco, obliga a la reorganización total de las estructuras de soporte de la relación involucrada. Aquí ello no es necesario, pues recuperando de los dominios inversos las referencias correspondientes a sendos valores de los atributos especificados en la segunda secuencia y efectuando luego sus intersecciones, se obtiene la o las instancias a las cuales añadir los valores de la primer secuencia, es decir, una vez obtenidas

las referencias internas asociadas a la secuencia ***2*** se procede para cada una de ellas análogamente a la operación de altas independientes.
Observaciones: de igual modo que en la operación anterior y como en todas las de este modelo en general, lo que se gana en libertad posibilitando el manejo discrecional de las relaciones, debe equilibrarse con la consecuente responsabilidad de los autorizados a dicho manejo. En este caso, tan enorme como el potencial de la función puede ser el perjuicio que ocasione su empleo inadecuado.
Las operaciones de modificación de instancias de relaciones implícitas existentes, tal como anunciamos en la introducción de las funciones, pueden implementarse naturalmente amalgamando esta operación con las bajas de las instancias preexistentes de los atributos especificados en la primer secuencia, determinadas por la primera.

4.1.7.2.7.3. BAJAS

Entradas: s-string, s-(string, D) donde la primer secuencia corresponde a los atributos cuyos valores habrán de ser eliminados; y la segunda constituye, igual que en las altas relacionadas, un criterio de selección para determinar las instancias a suprimir. Por ejemplo, siguiendo con lo utilizado en zaga podemos considerar;
(DNI, Dirección), (Persona, "Paredes Joaquín")
que constituye una baja parcial, o también
(DNI, Dirección, Persona), (Persona, "Paredes Joaquín")
que sería una baja total y completa las posibilidades de complementación en cuanto a las altas.

Proceso: consiste en hallar las referencias internas que verifiquen el criterio determinado por la segunda secuencia, del mismo modo que en las altas relacionadas; eliminar las instancias (elementos del dominio) de los atributos enumerados en la primer secuencia que ellas indiquen y finalmente eliminarlas de las listas pendientes de los elementos de dominios inversos que correspondan.

Salidas: ídem operaciones anteriores.

Observaciones: cabe señalar que la segunda secuencia de entrada podría ser vacía, lo que implicaría la eliminación total de los atributos listados en la secuencia inicial.

4.1.7.2.7.4. RECUPERACIONES

Estas operaciones, que constituyen el quid de toda base de datos, se pueden implementar con una amplísima gama de criterios, a partir de una primitiva análoga a las bajas, con prescindencia del borrado. Así, mediante iteración de la primitiva y unión de las referencias resultantes se satisfacen criterios de rangos de valores y mediante amalgamación con conjunciones o disyunciones se satisfacen criterios lógicos generales.

4.2. ESPECTROS DE EVIDENCIA TAXONOMICA

La **clasificación** es una técnica de abstracción usada para agrupar objetos con propiedades comunes. Esta permite delimitar el dominio de objetos, bajo la hipótesis de que cada objeto pertenece a una clase (y solo una) y de que, para cada clase, hay al menos un objeto que pertenece a ella.

La búsqueda de conceptos clasificatorios que permitan una estructura de clasificación que no se modifique por el agregado de nueva información (estabilidad de la clasificación), ni se altere por la incorporación de nuevas entidades, nos motiva a buscar nuevas herramientas analíticas [Perichinsky,G. et al, Innsbruck. 1996] [Perichinsky,G. et al, Innsbruck.1997].

Este trabajo es un nuevo avance en la problemática de la taxonomía numérica utilizando las que hemos denominado Bases de Datos Relacionales Dinámicas, que constituyen un modelo que cambia la estructuración tradicional de agrupamiento estático de valores de atributos en registros por la creación de dominios independientes dinámicamente integrables por medio de estructuras dinámicas (v.g.: arborescente). El nuevo enfoque consiste en un mayor nivel de abstracción para conseguir la máxima independencia posible, ya que las referencias a los datos almacenados están aislados de los cambios y de los

diferentes usos en el entorno de los datos [Perichinsky,G. et al. Montevideo.1996] [Perichinsky,G. et al., Buenos Aires. 1997].

La aplicación de estos conceptos en los expresados en la semántica de las bases de datos relacionales dinámicas es directa.

La asociación de conceptos en forma sistemática, para clasificar, utilizando variables numéricas, son técnicas matemáticas que conforman la **Taxonomía Numérica**, disciplina definida como la evaluación numérica de la afinidad y similitud entre unidades taxonómicas y el agrupamiento de esas unidades en **taxones** (taxa como plural de taxon, cluster, familia), basada en el estado de sus caracteres.

Considerando la clasificación como el proceso de agrupamiento de objetos en clases teniendo en cuenta sus atributos comunes y relaciones, la taxonomía es el estudio teórico de la clasificación incluyendo sus principios, procedimientos y reglas.

Los objetos, a los que se denominan unidades taxonómicas operacionales (OTU), se clasifican según un proceso basado en las propiedades de los mismos. La diferencia entre ellas es la fuente de la evidencia taxonómica.

Un carácter podría ser definido como cualquier propiedad que caracteriza la OTU en estudio. Los ***estados*** son el conjunto de valores posibles de sus caracteres.

Para estimar la similitud taxonómica usada para agrupar las OTU´s tratamos de expresar esta semejanza en una manera cuantitativa. Usamos el coeficiente de similitud para cuantificar esta semejanza, es decir para obtener la semejanza para cada par de OTU´s de la matriz básica de datos.

A fin de trabajar de un modo novedoso con esta matriz introduciremos aquí Técnicas de Teoría de la Información.

La Taxonomía Numérica permite agrupar, a través de métodos numéricos llamados análisis de clusters, unidades taxonómicas operacionales (OTU´s) en taxa o grupos de OTU´s en función de sus estados característicos: valor, dominio, atributo. Los clusters constituyen familias cuyo primer análisis estructural es en base a su característica fenotípica, estructuras taxonométricas que muestran la relación en grado de similitud entre dos OTU´s o grupos de OTU´s.

Las OTU´s toman valores de los dominios dinámicos de atributos que forman entidades que van cambiando de acuerdo a las necesidades taxonómicas: clasificar para formar familias o clusters [Perichinsky et al., Innsbruck. 1997].

La conformación de grupos (clusters*) e*n base a la matriz de similitud, distancia Euclídea entre OTU´s, métrica de Minkowski y Manhattan y la aplicación de técnicas de *"vecinos próximos"* se realiza mediante OTU´s que se asocian por su grado de similitud. El objetivo no es mostrar sólo relaciones de similitud entre pares (*pair group method*) de OTU´s, sino entre todas las OTU´s (construcción de la matriz de similitud).

Se logra la fuente de la **evidencia taxonómica** al cuantificar a partir del coeficiente de similitud, la semejanza para cada par de OTU´s de la matriz básica de datos.

Surge el concepto de espectro de los estados de los caracteres de los pares de OTU´s respecto al total y la formación del espectro de familias por el principio de superposición al procesar los espectros a derecha e izquierda de los pares de OTU´s y la obtención de ***Invariantes*** (centroide, varianza y radio).

La secuencia algorítmica está dada por la conformación de la matriz de datos, su normalización, la construcción de la matriz de similitud, los espectros de totales y de familias formadas por clustering y el análisis de invariantes.

Por todo lo expuesto se desarrollaron técnicas para atacar la problemática desde la **Teoría de la Información** y se operó sobre técnicas de clasificación cuyo fundamento está en la **Taxonomía numérica** según sus métodos y principios.

4.2.1. MATRIZ DE DATOS

Para estimar la semejanza entre pares de OTU´s adoptamos la convención del arreglo de datos para la taxonomía numérica en la forma de una matriz de n x t , cuyas t columnas representan t OTU´s a ser agrupados sobre la base de semejanzas y cuyas n filas son unidades de caracteres. Cada entrada Xij en tal matriz es el valor del OTU j para el carácter i.

Las operaciones matemáticas permiten calcular la afinidad entre unidades taxonómicas en base al estado de sus caracteres.

Los pasos comunes a casi todos los métodos numéricos son los siguientes:

1. Elección de OTU´s.
   - Son los objetos a estudiar.

2. Elección de caracteres.
   - Describen las propiedades de cada OTU. Al comenzar la clasificación tienen todos el mismo peso o importancia. Se registra el estado (valor que puede tomar cada carácter) de los caracteres de cada OTU. La taxonomía numérica exige que todos los estados sean expresados en forma cuantitativa de manera de poder ser computables.

4.2.2. CONSTRUCCIÓN DE LA MATRIZ DE DATOS

Si tengo "t" OTU´s y "n" caracteres la matriz de datos visualiza los estados de los caracteres para cada OTU.

	OTU_1	OTU_2	OTU_3	...	OTU_t
$carácter_1$	x_{11}	x_{12}	x_{13}	...	x_{1t}
$carácter_2$	x_{21}	x_{22}	x_{23}	...	x_{2t}
...	...	...	...	...	...
$carácter_n$	x_{n1}	x_{n2}	x_{n3}	...	x_{nt}

(n x t)

Siendo x_{nt} el valor del carácter "n" para la OTU "t".

1. Columnas:
   - Representan "t" OTU´s a agrupar en base a su similitud.
2. Filas:
   - Representan "n" caracteres de OTU´s.
3. Espacio A:
   - Espacio de caracteres que tiene "n" dimensiones.
4. Espacio I:
   - Espacio individual o de objeto que tiene "t" dimensiones.

Esta matriz puede ser estudiada desde dos puntos de vista:

a) el de la asociación de caracteres, llamada *Técnica R.*

b) el de la asociación de las OTU´s, llamada *Técnica Q.*

4.2.3. NORMALIZACIÓN

Es útil considerar los efectos producidos por las operaciones de cambios de escala para operadores de distancias y ángulos respecto a coordenadas de referencia y su correlación respecto al cambio del origen. En taxonomía numérica los cambios de escala interfieren con el estudio de la evidencia taxonómica en cuestión [Crisci y López Armengol, .1983].

Otra clase de método de operación sobre la escala apunta a hacer la variación de los caracteres transformados tan igual como sea posible. El intento está en permitir que cada carácter contribuya a la semejanza total en proporción inversa a su variabilidad en el conjunto entero de OTU´s. Así un carácter con un pequeño rango de variación contribuye tanto como otro carácter con un gran rango de variación. Aunque se ha investigado mucho sobre los efectos de tales procedimientos de escala, la lógica detrás de ellos no está completamente explorada. Ambos procedimientos eliminan los efectos de tamaño y la variabilidad de un carácter, pero uno podría producir la eliminación de uno solo de estos. Nosotros podemos considerar por lo tanto estas transformaciones bajo tres casos, (1) esas que igualan el tamaño bruto de cada carácter, (2) esas que igualan la variabilidad de cada carácter, y (3) esas que hacen ambas transformaciones.

El tamaño bruto de cada carácter puede igualarse de varias maneras. El más lógico que es restar el promedio, no aparece como propuesto. Sin embargo, tal procedimiento no resuelve el efecto de una varianza grande (rango amplio).

El método es dividir cada valor por el estado máximo de ese carácter, cociente mediante $X' = X / Xmax$, se obtiene una escala de los estados de los caracteres en el intervalo 0 a 1.

Nótese, sin embargo, que aunque el máximo de estado se marca como 1, el más pequeño no se puede marcar como cero. El método es dejar sin cambiar la variabilidad relativa (el rango dividido por el máximo) y los tamaños de carácter se igualan a la norma del estado más grande de cada carácter. Así no aparece ningún empleo en taxonomía numérica de las transformaciones que igualan la variabilidad mientras dejan el tamaño bruto sin cambiar.

La forma simple de igualar ambos, el tamaño y la variabilidad, es operar sobre el rango, como se usa en el coeficiente de Gower Sg [Sokal y Sneath, 1973], donde

el valor más pequeño para el carácter se resta de cada valor y el resultado es dividido por el rango:

X´= (X-Xmin) / (Xmax-Xmin).

De esta manera el estado más pequeño de los OTU´s tiene el valor cero y el estado más grande tiene el valor 1. Los caracteres normalmente distribuidos, X´ se aproximan a una función monotónica de la desviación estándar (la relación del rango a la desviación estándar esperada, para tamaños diferentes de muestreo, se puede encontrar en textos de estadística).

En la normalización de caracteres se computan la media y la desviación estándar de cada fila (los estados de cada carácter) y expresamos cada estado como una desviación de la media en unidades de desviación estándar. La normalización de los estados del carácter hace que la media de todo carácter sea de valor cero y varianza de valor unitario.

$$\overline{Xj} = \left(\sum_{i}^{n} Xij \right) / n$$

$$\sigma j = \left(\left(\sum_{i}^{n} (Xij - \overline{Xj})^2 \right) / (n - 1) \right)^{1/2}$$

$$\overline{X'ij} = (Xij - \overline{Xj}) / \sigma j$$

Si nosotros deseamos agregar un OTU nuevo podemos calcular los valores estandarizados desde las medias previas y las respectivas desviaciones estándar, aunque el valor resultante no se corregirá realmente, sin embargo, porque ambos la media y la desviación estándar de los estados del carácter no han cambiado con la adición del OTU nuevo. Cuando se agregan unos pocos OTU´s no constituyen un problema serio, puesto que la media y la varianza no se verían apreciablemente alteradas.

Cuando se agrega un número más grande de OTU´s nuevos, será necesaria una normalización nueva de los caracteres afectados.

El uso de orden de rango de valores de estado del carácter, v.g.: el jth rango de n valores de un carácter determinado, merece análisis, como evitar problemas ciertos asociados con valores extremos o frecuencia muy anormal en las distribuciones. Se observa que las matrices con base en los logaritmos de métricas de caracteres o en caracteres normalizados son sumamente

congruentes. Sin embargo, uno se debería proteger contra caracteres que naturalmente pueden asumir el valor cero.

Si se desea, que una transformación normalmente distribuida con media cero y varianza unitaria pueda ser efectuada por medio de ordenamientos (tenga Orden, donde Orden es la desviación promedio del rth más grande de un muestreo de n observaciones sacadas al azar desde una distribución normal con una media cero y una varianza unitaria).

Hay algunas razones para considerar que el peso de un carácter debe ser inversamente proporcional a su variabilidad. Para una distribución normal de caracteres cuantitativos su información (en el sentido de teoría de la información) es proporcional a la varianza, tal que si las varianzas se hacen iguales, entonces cada carácter contribuye información por igual. En un sentido más general nosotros podemos argumentar que la variación contribuye a la mayoría de la información, y que el tamaño bruto del carácter y el rango de variación debería contribuir poco a la semejanza fenotípica, desde el punto de vista de la información relativa a la taxonomía (la ***equiprobabilidad*** produce la ***entropía*** máxima).

4.2.4. MATRIZ DE SIMILITUD

Una **clase** es definida ordinariamente por la referencia al conjunto de propiedades que es necesario y suficiente (se postula) para los miembros de la clase. Nosotros definimos un grupo **K** desde el punto de vista de un conjunto **G** de propiedades $\mathbf{f_1, f_2,...,f_n}$ en una manera diferente. Supongamos que nosotros tenemos una agregación de objetos (nosotros no los llamaremos aún una clase) tal que:

1. Cada uno posee un gran (pero no especificado) número de propiedades en G.
2. Cada f en G es una propiedad que poseen un gran número de estos objetos; y
3. Ninguna f en G es una propiedad de cada objeto en el agregado.

Podemos entonces decir que una clase es politética si las primera dos condiciones se cumplen y es politética completa si la condición **-3-** también se cumple.

Durante un procedimiento de clustering secuencial acumulativo los valores arbitrarios se reducen de una manera predeterminada; y nosotros extendemos este método para definir una función generalizada de distancia

$\mathbf{D_{j-k} = [(\overline{X}_k - \overline{X}_j)' S_J^{-1} (\overline{X}_k - \overline{X}_j) |S_J|]^{(1/2)}}$ donde $\overline{X}_j$, y $\overline{X}_k$ son los vectores columna que representan las medias para los clusters J y K, respectivamente, para n de variables, S_j es la matriz de varianza de estas variables para el cluster J, y $|S_J|$ es su determinante, la varianza generalizada, para varios OTU´s.

Así, nosotros podemos estimar la semejanza entre pares de OTU´s, adoptando la convención de arreglos de datos para la taxonomía numérica en forma de una matriz de (n x t) , cuyas t columnas representan t OTU´s a ser agrupados en base a la semejanza y cuyas n de filas son n caracteres {Xij}.

El **espacio A** es un espacio de los atributos de n dimensiones.

El **espacio I** es un espacio de individuos u objetos de t dimensiones.

Hay dos técnicas de relación de OTU´s, la asociación de pares de caracteres (filas) pueden examinarse sobre todos los OTU´s (columnas). Esta se denomina **técnica R**. O la asociación de pares de OTU´s (columnas) sobre todos los caracteres (filas), se denomina **técnica Q**.

El espacio no es necesario que sea Euclídeo en sentido estricto, pero debería si es posible ser métrico, que significa que las medidas de las funciones de similitud respetan las propiedades de una métrica, satisface axiomas sobre todos los OTU´s:

1. $F(a, b) \geq 0$, y $F(a, a) = F(b, b) = 0$
2. $F(a, b) = F(b, a)$
3. $F(a, c) \leq F(a, b) + F(b, c)$
4. Si $a \neq b$, entonces $F(a, b) > 0$.

Que establece que idénticos OTU´s son indistinguibles mientras que pueden ser o no indistinguibles, tener relación de simetría y son una semimétrica y una ultramétrica, familiares para distancias Euclídeas que poseen la propiedad de obedecer al teorema de Pitágoras.

Usando distancias Euclídeas (o con métrica de Manhattan) nosotros podemos computar la matriz de Distancia Taxonómica o de Similitud o de Semejanza o Matriz de Coeficientes de Similitud, Matriz en mediante la cual nosotros deseamos encontrar la estructura taxonómica {Sij} de dimensiones (t x t) donde t es el número de OTU´s.

Los clusters son los conjuntos de OTU´s en el hiperespacio, fenotípicos en término de patrones.

El centro del cluster o centroide representa un objeto promedio, que es simplemente una construcción matemática, que permite la caracterización de la Densidad y la Varianza, y el radio y rango del taxon.

Se postula que los coeficientes de correlación o de asociación pueden ser relacionados con las distancias.

En un hiperespacio ***A*** se pueden representar las posiciones de las t OTU´s en un sistema de coordenadas, si dichas posiciones son cercanas, la distancia disminuye hasta hacerse cero si coinciden, así la distancia puede ser vista como el complemento de la similitud, pudiéndose probar algebraicamente que los teoremas de la geometría se cumplen en un hiperespacio Euclídeo de n dimensiones.

A partir de los dominios normalizados se calculan la diferencia media entre caracteres, tomándose el valor absoluto de la diferencia pues esta puede ser negativa, y la distancia taxonómica donde se pueden considerar las métricas de Minkowski y de Manhattan.

$\overline{D}$ = ($\sum_{i}^{n}$ | Xij - Xik |) / n diferencia media entre caracteres.

$\Delta jk = [\sum_{i}^{n} (Xij - Xik)^2]^{1/2}$ distancia Δjk entre OTU´s.

$djk = ((\sum_{i}^{n} (Xij - Xik)^2 / n))^{1/2}$ donde se calcula djk como promedio pues Δjk crece con el número de caracteres.

El valor esperado para (d) djk para una distribución normal de media cero y varianza unitaria es:

$E(d) = ((n - 1)! (\pi / n)^{1/2}) / (2^{n-2} [((n/2) - 1)!]^2)$

Utilizando la fórmula de Stirling:

$E(d) \approx \sqrt{2} (1 - 1/n)^{½} ((1 + (1 / (n - 2)))^{½} (1 / e))$

El valor esperado de la varianza para (d) djk:

$E(\sigma^2_d) = 2 - [E(d)]^2 \approx 1/n.$

Así se llega a la matriz de similitud:

	OTU_1	OTU_2		OTU_t
OTU_1	0	S_{12}		S_{1t}
OTU_2	S_{21}	0		
OTU_j	S_{j1}	S_{j2}	0	S_{jt}
OTU_t	S_{t1}		S_{tt-1}	0

(t x t)

Características:

1. Diagonal Principal:
   - Cada valor de esta diagonal representa cada OTU comparado consigo mismo. Este valor corresponde al de máxima similitud ($S_{ii} = 0$).
2. Matriz Simétrica:
   - La similitud entre OTU_1 y OTU_2 es la misma que entre OTU_2 y OTU_1.

4.2.5. ESPECTROS DE SIMILITUD

Cada fila j de la matriz de similitud contiene las distancias entre el OTU_j y todos los t-1 OTU´s restantes [Perichinsky, Partenkirchen.Germany, 1998 a y b] [Perichinsky, Innsbruck. Austria, 1999, 2000] [Perichinsky, Central Michigan University at Foz do Iguazú. Brazil., 2002] [Perichinsky, Central Michigan University at Río de Janeiro. Brazil., 2003].

Estas similitudes dependen de los valores o estados de los caracteres por el aporte que hacen a la distancia entre las OTU´s.

La distribución de los OTU´s en el hiperespacio taxonómico nos permite visualizar la acumulación de los mismos, por vecindad es decir, vecinos próximos o cercanos (nearest neighbor) por el método de relaciones de similitud entre pares (pair group method) de OTU´s.

En el análisis de estructuración, clustering, las familias por **aglomeración** o reunión de los OTU´s, se van produciendo mediante la agregación, hasta cubrirlos integralmente, de una cantidad de subconjuntos menor que t. Este método es más congruente que el **análisis por asociación** o método por **división**, en una cantidad de conjuntos parte, menor que t, sobre todo por la ubicación de objetos, por **identificación,** en una familia.

El método debe ser **jerárquico**, en secuencias de agrupamientos, clustering, C_0, C_1, C_2,...,C_w donde C_0 es un conjunto de particiones disjuntas y en cada secuencia C_j se forma un conjunto k_j de particiones asociadas que en todos los casos son **no solapadas**, es decir que si un OTU pertenece a una partición disjunta o asociada no puede pertenecer a otra; la secuencia de los agrupamientos debe ser recursiva es decir operaciones **secuenciales,** con criterio **global**, es decir que se considera que todos los caracteres aportan a la similitud y no **locales**, donde existen caracteres predominantes, aunque la solución es **directa**, que significa que el clustering en cualquier nivel se obtiene por soluciones óptimas, sin embargo, una vez que la estructura está establecida en un nivel de clasificación, esta no cambia en etapas posteriores, hay que hacer mención que los procedimientos para alcanzar soluciones óptimas, es difícil que permitan alcanzar una clasificación completa, exigiendo un proceso de **estabilización global**, **local o ambos**, en los cuales se pueden modificar la cantidad de caracteres, y los OTU´s pueden cambiar de un grupo, o partición o cluster o familia, a otro. Esto implica que el clustering es **no adaptativo**, pues el método es fijo y los grupos se forman interactuando con todos los puntos del **espacio A**; y es **pesado**, pues está basado en la cantidad de OTU´s y por lo tanto en la densidad de los clusters, que pueden ser visualizados como densas nubes elipsoidales, cuyos ejes mayores son las distancias, en el hiperespacio taxonómico.

El método de clustering cuyo acrónimo es **SAHN** resume lo expresado anteriormente: **Sequential, Agglomerative, Hierarchic and Nonoverlapping.**

Establezcamos a **C**jk, como un coeficiente general de similitud y como un ejemplo especial a la distancia taxonómica **d**jk. Las distancias Euclideas se usarán en la explicación de técnicas de agrupamiento, porque ellas son fáciles de visualizar geométricamente, aunque no todos los coeficientes de similitud sean necesariamente métricos.

Para discutir procedimientos de agrupamientos aglomerativos nosotros hacemos una distinción útil, entre los siguientes tres tipos de medida: las **medidas-(J)** son aquellas que definen una propiedad de un grupo único o cluster, tal como su centroide, su dispersión, su forma, etc.; las **medidas-(J,K)** son las que estiman la similitud o disimilitud entre dos grupos o entre un OTU y un grupo y; finalmente, las **medidas-(JK,L)** que describen los cambios en algunas medidas cuando dos grupos se fusionan. Un ejemplo sería el aumento en la información que resulta de la fusión de dos clusters separados.

En todo método de agrupamiento SAHN dos consideraciones gobiernan cada paso. Una es el nuevo cómputo del coeficiente de similitud entre clusters nuevos establecidos y candidatos potenciales para ingresos futuros y el otro es el criterio de ingreso para miembros nuevos al cluster establecido. Para todos los métodos **pair-group** el criterio es el mismo y es en base a una **medida-(J,K)**.

Para evaluar el primer criterio nosotros adoptaremos el simbolismo uniforme siguiente. Tendremos en cuenta los clusters **J**, **K** y **L** que contienen $\mathbf{t_j}$, $\mathbf{t_k}$ y $\mathbf{t_l}$ OTU´s, respectivamente, donde $\mathbf{t_j}$, $\mathbf{t_k}$ y $\mathbf{t_l}$ son todos $\geq \mathbf{1}$. Los OTU´s **j** y **k** pertenecen a los clusters **J** y **K**, respectivamente, y $\mathbf{l \in L}$. Dados estos clusters **J** y **K** unidos, el problema está en evaluar la disimilitud entre el cluster fusionado y los candidatos adicionales **L**, para la fusión. El cluster fusionado se denomina **(J, K)**, con $\mathbf{t_{(j,k)}=t_j+t_k}$ OTU´s. Los distintos métodos de clustering difieren en el cómputo propio del coeficiente $\mathbf{C_{(J,\ K),L}}$ de similitud. Nosotros consideramos dentro del SAHN el procedimiento **combinatorio** de clustering, donde el coeficiente $\mathbf{C_{(J,\ K),L}}$ de similitud puede computarse de las similitudes previamente evaluadas $\mathbf{C_{J,L}}$, $\mathbf{C_{K,L}}$, $\mathbf{C_{J,K}}$ y el tamaño de la muestra $\mathbf{t_j}$ y $\mathbf{t_k}$. Con las técnicas combinatorias los clusters más burdos pueden siempre computarse de los clusters más finos previos.

Otro criterio que nosotros tomamos en cuenta es la estrategia de **compatibilidad** de clustering donde la métrica entre clusters más burdos es igual que entre clusters más finos o aún entre OTU´s originales. Así la dimensión del espacio original se mantiene y es simple, para representar clusters en el **espacio-A** original. Denominamos a esta estrategia de clustering de **conservación del espacio**. En las técnicas en las cuales la estrategia es la **distorsión del espacio** parece como si el espacio, en la inmediata vecindad de un cluster se ha contraído o dilatado. Si volvemos al criterio de admisión para un candidato que se une a un cluster existente, este espacio vecino es constante sobre todo en el **método pair-group**. Todo OTU o cluster **J** se unirá a todo OTU o cluster **K** si y sólo si $\mathbf{C_{JK}<C_{JL}}$ y $\mathbf{C_{JK}<C_{KL}}$, donde **L** es cualquier OTU o cluster en estudio (en el nivel actual de clustering) a excepción de **J** o **K**. Esto significa que **J** y **K** son el par mutuamente más cercano de OTU´s o clusters. Los enlaces (links) con clusters contenidos en otros, se hacen con el primer enlace, por convención, para procedimientos computarizados.

Las técnicas de enlaces (links) para formar clusters pueden ser **simples**, con el vecino más próximo (**nearest neighbor**), **completo**, con el vecino más lejano o máximo (**farthest neighbor**), con variantes de encadenamiento y; **enlaces entre promedios** pudiéndose tomar los centroides como par virtual.

4.2.6. CARACTERIZACIÓN

Los estados de los caracteres taxonómicos en una **clase,** definida ordinariamente por la referencia al conjunto de sus propiedades, permiten calcular las distancias y a partir de las distancias se puede establecer por la relación de similitud entre individuos, por interferencia o superposición, que para ese dado conjunto de individuos, en una distribución dada en un hiperespacio, hay una constancia en los parámetros para identificar las características de un cluster e identificar un cluster para cada individuo según sus parámetros.

Considerando características espectrales [Frank,N.H.,1949] [Sawyer, R.A.,1963] (ver Capítulo **5.-**) a los estados de los caracteres o atributos de los OTU´S, en condiciones definidas de los **PRINCIPIOS DE SUPERPOSICIÓN E**

INTERFERENCIA, se introducen los nuevos conceptos de **ESPECTROS DE OBJETOS** y **ESPECTROS DE FAMILIAS**.

En el espacio taxonómico este método de clustering delimita grupos taxonómicos que pueden ser visualizados como espectros característicos de un OTU y espectros característicos de las familias.

Definimos como **espectro individual taxonómico** al conjunto de distancias de un OTU respecto a los demás OTU´s del conjunto, donde cada uno aporta los estados de los caracteres y por lo tanto es constante para cada OTU, en las mismas condiciones taxonómicas (en analogía con los fasores).

Definimos como **espectro de similitud taxonómica** al conjunto de distancias de los OTU´s respecto a los demás OTU´s del conjunto, que determinan las características constantes de un cluster o familia, en las mismas condiciones taxonómicas.

La primera importancia de tener un espectro individual taxonómico, surge del hecho de poder estudiar las propiedades de los individuos a través del aporte de los estados de su caracteres y en segunda instancia, el estudio de la estructura taxonómica al conformar clusters o familias, con características constantes, por la relación métrica o no de los OTU´s, cada una con su espectro de similitud taxonómica.

Se resuelve de esta manera la evidencia taxonómica y se pueden encontrar invariantes, que caracterizan a cada cluster, tales como: la varianza, el radio, la densidad y el centroide.

Estas invariantes están asociadas a los espectros de similitud taxonómica, que identifican a cada familia.

4.2.7. DISPERSIÓN

Una vez conocido un valor típico de la variable de los estados de los caracteres es necesario tener un parámetro que dé una idea de cuan esparcidos, o concentrados, están sus valores respecto al valor medio [Cramer, Harald, 1958]. Se considera a la varianza como como momento de segundo orden y representa al momento de inercia de la distribución de objetos (masa) respecto a su centro de gravedad: centroide .

Cuando $\overline{X'ij}$ = (Xij - $\overline{Xj}$) / σj es una variable normalizada la cual representa la desviación de Xij respecto de su media en unidades de σj.

La normalización de los estados del carácter hace que la media de todo carácter sea de valor cero y varianza de valor unitario.

Si tomamos como valor de la dispersión a la varianza σ^2_d , expresamos el principio de mínimos cuadrados.

Sea g (Xij) una función no negativa de la variable Xij, para todo k > 0 se tendrá la función de probabilidad:

P [g (Xij) ≥ K] ≤ (E (g (Xij)) / K

Así llegamos al Teorema de Tchebycheff.

Si designamos por S al conjunto de todas las Xij que satisfacen la desigualdad

g (Xij) ≥ K

la verdad del teorema surge de la relación (válida para una variable de cualquier número de dimensiones) :

$$Eg(Xij) = \int_{-\infty}^{\infty} g(Xij)dF \geq K\int_{S} dF = KP(S)$$

Si g (Xij) = (Xij - $\overline{Xj}$)^2 , K = k^2 σj^2, obteniendo para todo k > 0 la desigualdad de Bienaymé-Tchebycheff:

P (| Xij - $\overline{Xj}$ | ≥ k. σj) ≤ 1 / k^2

Esta desigualdad muestra que la cantidad de (OTU´s) masa de la distribución debería estar situada en el intervalo:

$\overline{Xj}$ - k. σj < Xij < $\overline{Xj}$ + k. σj

es a lo sumo el valor maximal igual a 1 / k^2, dando una idea de utilización de σj como medida de la dispersión o concentración

En particular para una distribución de media $\overline{Xj}$ y desvío σj que tiene una masa 1 / 2.k^2 en cada uno de los puntos Xij = $\overline{Xj} \pm$ k. σj y [1 - (1 / k^2)] y en un punto de masa (OTU) Xij = $\overline{Xj}$ se tiene:

P (| Xij - $\overline{Xj}$ | ≥ k. σj) = 1 / k^2

valor maximal límite (superior) de la probabilidad que no puede ser mejorado.

La desigualdad permite fijar cotas a la distribución y nos permite fijar el radio de un cluster.

Si hacemos k=2 debemos multiplicar a la varianza por la raíz cuadrada de 2 en la iteración del algoritmo.

4.2.8. NORMALIZACIÓN DEL RANGO DE VARIACIÓN.

Hay algunas razones para considerar que el peso de un carácter debe ser inversamente proporcional a su variabilidad. Para una distribución normal de caracteres cuantitativos su información (en el sentido de Teoría de la Información) es proporcional a la varianza, tal que si las varianzas se hacen iguales, entonces cada carácter contribuye información por igual, la ***equiprobabilidad*** produce la ***entropía*** máxima [Jaynes, 1986] [Perichinsky, 2005, 2006, 2007].

En un sentido más general nosotros podemos argumentar que la variación contribuye a la mayoría de la información, y que el tamaño bruto del carácter y el rango de variación debería contribuir poco a la semejanza fenotípica, desde el punto de vista de la información relativa a la taxonomía.

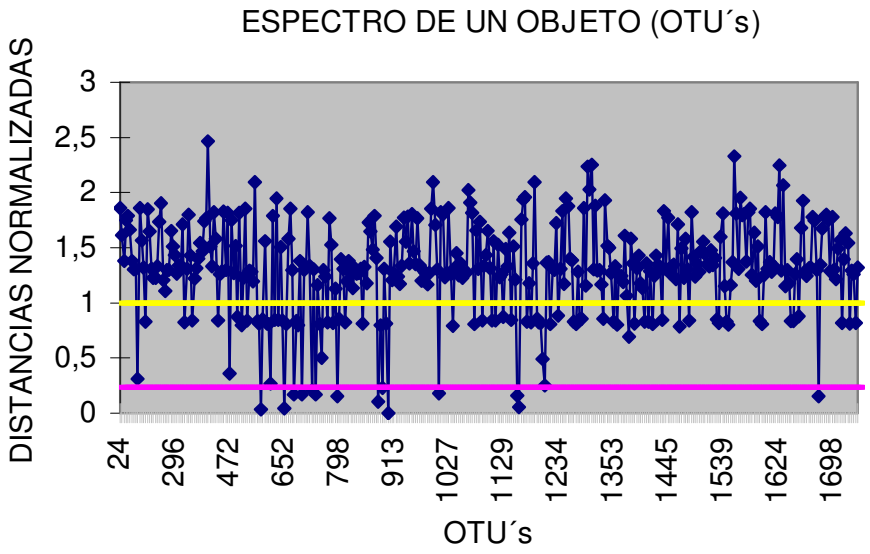
4.2.9. ALGORITMO.

La secuencia algorítmica está dada por la conformación de la matriz de datos, su normalización, la construcción de la matriz de similitud, los espectros de totales y de familias formadas por clustering y el análisis de invariantes.

1. Matriz de Datos
2. Selección de Caracteres
3. Construcción del Dominio de Atributos
4. Normalización
5. Análisis de Distancias
6. Matriz de Similitud
7. Análisis de la Dispersión
8. Identificación de OTU´s en las familias (clusters)
9. Análisis de Invariantes
10. Espectro Característico de los Objetos
11. Espectro Característico de Familia (Cluster)
12. Iteración alrededor del centroide

4.3. COROLARIO

El tratamiento dinámico e integrado de los dominios permite una fácil normalización, atributo - dominio - valor, y la implementación en el modelo de Base de Datos Dinámica y su utilización en Taxonomía Numérica.
La contribución teórica - empírica es la aglomeración de objetos formando clases producidas por pasos del método **(ALGORITMO)** obteniendo clusters y dominios con valores normalizados y la **densidad y el rango** en términos del **radio** del conjunto puede ser visualizado como una **INVARIANTE CARACTERÍSTICA de los OTU´s.**



Invariantes:

- Distancia media: 0.1321
- Densidad: 13
- Dispersión: 0.059
- Rango: 0.2343

Si se observa el gráfico para la recta de Invariante igual a uno (1) en el rango se tiene una región que claramente muestra los objetos que la forman, por encima de la recta hay objetos de otras regiones. Por debajo de la recta de valor del

rango 0.2343 se tienen objetos de una familia. Quedan objetos entre ambas rectas que hay que analizar a que familia pertenecen.

4.4. Testeo de Minería de Datos Inteligente (Intelligent Data Mining).

Un sistema del software fue construido para evaluar el algoritmo de C4.5 (ver el Capítulo de Estado del Arte 2.2.7.) [Michie, 1986, 1988] [Quinlan, 1986-1996] [Mitchell, 1997, 2000a, 2000b] [Perichinsky, 203, 2005, 2006, 2007]. Este sistema toma los datos de entrenamiento como una entrada y le permite al usuario escoger si él quiere construir un árbol de decisión según el C4.5. Si el usuario escoge el C4.5, el árbol de decisión se genera, entonces se recorta y las reglas de decisión se construyen.
Se evalúan el árbol de decisión y el conjunto de reglas generados por el C4.5 por separado cada uno de los otros.
Nosotros usamos el sistema para probar los algoritmos en los dominios diferentes, principalmente Elita, una base de asteroides.

4.4.1. Cómputo de la Ganancia de la Información.

En los casos, aquéllos en los cuales el conjunto T contiene ejemplos que pertenecen a clases diferentes, se ha llevado a cabo un testeo de los diferentes atributos y se ha logrado una partición congruente con el "mejor" atributo. Para encontrar el mejor atributo, se usa la teoría de la información que sostiene que la información es maximizada cuando la entropía se minimiza. La entropía determina la aleatoriedad o desorden de un conjunto.
Supongo que se tiene ejemplos negativos y positivos. En este contexto la entropía del subconjunto Si, H(Si), puede calcularse como:

$$H(S_i) = -p_i^+ \log p_i^+ - p_i^- \log p_i^- \quad \textbf{(4.4.1.1.)}$$

Donde p_i^+ es la probabilidad de que un ejemplo que fue tomado al azar de S_i será positivo. Esa probabilidad puede ser calculada como

$$p_i^+ = \frac{n_i^+}{n_i^+ + n_i^-} \quad \textbf{(4.4.1.2.)}$$

Siendo n_i^+ la cantidad de ejemplos positivos de S_i, y n_i^- la cantidad de ejemplos negativos.

La probabilidad p_i^- es calculada en forma análoga que p_i^+, reemplazando la cantidad de ejemplos positivos por la cantidad de ejemplos negativos y viceversa. Generalizando la expresión ***(4.4.1.1.)*** para cualquier tipo de ejemplos, se obtiene una ecuación general para la entropía:

$$H(S_i) = \sum_{i=1}^{n} - p_i \log p_i \quad \textbf{(4.4.1.3.)}$$

En todos los cálculos relacionados con la entropía, se define ***0log0*** *igual a* ***0****.*

Si el atributo ***at*** divide al conjunto *S* en los subconjuntos S_i, i = 1, 2, …, j , …, n, luego, la entropía total del sistema de subconjuntos será:

$$H(S, at) = \sum_{i=1}^{n} P(S_i) \cdot H(S_i) \quad \textbf{(4.4.1.4.)}$$

Donde $H(S_i)$ es la entropía del subconjunto S_i y $P(S_i)$ es la probabilidad del hecho de que un ejemplo pertenezca a S_i. Esa probabilidad puede ser calculada utilizando las dimensiones relativas de los subconjuntos, como:

$$P(S_i) = \frac{|S_i|}{|S|} \quad \textbf{(4.4.1.5.)}$$

La ganancia de información puede ser calculada como el decremento en entropía. Así:

$$I(S, at) = H(S) - H(S, at) \quad \textbf{(4.4.1.6.)}$$

Donde $H(S)$ es el valor de la entropía a priori, antes de realizar la subdivisión, y $H(S, at)$ es el valor de la entropía del sistema de subconjuntos generados por la partición acorde con ***at****.*

El uso de la entropía para evaluar el mejor atributo no es el único un método existente o usado en el Aprendizaje Automático. Sin embargo, fue usado por Quinlan al desarrollar el ID3 y su exitoso C4.5.

4.4.2. Datos numéricos

Los árboles de decisión pueden generarse tanto con atributos discretos como con atributos continuos. Cuando ha funcionado con atributos discretos, la partición del conjunto según el valor de un atributo es simple.

Para resolver este problema, puede recurrirse al método binario. Este método consiste en formar dos rangos de valores de acuerdo al valor de un atributo que pueden tomarse como simbólicos.

4.4.3. Resultados del C4.5

El C4.5 con post-poda resulta en árboles más pequeños y menos frondosos o espesos. Si se analizan los árboles obtenidos en el dominio tratado (ELITA), se ve que los porcentajes de error obtenidos con el C4.5 está entre un 3% y un 3.7%, puesto que los C4.5 generan árboles menores y los conjuntos de reglas también menores. Derivado del hecho que cada hoja en un árbol generó las envolventes de una distribución de clases.

4.4.4. Porcentaje de Error

Dado {ELITA}, aplicando C4.5, la Ganancia en los árboles, la Ganancia en las reglas, la Ganancia en la proporción de los árboles, la Proporción de reglas de la Ganancia de los árboles siempre da menor a un 3%.
Del análisis de este valor pude concluir que ningún método puede generar un modelo claramente superior para el dominio. Al contrario, se podría establecer que el porcentaje de error no parece depender del método usado, pero en el dominio que se analice.

4.4.5. Espacio de las Hipótesis

El conjunto de hipótesis para este algoritmo está completo según los atributos disponibles. Porque cualquier prueba de valor puede representarse con un árbol de decisión, este algoritmo evita uno de los riesgos principales del método inductivo, pues trabaja reduciendo el conjunto de hipótesis.
Un rasgo importante del algoritmo de C4.5 es que usa todos los datos disponibles en cada paso al escoger el "mejor" atributo; esta es una decisión que está hecha

con método estadístico. Este hecho favorece a este algoritmo por encima de otros algoritmos porque analiza como los conjuntos de datos de entrada se representan como árboles de decisión en forma consistente.

Una vez que un atributo ha sido seleccionado como un nodo de decisión, el algoritmo no va hacia atrás remontando por encima de sus opciones ya tomadas. Esta es la razón por qué este algoritmo puede converger a un máximo local [Mitchell, 2000]. El algoritmo de C4.5 agrega un cierto grado de reconsideración de sus opciones en la post-poda de los árboles de decisión.

No obstante, se puede establecer que los resultados muestran que la proporción de error depende del dominio de los datos. Para estudios futuros, se puede pensar en un análisis del conjunto de datos de entrada, agrupando con este método numérico y escoger dominios, si el método mantiene un porcentaje bajo de error, en bases de datos extendidas, es una prueba de la robustez del método.

FENOMENOLOGÍA FÍSICA

Tengo la impresión de que el intento de la naturaleza de crear ... en este mundo un ser pensante ha fracasado...

Max Born

CLXXIV

5. FENOMENOLOGÍA FÍSICA

5.1. PRINCIPIOS DE INTERFERENCIA Y SUPERPOSICIÓN.

El peso de Sir Isaac Newton ignorando la teoría ondulatoria de la luz cayó frente a la teoría ondulatoria del Doctor Thomas Young en el siglo XIX, a la cual le añadió además el nuevo concepto del ***principio de Interferencia*** [Hecht et al., 1977], [Feynman et al., 1971]***.***

Cuando dos ondulaciones de diferentes orígenes coinciden perfectamente en una dirección o casi coinciden, su efecto conjunto es una combinación de los movimientos que pertenecen a cada uno.

Agustín Jean Fresnel sintetizó la teoría ondulatoria y el principio de interferencia, conceptualizando que la propagación de una onda primaria como una sucesión de onditas secundarias estimuladas que se superponían e interferían para reformar en su avance a la onda primaria. Los problemas de isotropía y de patrones de difracción fueron explicados satisfactoriamente.
La asimetría lateral y el efecto de la polarización en la interferencia son una manifestación de dos vibraciones ortogonales y transversales a la dirección de la luz.

Primero Michael Faraday estableció la interrelación entre el electromagnetismo y la luz y luego James Clerk Maxwell resumió el conocimiento empírico y lo amplió y lo formalizó con un conjunto de ecuaciones matemáticas {ver [Hecht et al., 1977.], [Feynman et al., 1971.]}, llegando teóricamente a una expresión de la velocidad de la luz en términos de las propiedades eléctricas y magnéticas del medio

$c = 1/\sqrt{\varepsilon_\circ \mu_\circ}$. La combinación de la óptica con la electrónica llevó a la electro-óptica u opto-electrónica.

El siguiente aspecto a considerar es la base conceptual de que sucede cuando dos o más ondas de luz se superponen en la misma región del espacio. Las circunstancias que gobiernan la superposición determinan la perturbación óptica final o perturbación compuesta.

Cualquier combinación lineal de ondas individuales de $\Psi(r,t)$, $\Psi_1(r,t)$, $\Psi_2(r,t)$, $\Psi_3(r,t)$... es una solución $\Psi(r,t) = \sum C_i \Psi_i(r,t)$ que se conoce como ***principio de superposición***.

Responde a una ecuación diferencial de la forma:

$$(\partial^2\Psi / \partial x^2) + (\partial^2\Psi / \partial y^2) + (\partial^2\Psi / \partial z^2) = (1 / v^2) \cdot (\partial^2\Psi / \partial t^2)$$

La perturbación resultante en cualquier punto de un medio es la suma algebraica de sus ondas constitutivas separadas.

Según el método algebraico se produce una onda compuesta armónica y de la misma frecuencia que las constitutivas aunque su amplitud y fase son diferentes.

Si se aplica el método complejo se obtiene la amplitud compleja exponencial, pues matemáticamente se usa la representación compleja de las funciones trigonométricas cuando se está manejando la superposición de perturbaciones armónicas.

$$E \cdot e^{i\alpha} = \sum_{j=1}^{N} E_o \cdot e^{i\alpha j}$$

En ingeniería se conoce la amplitud compleja como ***fasor*** y se especifica por su magnitud y su fase. La suma compleja es una suma de vectores en el plano complejo.

En conclusión la inestabilidad del campo resultante en un punto donde hay superposición, es la suma vectorial de las perturbaciones constitutivas individuales.

El patrón de interferencia es deseable que sea estable aunque lo es por lapsos según los cambios de fases, si es observable.

Los patrones de interferencia más claros son los que surgen de amplitudes casi iguales en un campo de onda estacionario. En general cuando existe asimetría aparecen segundas y terceras armónicas con una inversión en ejes de coordenadas creando un centro de inversión o simetría.

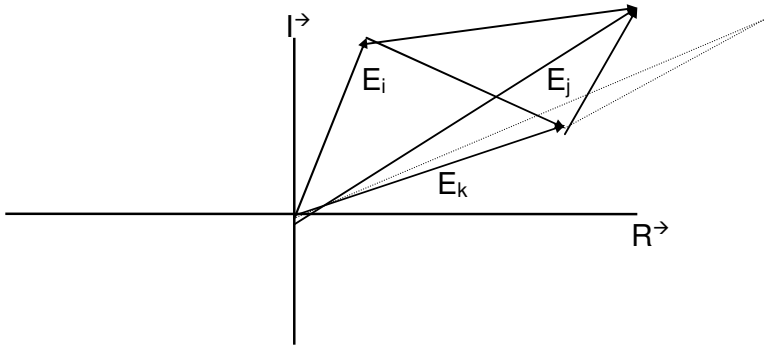


Fig. Fasores y sumas

El término de interferencia, debido al efecto de la interferencia, depende del coseno del ángulo de desfasaje (δ), lo cual puede producir interferencia positiva o negativa, y para determinados valores anular la amplitud resultante o compuesta.

$$E_{12} = E_1 + E_2 + 2\sqrt{E_1 E_2}\,.\cos\delta$$

En el caso del principio de superposición para sistemas complejos es que este sistema se puede descomponer de cualquier manera conveniente, en la suma de partes separadas y de un mismo modo simple cualquier parte especial, entonces el resultado es válido para el sistema total, pues podemos volver a juntar las partes como solución con carácter equivalente al sistema total.

Desde el punto de vista **experimental** Newton introdujo el primer concepto de espectro y de rayos homogéneos y heterogéneos de mezcla de rayos distinguibles como parte de otro o de clases, basado en el concepto de la naturaleza del color [Newton, I., 1672] en base a su teoría corpuscular, pero no reconoció las líneas espectrales y la teoría ondulatoria de la luz y por su prestigio e influencia científica produjo una demora de casi 150 años en la espectroscopía.

Recién con los trabajos de la naturaleza de los colores y de las líneas espectrales a partir de la teoría ondulatoria de la luz y los principios de superposición e interferencia Thomas Young y Joseph Fraunhofer aportan un gran desarrollo experimental detectando una gran número de líneas espectrales verticales débiles y fuertes, en particular dentro de las **Ciencias Astrofísicas** espectros de bandas y líneas de cuerpos celestes {ver [Hecht et al., 1977], [Feynman et al., 1971.], [Sawyer, R.A., 1963.]}.

La base científica de la espectroscopía es que los átomos y las moléculas tienen sus espectros característicos reconocibles.

La ley de Kirchhoff (G.R. Kirchhoff fundador del análisis espectroscópico junto con R. Bunsen y los trabajos experimentales de A.J. Ångström) es un hito importante, al establecer que la relación de las potencias de emisión y absorción de rayos de la misma longitud de onda es constante en cuerpos a la misma temperatura y son características de sus componentes y estructuras.

5.2. ANALOGÍAS

Los estados de los caracteres taxonómicos, en una **clase,** definida ordinariamente por la referencia al conjunto de sus propiedades, permiten calcular las distancias y a partir de las distancias se puede establecer por la relación de similitud entre individuos, por interferencia o superposición, que para ese dado conjunto de individuos, en una distribución dada en un hiperespacio, hay una constancia en los parámetros para identificar las características de un cluster e identificar un cluster para cada individuo según sus parámetros.

Considerando características espectrales [Frank, N.H., 1949] [Sawyer, R.A., 1963.] a los estados de los caracteres o atributos de los OTU´s, en condiciones definidas de los **PRINCIPIOS DE SUPERPOSICIÓN E INTERFERENCIA**, se

introducen los nuevos conceptos de **ESPECTROS DE OBJETOS** y **ESPECTROS DE FAMILIAS**.

En el espacio taxonómico este método de clustering delimita grupos taxonómicos que pueden ser visualizados como espectros característicos de un OTU y espectros característicos de las familias.

Definimos como **espectro individual taxonómico** al conjunto de distancias de un OTU respecto a los demás OTU´s del conjunto, donde cada uno aporta según los estados de sus caracteres y por lo tanto es constante para cada OTU, en las mismas condiciones taxonómicas. La figura geométrica de distancias es análoga a la que se ve en el párrafo anterior para los fasores.
Definimos como **espectro de similitud taxonómica** al conjunto de distancias de los OTU´s respecto a los demás OTU´s del conjunto, que determinan las características constantes de un cluster o familia (estructura), en las mismas condiciones taxonómicas.
La primera importancia de tener un espectro individual taxonómico, surge del hecho de poder estudiar las propiedades de los individuos a través del aporte de los estados de su caracteres y en segunda instancia, el estudio de la estructura taxonómica al conformar clusters o familias, con características constantes, por la relación métrica o no de los OTU´s, cada una con su espectro de similitud taxonómica.
Se resuelve de esta manera la evidencia taxonómica y se pueden encontrar invariantes, que caracterizan a cada cluster, tales como: la varianza, el radio, la densidad y el centroide.
Estas invariantes están asociadas a los espectros de similitud taxonómica, que identifican a cada familia.

SISTEMAS COMPLEJOS Y DINÁMICOS, MECÁNICA ESTADÍSTICA Y TEORÍA DE LA INFORMACIÓN

La ciencia... en cuanto se aplica al mejoramiento de nuestro medio natural y artificial, a la invención y manufactura de bienes materiales y culturales, se convierte en tecnología.
Mario Bunge
La ciencia, su método y filosofía

CLXXX

HERRAMIENTAS DE LA MECÁNICA ESTADÍSTICA, DE LA TEORÍA DE LA INFORMACIÓN Y DE LAS CIENCIAS DE LA COMPUTACIÓN.

6. SISTEMAS COMPLEJOS Y DINÁMICOS.

6.1. Conceptos

El estudio de los sistemas complejos en un esquema unificado ha sido reconocido como una nueva disciplina científica, dentro del contexto de los campos multidisciplinarios. Este tipo de sistemas incluye áreas tan diversas como ecosistemas, computadoras, la sociedad humana y su economía, el clima. Las herramientas para el estudio de estos sistemas, son variadas y en la presente tesis se utilizan en forma conjunta técnicas de la mecánica estadística, teoría de la información, ciencias de la computación e informática y ciencias sociales. Las áreas a cubrir son: descripción de sistemas humanos, descripción de sistemas físico-químicos, desarrollos de los fundamentos para las teorías utilizadas y desarrollo y aplicación de nuevas herramientas computacionales.

Para los sistemas humanos [Acedo, 1997.] [Weidlich, 1991] conviene definir que tipo de modelización se realiza del sistema. Si bien hay varias formas de definir que es un modelo, trataremos de dar la más simple que es decir que un modelo es dar un marco formal al conjunto de hipótesis que surgen de un determinado conjunto de observaciones. A menudo estas hipótesis buscan solo identificar las funciones que permiten reproducir los datos observados. Se dice entonces que es un modelo empírico. En otras se trata de identificar los mecanismos que se supone generan determinados datos, y generar "predicciones" sobre el comportamiento de un determinado sistema. En todos los casos el modelo es una representación del sistema. Un sistema es un conjunto de entidades vinculadas entre si por relaciones. El desarrollo de un modelo es un proceso característico de prueba y error, y se desarrolla sobre la base del mundo real. Mediante simulaciones y/o análisis del modelo, se busca obtener resultados que reproduzcan los datos. Así por ejemplo, para el caso

de los sistemas político-económico-sociales, es posible definir una estrategia general para la construcción del modelo cuantitativo, o al menos semicuantitativo de evoluciones macrodinámicas en la sociedad, utilizando diversos conceptos provenientes de la mecánica estadística. Las definiciones cualitativas sobre el comportamiento humano individual y colectivo pueden asociarse a micro y macro variables de los sistemas físicos, y ningún modelo cuantitativo puede prescindir de las definiciones cualitativas que caracterizan al comportamiento social o político cuya modelización se pretende hacer.

6.2. Objetivos

- Estudio de evolución temporal de sistemas complejos en un esquema unificado con herramientas computacionales y de la mecánica estadística y teoría de la información.
- Descripción dinámica de sistemas económico-político-sociales y empresariales.
- Aspectos relativos a la fundamentación de la mecánica estadística y la informática.
- Desarrollo y aplicación de nuevos algoritmos computacionales.
- Modelización de sistemas socioeconómicos y empresariales.

Los antecedentes de los trabajos permitieron el desarrollo de la tesis en las áreas propuestas, estableciendo que siempre es posible cumplir, con las tareas y disciplinas relacionadas y mencionadas en un programa de investigación científica.

6.3. Mecánica estadística y Teoría de la información.

6.3.1. Mecánica estadística

La mecánica estadística es una rama de la Física que estudia los sistemas contienen un elevado número de partículas, tantas que, normalmente, no es posible determinar las propiedades globales del sistema evaluando la contribución de cada partícula individualmente. En lugar de hacer esto, las propiedades totales se determinan a partir del comportamiento medio de un

cierto número de sistemas idénticos. Una colección de sistemas idénticos se denomina agrupación, **"*ensemble"*,** y está caracterizada por la media de los sistemas que la forman con las fluctuaciones estadísticas respecto a la media. Por ejemplo la energía media de una agrupación es la media estadística de las energías de les constituyentes de la agrupación, cada una de las cuales está ponderada por la probabilidad de que el sistema tenga una cierta energía. Si P_r es la probabilidad de que un sistema tenga la energía E_r entonces la energía media de la agrupación de estos sistemas es:

$$< E >= \sum_r P_r E_r \qquad \textbf{6.3.1.1.}$$

Una clase muy importante de sistemas es la formada por los que están en contacto con un sistema mucho mayor denominado fuente de calor *(heat reservoir).* Las fuentes de calor se caracterizan por el hecho consistente en que toda interacción térmica con el sistema (más pequeño) en cuestión da lugar únicamente a cambios infinitesimales de las propiedades de la fuente, mientras que el sistema pequeño puede sufrir cambios de importancia, hasta que se alcancen unas condiciones de equilibrio. Un ejemplo sería un objeto en un cluster de características disímiles. A medida que se alcanza el equilibrio, la ubicación del objeto puede sufrir cambios notables, mientras que las características del cluster global cambiará únicamente en una cantidad imposible de medir.

Si examinásemos un gran número de objetos, agrupados en clusters idénticos (o en el mismo cluster, si fuera muy grande en comparación con el conjunto de objetos total), encontraríamos algunas variaciones en la energía total de los objetos. Además, encontraríamos que la probabilidad *Pr* de que un objeto tuviese una cierta energía *Er* iba a ser proporcional a un factor exponencial:

$Pr = Ce^{-\beta Er}$,

donde p es un parámetro que depende de la temperatura del cluster. Dado que la suma de todas estas probabilidades debe ser igual a uno (1), $\sum_r P_r E_r$ **= 1**, la constante de proporcionalidad tiene que ser

$C = (\sum_r e^{-\beta Er})^{-1}$

y, por tanto,

$$P_r = \frac{e^{-\beta E_r}}{\sum_r e^{-\beta E_r}} \qquad \textbf{6.3.1.2.}$$

La magnitud C^{-1} tiene un nombre y un símbolo especial en mecánica estadística, se denomina función de partición y el símbolo que se le suele asignar es Z:

$$Z = e^{-\beta Er} \qquad \textbf{6.3.1.3.}$$

Z es el factor de Boltzmann y la distribución de probabilidades está dada por la ecuación **6.3.1.2.** es la distribución de Boltzmann y β la constante de Boltzmann, aplicable en la Teoría de la Información, permite calcular la entropía y la entropía máxima para la información está dada para símbolos equiprobables, como se ve también en los capítulos de la solución propuesta (4) y de la aplicación (7) en un caso de uso, respectivamente, de la tesis.
La mecánica estadística emplea principios estadísticos para predecir y describir el movimiento de las partículas.

Esta mecánica fue desarrollada en el siglo XIX, fundamentalmente por James Clerk Maxwell, Ludwig Boltzmann y J. Willard Gibbs. Estos científicos trataron de explicar el "***problema de muchos cuerpos"*** (Many Body Problem, objects) conjeturan que los sistemas están compuestos por caso la materia se compone de muchas partículas minúsculas (átomos y moléculas) en movimiento constante. Sabían que era imposible determinar los movimientos de las partículas suponiendo que cada partícula individual se comporta según la mecánica newtoniana, ya que cualquier muestra de materia contiene un número enorme de partículas. Hallaron que la entropía es proporcional al logaritmo del número de formas en que se puede ordenar microscópicamente un sistema macroscópico dado.

Como se expresara en otra sección en los primeros 30 años del siglo XX, se fijaron nuevos conceptos en todas las disciplinas, en este caso hubo que ampliar la mecánica estadística para incorporar los nuevos principios de la teoría cuántica. En esta teoría, la naturaleza de las partículas es diferente a la de la física clásica, que se basa en las leyes del movimiento de Newton. En

particular, dos partículas clásicas son distinguibles en principio; en cambio, dos partículas cuánticas idénticas son indistinguibles, incluso en principio, y eso exige una nueva formulación de la mecánica estadística. La formulación empleada para describir el comportamiento de un grupo de partículas clásicas se denomina estadística de Maxwell-Boltzmann. Las dos formulaciones de la mecánica estadística empleadas para describir las partículas cuánticas son la estadística de Fermi-Dirac, y la estadística de Bose-Einstein, según las características (propiedades) de las partículas (fermiones y bosones).

Los fermiones - partículas con espín no entero - cumplen el principio de exclusión de Pauli, que afirma que dos fermiones no pueden estar en el mismo estado cuántico. En cambio, los bosones - partículas con espín entero - no cumplen el principio de exclusión de Pauli.

6.3.2. Teoría de la información.

Es la teoría relacionada con las leyes matemáticas que rige la transmisión y el procesamiento de la información. Más concretamente, la teoría de la información se ocupa de la medición de la información y de la representación de la misma, como, su codificación, y de la capacidad de los sistemas de comunicación para transmitir y procesar información.

La codificación puede referirse tanto a la transformación de voz o imagen en señales eléctricas o electromagnéticas, como al cifrado de mensajes para asegurar su privacidad.

La teoría de la información fue desarrollada inicialmente, por Claude E. Shannon, en su artículo, A Mathematical Theory of Communication [Shannon, 1948, 1963] [Abramson, N., 1966]. La necesidad de una base teórica para la tecnología de la comunicación surgió del aumento de la complejidad y de la masificación de las vías de comunicación (teléfono, redes de teletipo, radio, teoría de la información también de todas las formas de transmisión y almacenamiento de información, impulsos eléctricos que se transmiten en las computadoras y en la grabación óptica de datos e imágenes). El término información se refiere a los mensajes transmitidos, información digital en

sistemas y redes de computadoras, e incluso a los impulsos nerviosos en organismos vivientes.

De forma más general, la teoría de la información ha sido aplicada en campos tan diversos como la cibernética, la criptografía, la lingüística, la psicología y la estadística.

El sistema de comunicación consta de varios componentes, tales como: (1) una fuente de información, que produce un mensaje o información que será transmitida, (2) un transmisor que convierte el mensaje en señales electrónicas o electromagnéticas, (3) las señales son transmitidas a través de un canal o medio, (cable o espacio, que puede tener distorsión o ruido por interferencias y degradación), (4) otro componente es el receptor, que transforma de nuevo la señal recibida en el mensaje original y finalmente (5) es el destinatario, sistema interno o externo o usuario de la información o mensaje.

Dos de las principales preocupaciones en la teoría de la información son la reducción de errores por interferencias en los sistemas de comunicación, y el uso más eficiente de la capacidad total del canal.

6.3.2.1. La cantidad de información

Un concepto fundamental en la teoría de la información es que la cantidad de información contenida en un mensaje es un valor matemático bien definido y medible. El término cantidad es la probabilidad de que un mensaje, dentro de un conjunto de mensajes posibles, sea recibido. En lo que se refiere a la cantidad de información, el valor más alto se le asigna al mensaje que menos probabilidades tiene de ser recibido. Si se sabe con certeza que un mensaje va a ser recibido, su cantidad de información es 0. Si, por ejemplo, se lanza una moneda al aire, el mensaje conjunto cara o cruz que describe el resultado, no tiene cantidad de información. Sin embargo, los dos mensajes por separado cara o cruz tienen probabilidades iguales de valor un medio. Para relacionar la cantidad de información (I) con la probabilidad, Shannon presentó la siguiente fórmula:

$I = \log_2 1/p$ donde p es la probabilidad del mensaje que se transmite.

Para transmitir s se tiene $I(s) = \log_2 1/p(s)$ con $s = (s_1, s_2, \ldots, s_n)$ de probabilidad fija de: $\{P(s_1), P(s_2), \ldots, P(s_n)\}$ alfabeto generado con una distribución de probabilidad.

Con esta fórmula, obtenemos que los mensajes cara y cruz tienen una cantidad de información de $\log_2 2 = 1$ para el primer caso y la cantidad de información recibida:

$$< I >= \sum_{i=1}^{n} P(s_i)\, I(s_i) = - \sum_{i=1}^{n} P(s_i) \log_2 \frac{1}{P(s_i)} \quad \textbf{6.3.2.1.1.}$$

La cantidad de información de un mensaje puede ser entendida como el número de símbolos posibles que representan el mensaje. En el ejemplo anterior, si cruz está representado por un 0 y cara por un 1, sólo hay una forma de representar el mensaje: 0 o 1. El 0 y el 1 son los dígitos del sistema binario, y la elección entre estos dos símbolos corresponde a la llamada unidad de información binaria o bit. La probabilidad de cada mensaje es el número de bits que se necesitan para representar cada mensaje.

6.3.2.2. Entropía

En la mayoría de las aplicaciones prácticas, hay que elegir entre mensajes que tienen diferentes probabilidades de ser enviados. El término entropía ha sido tomado prestado de la termodinámica, para designar la cantidad de información media de estos mensajes. La entropía puede ser intuitivamente entendida como el grado de 'desorden' en un sistema. En la teoría de la información la entropía de un mensaje es igual a su cantidad de información media. Si en un conjunto de mensajes, sus probabilidades son iguales, la fórmula para calcular la entropía total sería: $H = \log_2 N$, donde N es el número de mensajes posibles en el conjunto.

Se tiene de acuerdo a la ecuación **6.3.2.1.1.** la entropía**:**

$$< I >= \sum_{i=1}^{n} P(s_i)\, I(s_i) = - \sum_{i=1}^{n} P(s_i) \log_2 \frac{1}{P(s_i)} = \mathrm{H(S)} \quad \textbf{6.3.2.2.1.}$$

Para símbolos equiprobables en una fuente de n símbolos la entropía es máxima, esto surge de restar las entropías de dos fuentes, S_1 y S_2, con la misma cantidad de caracteres n:

$$H_1(S_1) - H_2(S_2) = -\sum_{i=1}^{n} P(s_{1i}) \log_2 P(s_{1i}) - P(s_{2i}) \log_2 P(s_{2i}) \quad \textbf{6.3.2.2.2.}$$

Como la sumatoria de las probabilidades es igual a uno (1) y sumando y restando dos entropías equiprobables en **6.3.2.2.2.** se tiene para los n símbolos que $H_2(S_2) = -\log_2 n$ por lo tanto la ganancia de información G es:

$$H_1(S_1) - (-\log_2 n) = \sum_{i=1}^{n} P(s_{1i}) \log_2 \frac{P(s_{2i})}{P(s_{1i})} = G \leq 0 \quad \textbf{6.3.2.2.3.}$$

La entropía máxima si la cantidad de caracteres es n y S_1 es también equiprobable y las sumatorias de las probabilidades es uno (1) y la resta de entropías da cero (0), si no $H_1(S_1) < H_2(S_2)$ siendo esta última equiprobable.
G, si son dos fuentes arbritarias, igual a la ecuación **6.3.2.2.2.**

6.3.2.3. Codificación y Redundancia

Si se transmiten mensajes que están formados por combinaciones aleatorias de un alfabeto, el espacio en blanco y signos de puntuación, y si suponemos que la probabilidad de cada mensaje es la misma, la entropía, H, significa cuantos bits se necesitan para codificar cada carácter o mensaje. Una transmisión y almacenamiento eficiente de la información exige la reducción del número de bits utilizados en su codificación. Esto es posible cuando se codifican textos donde la colocación de los símbolos no es aleatoria. La probabilidad de que un símbolo suceda en la secuencia de la información es muy alta.

Se puede demostrar que la entropía de un idioma es de un bit o muy pocos bits por palabra. Por lo tanto todo lenguaje natural tiene una gran cantidad de redundancia incorporada, que se denomina redundancia natural. Esta redundancia permite, por ejemplo, a una persona entender mensajes en los cuales faltan vocales, así como descifrar escritura poco legible. En los sistemas de comunicación modernos, se añade redundancia artificial (cantidad de bits) a

la codificación de mensajes, para reducir errores en la transmisión de los mismos.

6.4. Metodología: Principio de máxima entropía

El principio de máxima entropía (PME) es ampliamente aplicado no solo en física, sino también en meteorología, genética, y en general en procesos de cualquier índole donde se desea obtener información a partir de un conjunto incompleto de datos, o bien utilizando la menor cantidad de suposiciones previas. Para el caso de los sistemas físicos, el PME, provee una formulación alternativa de la Mecánica Estadística, elegante y compacta, formulada por Jaynes [Jaynes, 1957, 1986.]. En su propuesta, Jaynes, presenta al PME como un método canónico para construir matrices densidad, en términos de las variables cuyos valores medios son conocidos "a priori". Sin embargo esta forma constructiva no permite asegurar que la matriz densidad encontrada pueda reproducir valores medios de otras magnitudes del sistema en estudio, que no formen parte de la información a priori. Esta falencia fue la mayor critica a la reformulación de la ME en términos del PME. Posteriormente, [Alhassid y Levine, 1977.] remueven esta limitación (con una formulación cuántica, aplicable también al caso clásico) dando un método especifico para encontrar todos los operadores "relevantes" del sistema, esto es, todos los operadores necesarios para describir completamente la dinámica del sistema. El operador densidad así construido, vale para temperatura diferente de cero y fuera del equilibrio, tanto para el caso de sistemas clásicos, como cuánticos, dependientes o no del tiempo, siguiéndose así esta mitología.

6.4.1. Principio de Máxima Entropía (PME) y Mecánica Estadística

La Teoría de la Información, tal como la concibió Jaynes [Jaynes, 1957, 1986.], provee un criterio constructivo para hallar distribuciones de probabilidad sobre la base de un conocimiento parcial (que se tiene de un sistema en estudio) y, conduce a un tipo de inferencia que se denomina estimación de máxima entropía. A esta estimación se la considera como la conjetura más adecuada compatible con la información dada. Dentro del contexto Jaynesiano, se interpreta que la entropía es una medida de la cantidad de incerteza o falta de

información representada por la distribución de probabilidad; esto convierte a la entropía en un concepto primitivo más fundamental aún que la energía.

6.5. Principio de Máxima Entropía (PME)

La formulación de Levine se basa en que los operadores relevantes, son aquellos que satisfacen la relación de clausura del algebra, con el hamiltoniano del sistema. La principal critica, a esta metodología era que los hamiltonianos que la cumplen son generalmente simples.

Afortunadamente, se pudo remover esa limitación resolviendo, con el PME, la dinámica de hamiltonianos no triviales, como el de Jaynes-Cummings [Gruver, 1994] así como dos niveles acoplados a un baño térmico finito y discreto [Aliaga, 1991] En estos casos las algebras de operadores asociadas son infinitas. El PME permite encontrar la evolución temporal de valores medios de operadores, sin utilizar la función de onda, mediante la generalización del teorema de Ehrenfest [Proto, 1989] [Kowalski, A.N, 1994] que conduce a un sistema de ecuaciones diferenciales. Los hamiltonianos pueden además ser dependientes del tiempo [Proto, A.N, 1989]. Por la forma en que son construidos, estos sistemas de ecuaciones, no permiten el uso de condiciones iniciales arbitrarias, siendo necesario encontrar la función de partición mediante la diagonalización de la matriz densidad, [Aliaga, 1989] [Domany, Hemmen, Schulten, 1991.] [Erickson, y Ray Smith, 1989.] para poder construir un conjunto coherente de condiciones iniciales. Esta diagonalización permite además describir la dinámica del sistema en el espacio de los multiplicadores de Lagrange asociados a cada operador, espacio denominado "dual". Este espacio puede ser pensado como un espacio de las fases para sistemas cuánticos, conservándose en el (aun cuando los multiplicadores son numero s reales) las relaciones de conmutación [Aliaga ,1991 b] Todo el formalismo derivado para sistemas cuánticos se aplica en forma directa a sistemas clásicos reemplazando los conmutadores por corchetes de Poisson.

Como se ve en secciones anteriores, la Teoría de la Información desarrollada por Shannon para ser aplicada al campo de las comunicaciones, parte de la existencia de un conjunto de eventos numerables y de un espacio de probabilidad, en el que cada evento tiene una probabilidad definida $p= \{p_1 , p_2,\ldots, p_n\}$ que está normalizada

$$\sum_{i=1}^{n} p_i = 1. \qquad \textbf{6.5.1.}$$

Es posible entonces, definir la información (I) asociada con esta distribución de probabilidad, o la ignorancia relacionada con esta última, antes de conocerla *(S)*

$$I \equiv S = -\sum_{i=1}^{n} p_i \ln p_i. \qquad \textbf{6.5.2.}$$

y se denomina Entropía de Shannon. Si se considera el caso de un sistema físico, es posible expresar la entropía del sistema como

$$\boldsymbol{S = -k_B\, Tr(\rho \ln \rho) = -k_B \ln \rho}\ , \qquad \textbf{6.5.3.}$$

donde $\boldsymbol{k_B}$ es una constante que se agrega a la definición expresada por la ecuación ***6.5.3.*** a los efectos de darle unidades físicas a la entropía. Von Neumann fue el primero en asociar *S* con la entropía del estado descrito por el operador $\boldsymbol{\rho}$ al tomar $\boldsymbol{k_B}$ igual a la Constante de Boltzmann ($\boldsymbol{k_B =}$ 1.38 x 10^{-16} erg/ºK).

Como se indicó en el párrafo anterior, el Operador de Densidad $\boldsymbol{\rho}$ determina el estado físico y la entropía del mismo. Dicho estado queda parcialmente caracterizado por medio del conocimiento de un cierto número de observables relevantes para el problema físico de interés. Sólo en el caso de que los operadores formen un Conjunto Completo de Observables que Conmutan, la determinación del estado es unívoca y la entropía vale cero. En el caso de información parcial, el conocimiento de los valores medios de un número limitado de operadores implicará la existencia de distintos Operadores de Densidad que satisfagan las condiciones impuestas por las ecuaciones ***6.5.1. y 6.5.2.*** Surge, por consiguiente, el problema de la elección de uno de estos Operadores de Densidad como representación del estado físico. Es en este punto que E. T. Jaynes introduce en la teoría, el Principio de Máxima Entropía: dado un conjunto de observables $\{\hat{O}_1, \hat{O}_2, \ldots, \hat{O}_n\}$ cuyos valores medios,

$$\boldsymbol{(\hat{O}_i) = Tr(\rho\, \hat{O}_i),\ i=1,\ldots,n}\ , \qquad \textbf{6.5.4.}$$

son la única información que se tiene del sistema físico y que se denominarán Operadores Relevantes, el Operador de Densidad del sistema es aquel que maximiza la entropía, definida a través de la ecuación ***6.5.3.*** El Operador de Densidad que satisface esta condición se obtiene por el Método de los Multiplicadores de Lagrange.

$$\rho = \exp\left(- \sum_{i=0}^{n} \lambda_i \, \hat{\mathrm{O}}\mathrm{i} \right) \qquad \textbf{6.5.5.}$$

donde $\hat{O}_0$ es el Operador de Identidad que se agrega al conjunto inicial, a los efectos de satisfacer la condición

Tr ρ = 1. **6.5.6.**

Utilizando las ecuaciones ***6.5.5.*** y ***6.5.6.*** es posible entonces, relacionar la entropía del sistema con los valores medios de los operadores

$$S = k_B \sum_{i=0}^{n} \lambda_i \, (\hat{\mathrm{O}}\mathrm{i}) \qquad \textbf{6.5.7.}$$

De aquí en más se considerará k_B = 1. Los valores medios y los Multiplicadores de Lagrange están relacionados por la Ecuación

$$\lambda_0 = \ln Tr \left[\exp\left(- \sum_{i=0}^{n} \lambda_i \, \hat{\mathrm{O}}\mathrm{i} \right) \right] \qquad \textbf{6.5.8.}$$

Obteniéndose

$$\textbf{(}\hat{\mathbf{O}}_i\textbf{)=}\frac{\partial \lambda_0}{\partial \lambda_i} \qquad \textbf{, i=1,…,n ,} \qquad \textbf{6.5.9.}$$

de Lagrange.

Los resultados expuestos precedentemente fueron presentados por Jaynes para ser aplicados a un conjunto de variables del sistema, cuyos valores medios son de interés. Estos valores medios eran promedios de observables clásicos relacionados con el sistema. En la sección anterior se los ha denominado "operadores" porque estos resultados se pueden extender sin dificultad a operadores cuánticos. El conjunto de operadores utilizados por Jaynes se forma con las variables que, a priori, parecen relevantes. Si a posteriori del estudio del sistema se observa que es necesario incorporar algún operador a este conjunto para permitir una descripción más acertada, se redefine el conjunto inicial. Este método hace imposible la deducción de resultados, ya que no permite distinguir cuando un resultado no esperado es producto de la falta de algún operador o constituye un resultado nuevo del modelo en estudio. Estas limitaciones de la teoría fueron superadas por Y. Alhassid y R. D. Levine, ya que la extendieron a conjuntos de operadores cuánticos que pueden o no conmutar entre sí y

además elaboraron un método constructivo que permite, no solo determinar cuál es el conjunto de interés asociado con un sistema físico dado, sino también dotar a la dinámica, de una estructura de grupo de Lie.

Para introducir estos nuevos conceptos es conveniente trabajar con el logaritmo de la Matriz de Densidad,

$$\ln \rho = -\sum_{i=0}^{n} \lambda_i \, \hat{\mathrm{O}}\mathrm{i} \qquad \textbf{6.5.10.}$$

que también cumple con una ecuación del tipo,

$$i\frac{d \ln \rho}{dt} = \left[\mathrm{H}(t), \ln \rho\right] \qquad \textbf{6.5.11.}$$

Reemplazando la ecuación ***6.5.10.*** en la ecuación ***6.5.11.*** se comprueba que ésta será válida para todo tiempo, si el conmutador de los observables $\{\hat{O}_1, \hat{O}_2,\ldots, \hat{O}_n\}$ con el Hamiltoniano *H(t) satisface*

$$\left[H(t), \hat{\mathrm{O}}\mathrm{i}\right] = i\sum_{i=0}^{n} \hat{\mathrm{O}}\mathrm{i} \, g_{li} \qquad \textbf{i=1,…,n} \qquad \textbf{6.5.12.}$$

donde g_{li} son números complejos que se interpretan como las constantes de estructura de una semi-álgebra de Lie. Si el conjunto inicial no cumple con la condición ***6.5.12.***, se incorporarán a él todos los operadores necesarios para satisfacerla. Los *(n* + 1) x *(n* + 1) elementos g_{li} conforman la matriz $\underline{G}$, y establecen la dinámica del sistema físico, ya que como se verá, determinan las ecuaciones de evolución de los Multiplicadores de Lagrange y de los valores medios de los Operadores Relevantes. El agregar la condición de cierre de la semi-álgebra a la maximización de la entropía tiene un efecto importante ya que permite obtener, para un Hamiltoniano de un sistema físico de interés, un conjunto completo de Operadores Relevantes mediante la aplicación de un procedimiento canónico. Las ecuaciones ***6.5.11.*** y ***6.5.12.*** forman un conjunto acoplado de ecuaciones diferenciales para los Multiplicadores de Lagrange,

$$\frac{d \lambda_i}{dt} = \sum_{j=0}^{n} g_{ij} \, \lambda_j \;, \qquad \textbf{i = 1,2,…,q ,} \qquad \textbf{6.5.13.}$$

a las que se le agregan las condiciones iniciales $\lambda_j(t_0)$, compatibles con las

ecuaciones ***6.5.4.*** y ***6.5.9.*** Para el caso de Hamiltonianos independientes del tiempo, los coeficientes g_{ij} también son independientes del tiempo y la ecuación ***6.5.13.*** se transforma en un sistema de ecuaciones diferenciales a coeficientes constantes. En este caso las soluciones son del tipo:

$$\lambda_j(t) = \sum_{i=1}^{K} \exp(r_i t) \sum_{m=0}^{\gamma} a_{im}^{j}\, t^m \quad , \qquad \textbf{6.5.14.}$$

donde K es el número de raíces (r_i) diferentes de la ecuación secular correspondiente, a_{im}^{j} son constantes a determinar a partir de las condiciones iniciales y $\gamma + 1$ es la multiplicidad de r_i. Esta misma discusión puede aplicarse a los valores medios de los operadores utilizando el Teorema de Ehrenfest (ecuación ***6.5.5.)***. Si el Hamiltoniano es independiente del tiempo, al utilizar la ecuación ***6.5.12.*** se obtiene

$$\frac{d < \hat{O}_i >}{dt} = -Tr\left(\rho \sum_{j=0}^{N} \hat{O}_j\, g_{ji} \right) = -\sum_{j=0}^{N} < \hat{O}_j > g_{ji} \qquad \textbf{6.5.15.}$$

es decir, el Teorema de Ehrenfest en función de las constantes de estructura del álgebra g_{li}.

6.5.1. Definición de falta de información

Dada una situación en la cual la única información de que se dispone es la existencia de N posibilidades mutuamente excluyentes tal que, al menos una de ellas es cierta, se está en presencia de cierta falta de información ***I***(N). Se desarrolló una forma de medir la falta de información de una situación de N posibilidades, cada una de ellas con probabilidad preasignada p_i, no necesariamente iguales es [Shannon, 1948] [Abramson, 1966.]:

$$\boldsymbol{I} = \sum_{\boldsymbol{I}=1}^{N} \boldsymbol{p}_i \ln \boldsymbol{p}_i \quad , \qquad \textbf{6.5.1.1.}$$

6.5.2. Aplicación del PME al cálculo de distribuciones

Como la ecuación (**6.3.1.1.**) es la expresión para la entropía dada por la Mecánica Estadística, [Shannon, 1948] [Abramson, 1966.], se la denominó entropía S de la distribución de probabilidad:

$$S = -K\sum_{i=1}^{N} p_i \ln p_i \ , \qquad \textbf{6.5.2.1.}$$

donde K es una constante positiva que depende del problema.
por lo tanto, en lo sucesivo, se considerará a las expresiones "entropía" y "falta de información" como sinónimos y cualquier distribución de probabilidad tendrá una entropía asociada en el sentido de la Teoría de la Información.
El problema consiste en hacer uso del postulado, para encontrar la función distribución correspondiente a estados termodinámicos específicos: estados de equilibrio, dado que, la entropía termodinámica se define sólo para estados de equilibrio. En un sistema aislado, la entropía crece hasta que, en el equilibrio, resulta ser un máximo. Es por esto que la función distribución que buscamos es la que maximice la entropía **6.5.2.1.** sujeta a las condiciones de vínculo que presente el sistema (será hallada sobre la base de un conocimiento parcial que se tenga del sistema).
Generalmente, la falta de información que se tiene sobre el problema a estudiar no es total, sino que se cuenta con cierta información a priori dada por M valores de expectación F_α de ciertas magnitudes a las que llamaremos vínculos:

$$F_\alpha = \sum_{i=1}^{N} p_i f_\alpha(x_i) \qquad \textbf{6.5.2.2.}$$

con α =1,...,M. Una conjetura básica será suponer que la función distribución está normalizada, es decir, que se cumple:

$$\sum_{i=1}^{N} p_i = 1 . \qquad \textbf{6.5.2.3.}$$

la ecuación **6.5.2.3.** se denomina condición de normalización.
La función distribución se obtiene introduciendo los multiplicadores de Lagrange (λ=1,...,M) en la forma usual, es decir, hallando el máximo de la función Lagrangiano L:

$$L = -\sum_{i=1}^{N} p_i \ln p_i - \lambda_0 \sum_{i=1}^{N} p_i - \sum_{\alpha=1}^{M} \lambda_\alpha \sum_{i=1}^{N} p_i f_\alpha(x_i) \ , \qquad \textbf{6.5.2.4.}$$

siendo λ_0 el multiplicador de Lagrange asociado a la condición de normalización **6.5.2.2.**

La distribución de probabilidad que maximiza la entropía y satisface condiciones de normalización **6.5.2.2.** y vínculo resulta ser:

$$p_i = e^{-\lambda_0} \exp\left(- \sum_{\alpha} \lambda_\alpha f_\alpha(x_i) \right), \qquad \textbf{6.5.2.5.}$$

con

$$e^{\lambda_0} = \sum_i \left[\exp\left(- \sum_{\alpha} \lambda_\alpha f_\alpha(x_i) \right) \right], \qquad \textbf{6.5.2.6.}$$

que es la llamada función de partición generalizada y, donde los $\{\lambda_\alpha\}$ están determinados por el conjunto de ecuaciones acopladas

$$F_\alpha = \frac{\partial \lambda_0}{\partial \lambda_\alpha} \ . \qquad \textbf{6.5.2.7.}$$

Esta deducción es válida cualquiera sea la naturaleza de los valores de expectación. En síntesis: ***el PME provee de una metodología canónica que permite inferir distribuciones de probabilidad a posteriori cuando la información a priori son valores de expectación.***

6.5.3. Base del algoritmo

En general, las magnitudes de interés son una distribución que cumple con la ortodoxia matemática. Todas estas magnitudes representan una "escena" (lo que se mide) de la realidad. La **escena** es una distribución continua **f de un argumento x** (espacial, temporal, multidimensional) y, lo que se busca, es el modelo. Para proceder, el PME necesita de un modelo a priori (la espacial de los f). **El PME nos puede dar dos cosas: la mejor escena compatible con la información disponible y el mejor conjunto de parámetros del modelo conceptual que corresponde a esa escena.**
Mediante una formulación axiomática establece, partiendo de una S(f) desconocida (f(x) escena), que la funcional a maximizar para obtener el mejor modelo compatible con los datos es:

$$(f, m) = \int \left(\frac{f(x) - m(x) - f(x) \log\left[f(x) \right]}{m(x)} \right), \qquad \textbf{6.5.3.1.}$$

$$S(f, m) = \sum_i \left[f_i - m_i - f_i \log(f_i / m_i) \right], \qquad \textbf{6.5.3.2.}$$

donde m(x) es el modelo de la imagen. La ecuación variacional es

$$\delta(S - set\{vinculos\}) = 0 . \qquad \textbf{6.5.3.3.}$$

La selección de los multiplicadores de Lagrange optimiza el modelo.

6.5.4. Aplicación del PME a la dinámica de la evolución colectiva de sociedades

Como ya se mencionó anteriormente, la Teoría de la Información, provee un criterio constructivo para hallar distribuciones de probabilidad sobre la base de un conocimiento parcial y conduce a un tipo de inferencia que se denomina estimación de máxima entropía; a esta estimación se la considera como la conjetura más adecuada compatible con la información dada, es decir, la de concomitancia máxima respecto de la información perdida.
Si se considera a la Mecánica Estadística como una forma de Inferencia Estadística, más que como una teoría física se encuentra que sus reglas usuales, comenzando con la determinación de la función de partición, son una consecuencia inmediata del PME. Estas reglas están, entonces, justificadas independientemente de cualquier argumento físico y, en particular, independientemente de la verificación experimental; es decir, aunque los resultados inferidos concuerden o no con los medidos experimentalmente, aún seguirán representando la mejor estimación que se puede hacer sobre la base de la información disponible; la no concordancia con los datos experimentales implica, entonces, que la información de que se dispone no es la adecuada, o sea, no representa los vínculos a los cuales los datos experimentales están sujetos. Es decir, utilizaremos a la entropía en un proceso de inferencia, y según la propuesta de Shannon cuando demuestra que la expresión para la entropía, dada por la ecuación **6.5.1.1.**, representa una medida de la cantidad de incerteza representada por la distribución de probabilidad.
Nuestro análisis se basa en asumir que existe una función U(A), que representa la relación entre el crecimiento del PBI y el ingreso per capita. Siguiendo lo demostrado empíricamente supondremos que la U(A) tiene dependencia lineal, con las que se evaluara las distribuciones de probabilidad de las respectivas poblaciones. Es decir U(A) representa el vinculo de la distribución de población resultante.
De acuerdo con la metodología general, se introduce, como es usual, el Lagrangiano:

$$L = -K\sum_{i=1}^{N} p_i \ln p_i - \lambda_0 \sum_{i=1}^{N} p_i - \lambda A\,, \qquad \textbf{6.5.4.1.}$$

donde λ_o y λ son los multiplicadores de Lagrange asociados a la condición de vínculo dada por la ecuación **6.5.1.3.** y al nivel de ingresos A respectivamente. La ecuación **6.5.4.1.** es maximizada cuando se cumple

$$p(A) = p_0[U(A)]\exp(-\lambda_0 - \lambda A)\,, \qquad \textbf{6.5.4.2.}$$

donde $p_o[U(A)]$ es la distribución "a priori" del sistema. Se denomina a

$$\ln p = -\sum_{\alpha=0}^{N} \lambda_\alpha F_\alpha \qquad \textbf{6.5.4.3.}$$

como algo “sorprendente”; cuando se dispone de una cierta información "a priori" contenida en una p_0 , la diferencia entre la distribución real *p* del sistema o "a posteriori" y la distribución p_0 será medida por los vínculos faltantes en p_0 , por lo que se cumplirá

$$\ln\left(\frac{p}{p_0}\right) = -\sum_{\alpha=0}^{N} \lambda_\alpha F_\alpha \qquad \textbf{6.5.4.4.}$$

o bien

$$p = p_0 \exp\left(-\sum_{\alpha=0}^{N} \lambda_\alpha F_\alpha\right). \qquad \textbf{6.5.4.5.}$$

6.6. Distancia de Hamming.

6.6.1. Definiciones de memoria asociativa

Veamos un par de definiciones y conceptos básicos relacionados con las memorias asociativas. El primero es el tratamiento de la distancia de **Hamming,** no porque sea un concepto nuevo, sino porque deseo relacionarlo con la distancia euclídea, más conocida, con el objeto de que resulte más *plausible* la noción de distancia de Hamming. El segundo es tratar una memoria asociativa sencilla, denominada **asociador lineal** [Abramson, 1966] [Perichinsky, 1989a] [Freeman y Skapura, 1991].

6.6.1.1. Distancia de Hamming

En un espacio un conjunto de puntos forman un **cubo de Hamming** tridimensional. En general, el ***espacio de Hamming*** se puede definir mediante la expresión

$$\mathbf{H^n = \{x = (x_1, x_2, \ldots, x_n)^i \in R^n : x_i \in (\pm 1)\}} \qquad \mathbf{6.6.1.1.1.}$$

Dicho con palabras, el espacio de Hamming n-dimensional, es un hiperespacio, es el conjunto de vectores n-dimensionales cuyos componentes son elementos del conjunto de números reales **R,** estando sujetos los componentes a la restricción consistente en que su valor sólo puede ser ± 1. Este espacio contiene 2^n puntos, todos los cuales son equidistantes del origen del espacio euclídeo.

Hay muchos modelos de redes neuronales que emplean el concepto de distancia entre dos vectores. Sin embargo, hay muchas formas de medir la distancia. Se definirá la medida de distancia conocida con el nombre de ***distancia de Hamming****,* y se mostrará su relación con la ya familiar distancia euclídea entre puntos. En capítulos posteriores se explorarán otras medidas de distancia.

Sean $\mathbf{x = (x_1, x_2, \ldots, x_n)^i}$ **e** $\mathbf{y = (y_1, y_2, \ldots, y_n)^i}$ dos vectores del espacio euclídeo n-dimensional, sujetos a la restricción consistente en que x_i, $y_i \in (\pm 1)$, de tal forma que x e y son también vectores del espacio de Hamming n-dimensional. Entonces la distancia euclídea entre los extremos de los dos vectores es:

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}$$

Dado que $x_i, y_i \in (\pm 1)$, será $(x_i - y_i)^2 \in (0,4)$:

$$(x_i - y_i)^2 \equiv \begin{cases} 0 & x_i = y_i \\ 4 & x_i \neq y_i \end{cases}$$

Por tanto, la distancia euclídea se puede escribir en la forma:

$d = \sqrt{}$ 4(número de componentes distintos de x e y)

y la distancia de Hamming como:

h = número de componentes distintos de x e y **6.6.1.1.2.**

De esta manera están relacionadas por: **d = 2 √ h 6.6.1.1.3. y h = d2 / 4 6.6.1.1.4.** para componentes 0 y 1 serán bits.

Utililizaremos el concepto de distancia de Hamming algo más adelante, durante el tratamiento de la BAM. En la sección siguiente examinaremos la definición formal de la memoria asociativa y los detalles del modelo del asociador lineal.

6.6.1.2. El asociador lineal

Supongamos que se tienen L pares de vectores si se tiene $x_i \in R^n$ **e** $y_j \in R^m$ entonces **{(x_1, y_1), (x_2, y_2), (x_3, y_3), …,(x_L, y_L)}.** Estos vectores se denominan ejemplares, porque los utilizaremos como ejemplos de asociaciones correctas. Se pueden distinguir tres tipos tipos de memorias asociativas:

1. **Memoria heteroasociativa:** Establece una correspondencia **Φ** de **x** de tal manera que **Φ (x_i) = y_i** y, si un x arbitrario está más próximo a x_i que a cualquier otro x_j, con $j = 1, \ldots, L$, entonces **≥ (X) = Y_j**. En esta definición, así como en las siguientes, más próximo quiere decir con respecto a la distancia de Hamming.
2. **Memoria asociativa interpoladora:** Establece una correspondencia **Φ** de **x** e **y** de tal manera que **Φ (x_i) = y_i**, pero, si el vector de entrada difiere de uno de los ejemplares en el vector **d**, de tal modo que **x = x_j + d**, entonces la salida de la memoria también difiere de uno de los ejemplares en algún vector **e**: **Φ**

$(\mathbf{x}) = \Phi(\mathbf{x_j} + \mathbf{d}) = \mathbf{y_j} + \mathbf{e}$.

3. **Memoria autoasociativa:** Presupone que $\mathbf{y_i} = \mathbf{x_i}$ y establece una correspondencia Φ de $\mathbf{x}$ en $\mathbf{x}$ tal que $\Phi(\mathbf{x_i}) = \mathbf{x_i}$, y, si algún $\mathbf{x}$ arbitrario está más próximo a $\mathbf{x_i}$ que cualquier otro $\mathbf{x_j}$, con $\mathbf{j = 1, \ldots, L}$, es $\Phi(\mathbf{x}) = \mathbf{x_j}$.

No es difícil construir matemáticamente una memoria como ésta si se aplica la restricción adicional consistente en que los vectores $\mathbf{x_i}$ constituyan un conjunto ortonormal. Para construir una memoria asociativa interpoladora, se define la función:

$$\Phi(\mathbf{x}) = \Phi(\mathbf{y_1 x'_1}) + \Phi(\mathbf{y_2 x'_2}) + \Phi(\mathbf{y_3 x'_3}) + \ldots + \Phi(\mathbf{y_L x'_L})\,\mathbf{x} \qquad \textbf{6.6.1.2.1.}$$

Si x_i es el vector de entrada, entonces $\Phi(\mathbf{x_j}) = \mathbf{y_j}$, puesto que el conjunto de vectores $\mathbf{x}$ es ortonormal, donde $\mathbf{x}$'i $\mathbf{x_j} = \delta_{ij}$ con $\delta_{ij} = \mathbf{0}$ si $\mathbf{i \neq j}$ y $\delta_{ij} = \mathbf{1}$ si $\mathbf{i = j}$.

Este modelo es utilizado en procesamiento distribuido, memoria asociativa, combinar multiprocesamiento simétrico con procesamiento paralelo masivo, arquitecturas cc-NUMA (SPIDER) y bases de datos relacionales y OO con planos virtuales en analogía con las memorias asociativas y recordando que los vértices del hipercubo del hiperespacio de Hamming es el punto de convergencia por estabilidad, incluyendo el clustering de taxonomía.

APLICACIÓN

La norma es tener fe en los hechos que siempre superan las expectativas

Rodolfo Walsh

CCII

7. APLICACIÓN

7.1. CUERPOS CELESTES. FAMILIAS DE ASTEROIDES

7.1.1. INGENIERÍA DE REQUERIMIENTOS

Cuando desarrollamos software debemos garantizar la producción sistemática y controlada de los productos que satisfagan las necesidades de los usuarios, a tiempo y costos establecidos. Se deben aplicar principios, procedimientos, métricas y herramientas similares a las que se emplean en otras ramas de la ingeniería ya que se necesitan emplear estándares [Yourdon, 1993] [Sommerville, 1997] [Kotonya y Sommerville, 1998] [Robertson y Robertson, 1999] [IEEE Std. 1471, 2000] [Pressman, 2002] [Wiegers, 2003] [Brito y Moreira, 2004].

7.1.1.1. Por lo tanto aplicamos: Ingeniería de Software.

Definiciones del IEEE:

"El uso de métodos sistemáticos, disciplinados y cuantificables para el desarrollo, operación y mantenimiento del software."

"El estudio de técnicas relacionadas con el uso de métodos sistemáticos, disciplinados y cuantificables para el desarrollo, operación y mantenimiento del software."

Definición de Richard Fairley:

"La ingeniería de software es la disciplina tecnológica y de administración que se ocupa de la producción y evolución sistemática de productos de software que son desarrollados y modificados dentro de tiempos y costos estimados."

7.1.1.2. Ciclo de Vida de Software

[Brooks, 1987]

1. Un ciclo de vida está conformado por distintas etapas.

2. Cada ciclo de vida determina la forma en que se van desarrollando las etapas.
3. Se utilizará un ciclo de vida u otro dependiendo de ciertas características del producto a desarrollar, de los costos y los requerimientos del usuario
4. Organizar las actividades del administrador aumentando la probabilidad de que se aborden los problemas pertinentes en el momento adecuado.

7.1.1.3. Actividades genéricas:

1. Definición (qué debe hacerse)
2. Análisis del sistema y Planificación del proyecto
3. Análisis de Requerimientos
4. Desarrollo (cómo debe hacerse)
5. Diseño de software =>Codificación
6. Prueba de software y Mantenimiento:
7. Correctivo =>Adaptativo =>Perfectivo =>Preventivo

Un requerimiento es una característica del sistema o una descripción de algo que el sistema es capaz de hacer con el objeto de satisfacer el propósito del sistema [Jacobson, 1992].

El análisis de requerimientos es un proceso de descubrimiento, refinamiento, modelización y especificación.

Comienza con un refinamiento detallado del ámbito del programa establecido durante la ingeniería del sistema y refinado durante la planificación del proyecto de software.

Se crean los modelos del flujo de la información y del control, del comportamiento en operación y del contenido de los datos. Se analizan las soluciones alternativas.

El análisis y la especificación de requerimientos puede parecer una tarea sencilla, pero el contenido de la comunicación es muy denso y se puede llegar a malas interpretaciones, falta de información o ambigüedades.

El análisis de requerimientos es la tarea que establece un puente entre la asignación del software a nivel de sistema y el diseño del software.

El análisis de requerimientos proporciona al diseñador una representación de la información y de las funciones que se puede traducir en un diseño de datos, arquitectónico y procedimental.

Los Tipos de Requerimientos pueden ser funcionales, si describen una interacción entre el sistema y su ambiente, como debe comportarse el sistema ante determinado estímulo, o Requerimientos no funcionales, si describen una restricción sobre el sistema que limita nuestras elecciones en la construcción de una solución al problema.

A las **Actividades de Ingeniería de Requerimientos** usualmente podemos dividir las prácticas en cuatro actividades, a saber:

1. **Extracción**
2. **Análisis**
3. **Especificación**
4. **Validación.**

Como toda división de tareas, no es una estricta representación de la realidad, sino que se hace con el fin de sistematizar la realización de la **Ingeniería de Requerimientos**.

En general la delimitación entre una actividad y la otra no es tan clara, ya que están sumamente interrelacionadas, existiendo un alto grado de iteración y retroalimentación entre una y otra.

Extracción es el nombre comúnmente dado a las actividades involucradas en el descubrimiento de los requerimientos del sistema.

Es decir, obtener un conocimiento del área general de aplicación del sistema; comprender el problema en sí, lo que implica que se debe extender y especializar el conocimiento sobre el dominio general para que se aplique en particular.

Es importante, entonces, que la extracción sea efectiva, ya que la aceptación del sistema dependerá de cuanto ésta satisfaga las necesidades y de cuan bien asista a la automatización del trabajo.

Sobre la base de la **extracción** realizada previamente, usualmente se hace un **análisis** luego de haber producido un bosquejo inicial del documento de requerimientos. Esta es una fase sumamente compleja en un proyecto pues el dominio es desconocido, y se apunta a descubrir problemas con los requerimientos del sistema identificados hasta el momento.

Hay distintas técnicas y herramientas que se utilizan para llevar a cabo cada una de las actividades del proceso.

- Preparar hipótesis con un conjunto de preguntas que pueda influir en las respuestas, de dos categorías: abiertas y cerradas.

En las preguntas abiertas responden con su propia terminología y son reveladoras, ya que no están limitadas las respuestas; y si las preguntas son cerradas predeterminan las posibles respuestas.

- Visualizar Sistemas existentes consiste en analizar distintos sistemas ya desarrollados que estén relacionados con el sistema a ser construido, analizar las interfases, observando el tipo de información que se maneja y cómo es manejada.

Una ventaja que presenta esta técnica es que como estos sistemas ya están en producción, ya han pasado por la curva de aprendizaje del dominio del problema.

- Producir un Brainstorming (tormenta de ideas), modelo para generar ideas.

La intención en su aplicación es la de generar la máxima cantidad posible de requerimientos para el sistema, sin detenerse en pensar si la idea es o no del todo utilizable. Luego, se irán eliminando en base a distintos criterios. La regla básica es que los participantes deben pertenecer a distintas disciplinas [Arango, 2002].

- Hacer una Arqueología de documentos para determinar posibles requerimientos sobre la base de la investigación documental como toda etapa científica.

Del análisis documental se debe buscar el propósito del documento, quién lo usa, por qué, para qué, las tareas que realizan, la relación entre documentos y el proceso que realiza la conexión.

- Lo conceptual de la idea del maestro y el aprendiz, es una herramienta que permite una buena forma de observar el trabajo real.

El Analista representa al aprendiz y el usuario / especialista cumple el rol de maestro. El aprendiz se sienta con el maestro a aprender por medio de la observación.

- Observar como se hacen las cosas es una buena manera de entender lo que estas requieren. Conectarse íntimamente con la cultura de la organización, vivirla, es una herramienta que debe ser tomada en cuenta [Kotler, 1993].

- Realiza un Run Use Case WorkShop (Talleres de Trabajo basados en los Casos de Uso), entre especialistas y analistas, consiste en generar los escenarios, usando la información que tiene para brindar el especialista.

La idea es especificar por medio de los casos de uso y extraer las cosas esenciales que suceden cuando ocurre un evento determinado. Como resultado de este proceso se obtiene un excelente bosquejo del caso de uso [Jacobson, Booch y Rumbaugh, 2000] [Filman, Elrad, Clarke, y Aksit, 2005].

- Durante la actividad de extracción de requerimientos, puede ocurrir que algunos requerimientos no estén demasiado claros, para validarlos se construyen prototipos.

Los prototipos son simulaciones del posible producto, que luego de utilizados permiten conseguir una importante retroalimentación.

Los prototipos se pueden clasificar en Prototipo evolutivo, Prototipo bosquejado y Prototipo tangible.

- Hay que realizar un análisis para intentar identificar las principales fortalezas, oportunidades, debilidades y amenazas con las que se enfrenta el Problema: FODA [ISO/IEC: FCD 9126-1, 2001].

Las oportunidades y las amenazas se refieren a los factores externos futuros.

Las fuerzas y debilidades que son factores internos; señalan estrategias cuya aplicación podría conducir al éxito, mientras que las debilidades es lo que se debe corregir.

Esta herramienta es útil para analizar la situación del Problema y ver de qué forma se puede ayudar a disminuir las debilidades y amenazas: en positivo y en negativo.

- Todos los Problemas son una colección de actividades que se llevan a cabo para diseñar, producir, distribuir, entregar y apoyar a su producto. Esta colección es la Cadena de Valor de actividades estratégicas para comprender el comportamiento y las fuentes de diferenciación presentes y futuras. Con esta herramienta se puede analizar el flujo de información que interviene en las distintas actividades.

- El Modelo de clase conceptual, el Diagrama Conceptual y el Diagrama de Clases Conceptual, son modelos que representan conceptos del dominio del problema y permite mostrar conceptos, asociaciones entre conceptos y atributos de conceptos, ayuda a comprender la terminología del dominio y

comunica cuáles son los términos importantes y las relaciones existentes entre ellos.

Siendo un Concepto una categoría de ideas o cosas, para la descripción de sus atributos, operaciones y significado y una clase representa un concepto del dominio del problema, como caso particular el Diagrama de pescado (Ishikawa Diagram, Cause-and-Effect o Fishbone Diagram) [ISHIKAWA, 1969] [David, 1998].

Otras **Herramientas** son: (1) el **Glosario** que es una simple lista de términos en donde se explica su significado; (2) el **Documento de Concepto de Operaciones (DCO)** para comprender el entorno en el cual se encuentra el Problema, describiendo su funcionamiento interno y su relación con el ambiente; (3) el **Diagrama de Actividad** o diagrama de proceso, se asemeja a un mapa de procedimientos, mostrando el flujo de actividades; (4) el **Documento de Especificación de Requerimientos (ESRE)** para especificar los requerimientos del sistema **(Casos de Uso)**, que se pueden clasificar en categorías de no-funcionales y funcionales; (5) el **Caso de Uso** es un documento narrativo que describe la secuencia de eventos de un actor (agente externo) que utiliza un sistema para completar un proceso. Es una técnica diseñada para especificar el comportamiento de un sistema; (6) la **Casa de Calidad** es un esquema **QFD (Quality Function Deployment)** es una matriz que representa las casas de calidad, en las cuales las filas representan la lista de requerimientos, mientras que las columnas representan cómo se llevan a cabo los requerimientos **(casos de uso)** utilizando referencias y finalmente (7) la **Checklist** o lista de verificación para probar que no falta ningún caso de uso para los requerimientos, recorriendo el **ESRE** y los **Casos de Uso** [Jacobson, Booch y Rumbaugh, 2000] [ISO/IEC: FCD 9126-1, 2001] [Filman, Elrad, Clarke, y Aksit, 2005].

En **Conclusión**, finalizado el Proceso de Ingeniería de Requerimientos y las herramientas, que se pueden utilizar para realizar las actividades del proceso de ella, es útil compartir experiencias para su aplicación práctica en las diferentes etapas.

7.1.2. ESPECIFICACIÓN, REQUERIMIENTOS Y CATEGORÍAS DE CASOS DE USO: CUERPOS CELESTES

7.1.2.1. Hirayama

Examinando la distribución de los asteroides con respecto a sus elementos orbitales, en particular su movimiento principal, la inclinación y la excentricidad, se observan condensaciones en distintos lugares que parecen al azar, pero hay algunos casos en los cuales tener en cuenta solo las leyes de la probabilidad no es tan evidente [Hirayama, 1918a, 1918b, 1933].
Los asteroides están demasiado agrupados por tener inclinaciones cercanas o los planos de las orbitales tienen prácticamente el mismo polo (el de la órbita de Júpiter), otros agrupamientos no tienen el mismo centro pero el dibujo del diagrama tomando la excentricidad y la longitud del perihelio en lugar de la inclinación y la longitud del nodo se forma una distribución circunferencial.
Siguiendo el desarrollo de la teoría mencionada no existen dudas de que hay relaciones físicas que conectan a los asteroides. Por ello es que podemos aventurar que existen familias de asteroides asociados.
Si eliminamos las perturbaciones de Júpiter tenemos familias de elementos propios a las cuales denominamos KORONIS, EOS y THEMIS según los elementos (158), (221) y (24) respectivamente.

Si se supone que en un instante dado se rompió un asteroide en varios fragmentos y que la variación de velocidad es mínima, la variación máxima corresponde al semi-eje mayor y después de mucho tiempo la distribución de los fragmentos es irregular y la variación secular del punto que representa el centro de la circunferencia original es mínima.

La teoría queda verificada y así la formación de familias tales como KORONIS (158), EOS (221), THEMIS (24), FLORA (244), MARIA (170) y PHOCAEA (25).

7.1.1.2. Arnold

La distribución de elementos orbitales en cinturones de asteroides no es al azar mostrando la existencia de familias, tal que los grupos de asteroides cuyo semi-eje mayor, su excentricidad y su inclinación (o el seno de la misma) se aproximan a un cluster para ciertos valores especiales [Arnold, J.R.1969].

Se ha verificado la aglomeración en familias (clustering) corrigiendo la perturbación periódica producida por variaciones seculares debidas a los planetas mayores, tomando los elementos propios.
Se ha tenido en cuenta la correlación de elementos propios angulares como el argumento del perihelio y el argumento del nodo.

Otros agrupamientos han sido identificados por características propias de resonancia o corriente de asteroides impulsados (JET STREAMS) a través de la familia FLORA y objetos que cruzan MARTE en órbitas de excentricidad de orden superior.

Un interés especial es el lugar de origen de los meteoritos y el cálculo del ciclo de vida de los pequeños objetos del espacio que indica que el **tiempo esperado** de que un asteroide del cinturón golpee la TIERRA es del orden de 10^8 - 10^9 años y la **edad de bombardeo** es del orden de los 10^7 años o menos.

El problema es que la resonancia en las familias y jet streams no invaliden los cálculos, por ello es útil buscar evidencia de resonancias en ambos.
Considerando que cerca de 1969 había menos de 1735 objetos y con un semi eje mayor entre 2.15 - 4.00, la densidad de la distribución espacial, por antecedentes propios, es de 1735/(1.2 u.a. * 0.3 * 0.3) = 1.6. 10^4 $[u.a.]^{-1}$ donde u.a. = unidades astronómicas. Esto significa asumir que los objetos conocidos están distribuidos uniformemente entre 2.15 - 3.35 u.a. y con excentricidad entre 0.0 - 0.3.

Sobre esta base un cluster debe tener una densidad mayor, dado que esta tomada en el espacio de una región cuyo centro sea un asteroide. Es más claro y conveniente tomar una región elipsoidal que una rectangular pues esta resulta inconsistente. El paso siguiente es agrupar regiones de familias, que aunque se solapen, se logra que muchas familias sean descartadas entre sí pero identificadas en otras.

Según Arnold siguiendo la ley de Poisson el número de elementos de un conjunto debe ser menor que un cierto número esperado, con la cual no se concuerda en esta tesis pues los eventos no siguen esta ley por contradecir todo lo desarrollado hasta ahora: se basa en grandores físicos, en características fenotípicas de caracteres o atributos de los asteroides y finalmente de su genotípica u origen común.

Finalmente mediante una prueba estadística se debe encontrar el tamaño de una familia dada tal que la probabilidad de ocurrencia del mismo sea minimizada frente a la densidad de objetos esperados en la región.

Esto debe ser hecho de tal manera que el test estadístico rechace elementos errantes o vagabundos por no pertenecer a la estructura taxonómica. Esto es válido para los Jet Streams.

Toda esa conclusión parece ser arbitraria pues debe prevalecer el concepto conservativo de la masa es decir la densidad y la estabilidad del entorno.

Condiciones de vecindad cercana deben ser tenidas en cuenta y las familias de alta densidad son las más estables y menos azarosas.

Media, rangos, desviaciones estándar y otros estadísticos deben ser tenidos en cuenta para cada familia, no solo para las variables seleccionadas sino también para las distancias al afelio y al perihelio, valores propios sujetos a periódicas perturbaciones.

Se confirman las familias de Hirayama y las familias pequeñas son de baja densidad y la probabilidad de pertenezcan a las familias es alta y por lo tanto su acoplamiento por el método **pair-group**.

7.1.1.3. Carusi y Valsechi

Cerca de 1982 hay un registro de 2125 planetas menores, tipo asteroides, recolectados lo cual produjo discrepancias en los resultados de los métodos de clasificación computacionales basados en parámetros dinámicos y físicos [Carusi, A., Valsechi, G.B., 1982].

Esta discrepancia entre los métodos estadísticos es desconcertante pues la relación entre los miembros de una familia respecto a los parámetros dinámicos y cualquier estudio físico que se realice sobre los mismos deben ser concurrentes.

Investigadores han llegado a la conclusión que el problema de la clasificación de los asteroides en familias está claramente definido y prácticamente resuelto, visión simplista que en tal estado esta tesis no comparte.

Se puede observar que el crecimiento en observaciones entre 1969 [Arnold, J. R.1969] y 1982 [Carusi. A., Masaro, E., 1978] [Williams, J. G., 1979] traen discrepancias.

Las discrepancias surgen [Carusi, A., Valsechi, G.B., 1982] de los métodos de cómputo de los elementos propios, del criterio de rechazo de los objetos a ser clasificados, del tamaño de la muestra, de los métodos de identificación de familias y los criterios de rechazo de un miembro de una familia.
Los métodos de cómputo de los elementos propios tratan la eliminación de las perturbaciones seculares como ya se mencionara.
El segundo y tercer aspecto de las discrepancias obedecen a la historia compleja de la astronomía y la astrofísica, problemas de frontera en la vecindad de objetos y clusters y su estabilidad.
De los métodos de identificación de familias las discrepancias surgen por sus criterios probabilísticos y el futuro descubrimiento de nuevos asteroides que parecen contradictorios, pero a pesar de todo si son congruentes las familias sospechadas aparecen en la realidad, contrastación científica, pero si los

métodos son arbitrarios siempre son discutibles además de la duda metodológica [el autor].

7.1.1.4. Williams

El problema de Arnold (que Williams vuelve a tomar en 1982) ya fue discutido en función de su criterio de **densidad de distribución Poissoniana uniforme** y los **elementos propios**.

En los años 1980´s las técnicas de análisis por similitud y una distancia generalizada pero con el uso de **juicios personales o manejo manual** es lo usual y no una clasificación automática.

Por ello aparece la consideración de la varianza (σ_j) de los dominios y familias para el proceso de identificación de elementos dentro de la familia o la subsiguiente.

Las clases aceptadas han sido divididas en dos tipos: ***tipo 1***, si la clase ha sido identificada en dos intervalos, sin diferencias perceptibles y ***tipo 2***, si la clase fue encontrada mezclada con otras clases menos importantes en intervalos solapados, pudiendo existir familias enmascaradas o entornos poco confiables, estos aspectos deben surgir del propio método estadístico.

Los criterios de rechazo de un miembro de una familia no son claros, son arbitrarios o directamente no se exponen en los trabajos y por lógica consecuencia no son automáticos.

El número de objetos en una región no está probado ni tampoco la densidad constante, que hizo que algunos autores dividieran la familia FLORA (splitting), por más que se tomen familias pequeñas como Arnold y Williams con 20 años de diferencia entre varios de sus trabajos y de los descubrimientos de nuevos objetos.

La identificación de familias de asteroides y los elementos propios fueron aproximadamente detallados anteriormente y fueron recalculados en función de las oscilaciones de los asteroides y de los elementos propios que anteriormente tenían órbitas pobres [Williams., J.G., 1989.].

Estos trabajos realizados en el Jet Propulsion Laboratory, California Institute of Technology, dieron como resultado órbitas cruzando los planetas mayores y que se dividen en familias, por la característica del método y no por poseer un fundamento formalmente "duro" (splitting). Algunos elementos propios como el semi-eje mayor y el ángulo de inclinación o su seno no pudieron ser contrastados, la longitud propia del perihelio y el nodo en grados, fueron tomados desde el equinoccio de un elemento y se computó la fase libre de oscilaciones libres igual que la resonancia secular. Una característica es que las resonancias fuertes no aparecen en asteroides y las débiles son tomadas como ruido.

Las distancias son tomadas desde una línea recta SOL - PLANETA (Marte MXR, Júpiter JXR, Saturno SXR, etc.) y los valores propios son más exactos dentro del cinturón que fuera de él (lo cual abona la teoría del autor).

7.1.1.5. KnêzevÎc y Milani

Los elementos propios de asteroides de una teoría analítica de segundo orden [Knezevic, Z., Milani, A., 1990], de asteroides identificados en el cinturón principal (main-belt), son mucho más exactos que los de excentricidad e inclinación baja en la región de la familia Themis. Esto es porque las perturbaciones periódicas cortas son eliminadas y son tenidos en cuenta los principales efectos de segundo orden dependientes de la relación al cuadrado de [masa de Júpiter / masa del Sol], de acuerdo con los resultados del algoritmo consistente con las modernas teorías dinámicas de Kolmogorov-Arnold-Moser, son cerca de 3495 asteroides de la edición del Leningrad Ephemerides of the Minor Planets. Se descartaron Hildas, Troyanas y los cercanos a la Tierra (q<1.1 u.a.).

El algoritmo permite calcular un código de calidad (QC) que indica cuantas iteraciones hay que realizar para que converja.

La cantidad de asteroides se pueden recalcular en alrededor de 55 (¿por qué?) iteraciones, siempre que la inclinación no sea grande al igual que la excentricidad.

Todo este desarrollo aparece poco claro y arbitrario, no hay un sustento formal en la relación convergencia cantidad de iteraciones y el número de asteroides.

7.1.1.6. Zappala, Cellino, Farinella y Knêzevîc.

Este criterio es importante desde que una clasificación de los asteroides mejorada fue mencionada en familias dinámicas, al analizar una base de datos de asteroides numerados, cuyos elementos propios se usaron para computar en un nuevo segundo-orden, cuarto-grado de la teoría de perturbación secular [Knêzevîc, y Milani, 1990] [Knêzevîc, Milani, Farinella, Froechle, Ch., Froechle, C., 1991] [Zappala, Cellino, Farinella, Knêzevîc, Milani, 1990, 1994], y se verificó su estabilidad en términos largos. Este criterio multivariado, usa la técnica de análisis de datos que se agrupan jerárquicamente. Un "dendrograma", gráfico similar al "fenograma", fue aplicado para construir en cada zona del cinturón de los asteroides en el espacio de elementos propios, con una distancia en función relacionada con la velocidad incremental necesaria del cambio orbital después de la eyección del cuerpo del padre fraccionario [Hirayama, 1918a, 1918b, 1933].

Las familias se identifican entonces por comparación con el dendrograma similar, derivadas de una distribución "aleatoria" de elementos que se compara con una estructura en escala "gruesa" de la distribución real.

Los parámetros de importancia asociados con cada familia, medidos como resultados aleatorios de las concentraciones, (como transformar las zonas anisotropas e inhomogeneas en zonas homogéneas e isotropas de las zonas entre espacios (inter-gaps) del cinturón de asteroides, modificación deatributos mecánicos, como el eje semimayor y la inclinación) y parámetros fuertes (estabilidad), se obtuvieron repitiendo el procedimiento de clasificación después de variar los elementos de velocidad en cantidades pequeñas al recomputar las zonas reales, con cálculos con cambios artificiales de los coeficientes de la función distancia.

Para de tomar promedios de variación de distancias fueron armadas las denomidas estalactitas, mientras se toma el ancho y la profundidad en la función de la velocidad modificada. Siendo un criterio innovador es importante analizarlo, aunque no está claro la técnica implementada, el clustering, dentro de las zonas y la variación de los promedios de las velocidades, antes mencionados, y por otro lado son ignoradas familias de hasta cinco elementos, todas síntesis de una instrumentación arbitraria.

Las familias más importantes y saludables son como de costumbre Themis, Eos, y Koronis que juntas contienen el 14% del cinturón principal conocido, de la población; pero las12 familias más confiables y saludables que se encontraron a lo largo del cinturón, la mayoría comparten parcialmente clasificaciones anteriores.

Es el caso de FLORA en la región interior del cinturón, mientras da lugar a una identificación muy difícil de familias confiables, principalmente cuando tienen una densidad y una exactitud alta de inclinaciones y excentricidades propias es pobre principalmente a causa de la proximidad de una fuerte resonancia secular.

Han llegado a constituir 21 familias [Zappala, Cellino, Farinella, Milani, 1990, 1994] [Perichinsky, Orellana, Plastino, 2000], con un método realmente importante y métodos totalmente automatizados.

Usando una base de datos actualizada de asteroides puedo determinar la robustez del método.

Por otro lado: (i) el trabajo de [Arnold, 1969] [Carusi, Massaro, 1978] [Knêzevîc, Milani, Farinella, Froechle, Ch., Froechle, C., 1991] ha clarificado los puntos sutiles acerca de la evolución dinámica en largo plazo de las órbitas de los asteroides cuyo planear es un requisito previo esencial para los elementos apropiados que derivan (para la clasificación en las familias); (ii) la disponibilidad de datos físicos de tamaños, formas, taxonomía numérica y velocidad de rotación de muchos cientos de asteroides ha provocado nuevos análisis de familias, la búsqueda de congruencia en las técnicas que usan las propiedades, en particular, la estructura interior, del cuerpo del padre y el mecanismo de su fragmentación [Zappala, Cellino, Farinella, Milani, 1990, 1994], y (iii), cuando se cree que las familias representan los resultados del impacto de asteroides con una energía muy fuerte, su abundancia y las propiedades proporcionan un complemento observacional que permite obviar a los modelos teóricos de la historia colisional de los asteroides, de los parámetros principales (fuerza de impacto, energía de fragmentación, etc.) y la abundancia inicial de material sólido en el cinturón de asteroides [Zappala, Cellino, Farinella, Milani, 1990, 1994].

Otros problemas han provenido de los estudios físicos de las familias. Primero, mientras las familias más pobladas aparecen en ambos criterios en forma

bastante homogénea, el criterio de la composición y precedentes físicos y cosmoquícos [Zappala, Cellino, Farinella, Milani, 1990, 1994], es un criterio con más o menos dificultad y el criterio que con menos dificultad han identificado a las familias es ese que usa los datos de la Mecánica Celeste. Segundo, algún análisis de Williams [Williams, 1992a, 1992b] [Zappala, Cellino, Farinella, Milani, 1990, 1994] ha concluido que familias pequeñas son originadas por uniones de tipo taxonómico que se corresponden con la materia. De un punto de vista del cosmoquímico no es probable tener objetos juntos que pertenezcan a un solo padre. El criterio de Chapman (1989) [Zappala, Cellino, Farinella, Milani, 1990, 1994], es diferente, las familias de Williams con 12 o más miembros son taxonómicamente diferentes de los precedentes, aquéllos con menos de cinco miembros no serán definitivamente diferentes (algo que no implica que ellos necesaria y genéricamente sean "irreales").

Un problema adicional evidenciado por los estudios físicos se relaciona a las propiedades geométricas, del campo de velocidades de eyección, de los miembros de las familias del cuerpo del padre, como deducido de distribuciones de elementos propios. Según Brouwer (1951) y Zappala (1984) [Zappala, Cellino, Farinella, Milani, 1990, 1994], la mayoría de las familias derivadas por Brouwer y Williams [Knêzevîc, Milani, 1990] muestran una anisotropía sistemática de distribución tridimensional de velocidades de eyección de una órbita. La anisotropía más probablemente se relaciona con una exactitud de inclinación limitada y excentricidades propias que a los eventos de desintegración, una conclusión apoyada en experimentos numéricos de elementos propios usados por Brouwer y Carpi (1986) [Williams, 1989] [Williams, 1992a, 1992b].

7.1.1.6. Criterio de Clasificación por Análisis Espectral

Con estas motivaciones, he logrado con el criterio de análisis espectral [Feynmann, 1971] [Hetcht, 1977] [Hamming, 1980] [Perichinsky, 2000 - 2007] (ver Capítulo **4**, de solución propuesta) que las clasificaciones de bases de datos extendidas con elementos propios de asteroides (método corroborado en Cúmulos [2007]) formen familias [Hirayama, 1918a, 1918b, 1933]. Reconozco que los trabajos de Zappala son muy importantes (clasificación automática y

con el método jerárquico), es un punto de inflexión en los tempranos 90´s, pero es diferente el acercamiento, porque trabajo en Taxonomía Computacional, en un hiperspacio taxonómico, y no en un criterio de composición y precedentes físicos y cosmoquícos. Zappala se confundió con una metodología, que usa sólo una variable de velocidad, en un espacio transformado no claramente unívoco.

Incorporando así un actualizado conjunto, más grande, de elementos oscilantes que derivaron de la teoría de perturbación secular, cuya exactitud (específicamente, la estabilidad con el tiempo) se ha verificado extensivamente por integración numérica en términos de largo plazo; en forma automática, y prejuzgar la técnica de análisis de datos en grupos no azarosos, no usados en el espacio de elementos propios, como en el criterio de Zappala [Zappala, Cellino, Farinella, Milani, 1990, 1994] y cuantitativamente la importancia estadística de estos grupos; con la robustez de las estadísticas para familias importantes con respecto a las variaciones aleatorias pequeñas de elementos propios, todos basados en un análisis de Taxonomía Computacional.

No consideramos la transformación de conjuntos isotrópicos y homogéneos, mientras cambian los valores de la excentricidad y el semieje al volver a computar valores de las zonas de entre-espacios del cinturón de asteroides en velocidades en promedio, o eliminar a los grupos de 5 o menos objetos, todos de los cuales se considera que están fuera de un criterio Computacional.

Así, un nuevo acercamiento a la Taxonomía Computacional se presenta, el cual ya ha sido empleado con referencia a la Minería de Datos (Data Mining).

7.2. Implementación

7.2.1. Matriz de Datos.

Atributos.

Del total de asteroides se calcularon las correcciones de estabilidad para 1796 y considerando solo el tratamiento de elementos propios, se ha tomado una prueba sobre 379 que se estabilizaron en las familias de Hirayama.

Se muestran a continuación los números de identificación de los asteroides sobre los que se trabajaron y cuyo detalle, **MATRIZ DE DATOS** se ve en el **ANEXO I,** de OTU´s y caracteres, que son los datos de entrada del algoritmo y donde cada columna es un dominio de un atributo.

THEMIS
Cantidad de asteroides 67, elemento de cabecera THEMIS.24.

24	62	90	104	171	184	222	223	268	316
379	383	431	461	468	492	515	526	555	621
637	656	710	767	846	848	936	938	946	954
981	988	991	996	1003	1027	1061	1073	1074	1082
1142	1171	1229	1247	1253	1259	1302	1331	1340	1440
1445	1462	1487	1539	1576	1581	1615	1623	1624	1633
1669	1674	1684	1686	1687	1691	1698			

KORONIS
Cantidad de asteroides 37, elemento de cabecera KORONIS.158.

158	167	208	243	263	277	311	321	452	462
534	658	720	761	811	832	962	975	993	1029
1079	1100	1223	1245	1289	1336	1350	1363	1389	1423
1442	1482	1497	1570	1618	1635	1725			

MARÍA
Cantidad de asteroides 20, elemento de cabecera MARÍA.170.

170	472	575	616	652	660	695	714	727	751
787	875	879	897	994	1158	1160	1215	1379	1677

EOS
Cantidad de asteroides 66, elemento de cabecera EOS.221.

221	320	339	450	513	520	529	562	573	579
590	608	633	639	651	653	661	669	742	766
775	798	807	833	876	890	1033	1075	1087	1105

1112	1129	1148	1174	1186	1199	1207	1210	1220	1234
1286	1287	1291	1297	1339	1353	1364	1388	1410	1413
1416	1434	1464	1485	1532	1533	1552	1557	1604	1605
1641	1649	1654	1711	1723	1732				

PHOCAEA
Cantidad de asteroides 34, elemento de cabecera PHOCAEA.24.

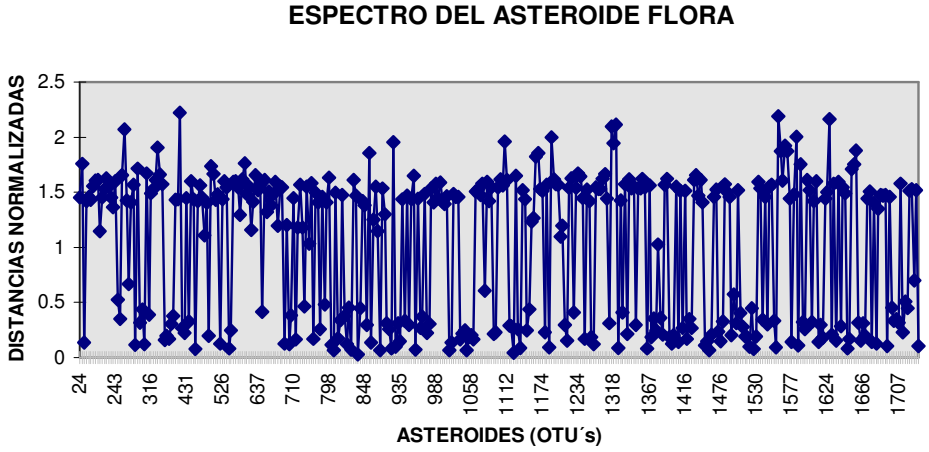
25	105	265	273	290	323	326	391	502	587
654	852	914	950	1090	1108	1164	1170	1192	1310
1316	1318	1322	1342	1367	1565	1568	1573	1575	1584
1591	1626	1657	1660						

FLORA
Cantidad de asteroides 156, elemento de cabecera FLORA.8.

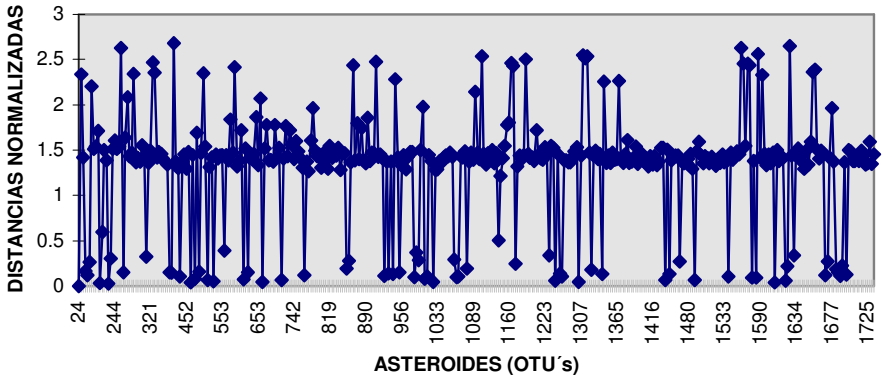
8	43	244	254	270	281	291	296	298	315
341	352	364	367	376	422	428	440	453	496
525	540	553	641	685	700	703	711	736	763
770	782	800	802	809	810	819	823	825	831
836	841	851	871	883	901	905	913	915	929
935	937	939	951	956	960	963	967	1016	1026
1037	1047	1052	1055	1056	1058	1060	1078	1088	1089
1117	1120	1123	1130	1133	1150	1153	1185	1188	1214
1216	1219	1225	1249	1270	1274	1307	1324	1335	1338
1344	1365	1370	1376	1377	1382	1387	1396	1399	1405
1412	1415	1418	1419	1422	1446	1449	1451	1455	1472
1476	1480	1492	1494	1496	1500	1513	1514	1518	1523
1526	1527	1530	1536	1549	1562	1563	1577	1590	1601

1602	1608	1619	1621	1622	1631	1634	1636	1651	1652
1661	1663	1666	1667	1675	1682	1696	1699	1703	1704
1707	1713	1717	1720	1729	1733				

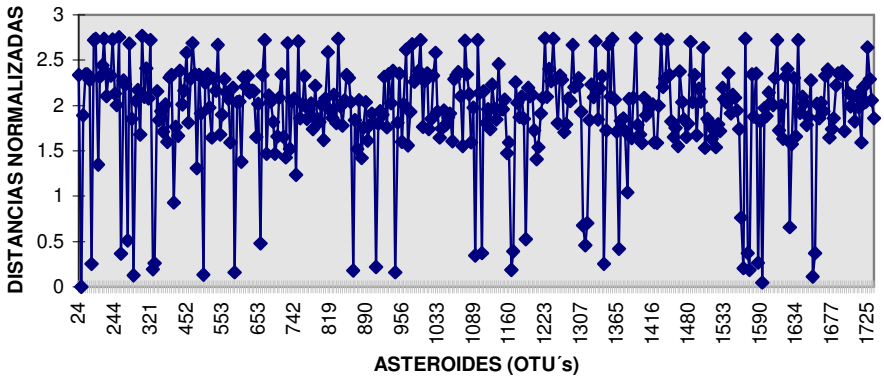
Espectros característicos de los Asteroides cabeza de familia.



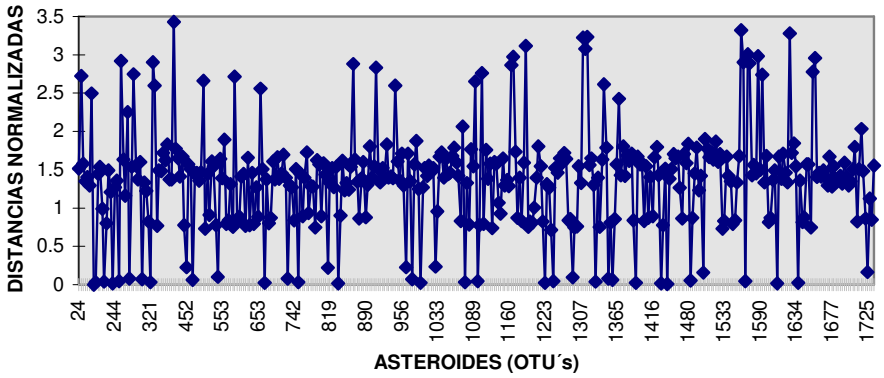
ESPECTRO DEL ASTEROIDE THEMIS



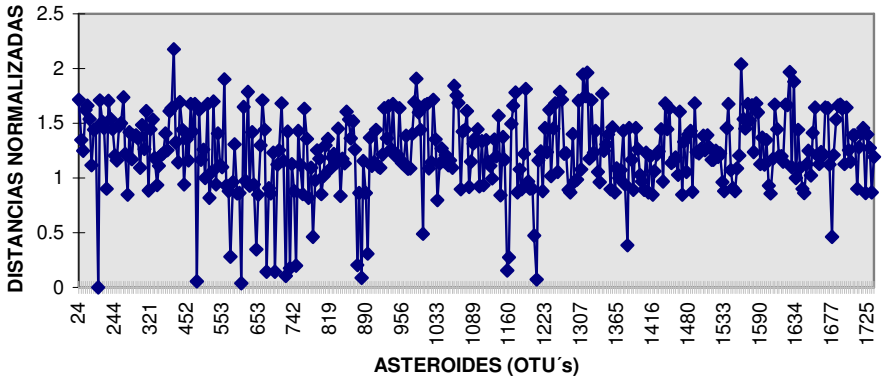
ESPECTRO DEL ASTEROIDE PHOCAEA

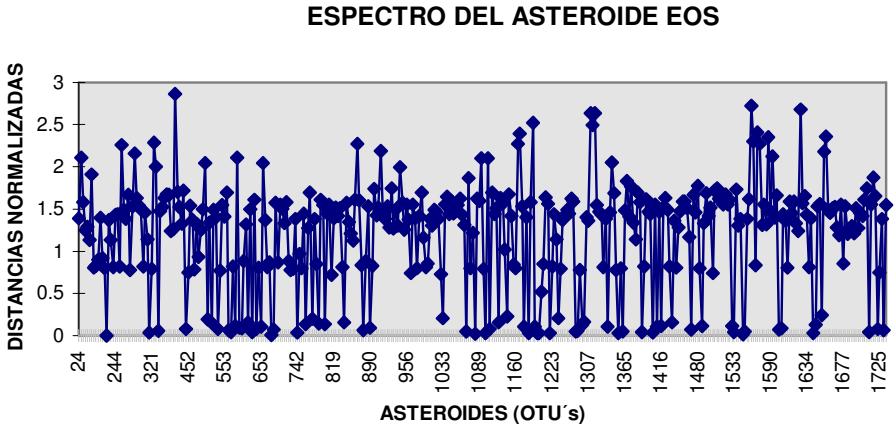


ESPECTRO DEL ASTEROIDE KORONIS



ESPECTRO DEL ASTEROIDE MARIA





Diseño de Registros (lenguaje ADA)

7.2.1.1. REGISTRO DE ELEMENTOS IMPROPIOS

Es el registro utilizado para todas las efemérides de los asteroides, cerca de 5.000 , con elementos cuya medición no elimina la perturbación.
Se utilizaron distintas muestras para ajustar el método y gran parte del algoritmo.

```
type aste is record
    numero: integer;
    nombre: string(1..17);
    h,g,anomalia,argumento_perihelio,longitud_nodo,
    inclinacion,excentricidad,movimiento_angular,
    semieje:float;
    end record;
```

7.2.1.2. REGISTRO DE ELEMENTOS PROPIOS

Este registro corresponde al archivo de elementos propios, donde se incluye además la calidad del algoritmo, que como se explicara es ambigua y hasta

errónea, al igual que la familia, ya discutida en este párrafo y la referencia que es un dato bibliográfico.

```
type aste is record
 numero: integer;
 semieje,excentricidad,seno_inclinacion,argumento_perihelio,
 longitud_nodo,relacion_perihelio,relacion_nodo: float;
 resonancia: string(1..3);
 distancia_marte,distancia_jupiter: float;
 calidad: string(1..1);
 familia: string(1..3);
 referencia: string(1..1);
    end record;
```

7.2.1.3. REGISTRO DE ELEMENTOS PROPIOS CON ALGUNOS DATOS IMPROPIOS

Se incluyó el nombre, que figuraba en el archivo original, la inclinación se obtuvo del archivo de elementos impropios por considerarlo un dato significativo ante el seno del mismo.

Es la estructura que se utiliza en la matriz de datos de [1] a [10].

```
type aste is record
 numero: integer;
 nombre: string(1..17);
 semieje,excentricidad, inclinacion,
 seno_inclinacion,argumento_perihelio: float;
 resonancia: string(1..3);
 distancia_marte,distancia_jupiter: float;
end record;
```

7.2.2 Matriz de similitud

La matriz de similitud surge del cálculo de las distancias una vez normalizados los dominios de los caracteres, y tiene una dimensión de t x t que es más de 10^5 de elementos (379 x 379) por la muestra de asteroides que se tomó y el total de información normalizada tiene más de 6 . 10^6 de almacenamiento y sobre todos los asteroides 1,5 Gigab.
Por ello es que se representa solo una parte de la matriz de similitud, para mostrar la simetría y que la diagonal principal es nula pues corresponde a la identidad.

[8]	[24]	[25]	[43]	[62]	[90]	[104]	[105]	[158]	[167]	[170] →
0	1.4537	1.7589	0.1346	1.4169	1.4558	1.4294	1.5548	1.6052	1.6124	
1.4537	0	2.3357	1.4222	0.1677	0.119	0.2623	2.2025	1.512	1.5422	
1.7589	2.3357	0	1.8883	2.3425	2.3461	2.2828	0.2493	2.722	2.7367	
0.1346	1.4222	1.8883	0	1.3846	1.4252	1.4071	1.6862	1.5736	1.5803	
1.4169	0.1677	2.3425	1.3846	0	0.0642	0.1319	2.196	1.3469	1.3774	
1.4558	0.119	2.3461	1.4252	0.0642	0	0.1545	2.2054	1.4065	1.4373	
1.4294	0.2623	2.2828	1.4071	0.1319	0.1545	0	2.135	1.296	1.3275	
1.5548	2.2025	0.2493	1.6862	2.196	2.2054	2.135	0	2.4988	2.5123	
1.6052	1.512	2.722	1.5736	1.3469	1.4065	1.296	2.4988	0	0.0343	
1.6124	1.5422	2.7367	1.5803	1.3774	1.4373	1.3275	2.5123	0.0343	0	
1.1488	1.7116	1.3457	1.2467	1.6234	1.6592	1.5402	1.1127	1.4402	1.4499	
1.4539	0.0329	2.3151	1.4246	0.1927	0.1422	0.2766	2.1844	1.5359	1.5661	
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓→	

Algunos asteroides que se ven en la representación:

8	**Flora**
24	**Themis**
25	**Phocaea**
43	**Ariadne**
62	**Erato**
90	**Antiope**
104	**Klymene**
105	**Artemis**
158	**Koronis**
167	**Urda**
170	**Maria**

7.2.3. Estructuración

A partir de la matriz de similitud se obtienen los espectros característicos de los asteroides (OTU´s), número del elemento funcional a la distancia, que muestra mucho más claramente que otras formas geométricas (v.g.: los fenogramas) cual es la estructura de la distribución y el aporte de los demás elementos a la agregación o agrupamiento de los mismos en clusters, familias.

En la algorítmia, luego de la normalización, se forman conjuntos parte C_i, en función de la varianza unitaria $\sigma = 1$ y según lo desarrollado en la metodología se ajustan los conjuntos partes a partir del maximal k. σ y de un radio de la densidad de distribución de cada C_i.

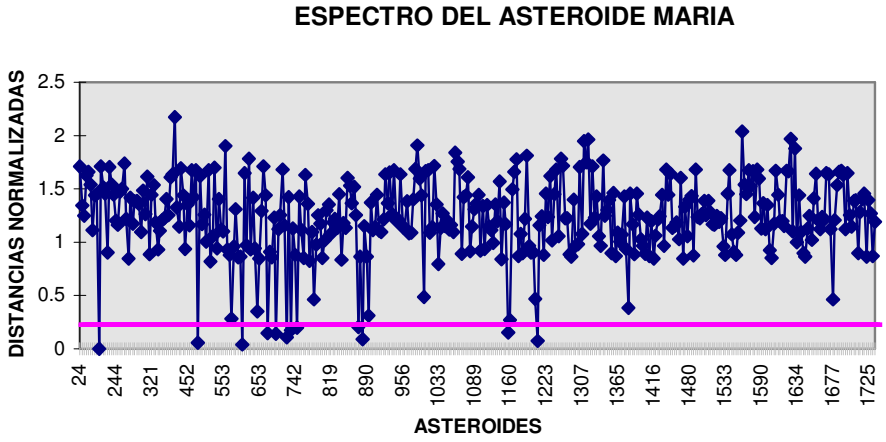
Si se observan los agrupamientos hay elementos que forman parte de una intersección de los conjuntos partes, mostrándose que por superposición se agrupan por vecino más cercano y con la técnica de pair-group.

De esta manera van quedando conjuntos vacíos en la estructuración por interferencia negativa. Están los elementos más cercanos a un conjunto que a otro aunque el rango sea aparentemente mayor.

Al hacer el gráfico del espectro característico incluyendo la recta de distancia igual al rango, los elementos que quedan por encima de la recta no pertenecen al conjunto parte mientras que los que están por debajo si, formándose así agrupamientos o clusters, ya que se observa que en todos los espectros de los elementos que están por debajo del rango son comunes y más cercanos, esto se puede observar hasta en los gráficos (en los fenogramas nada se observa en este sentido).

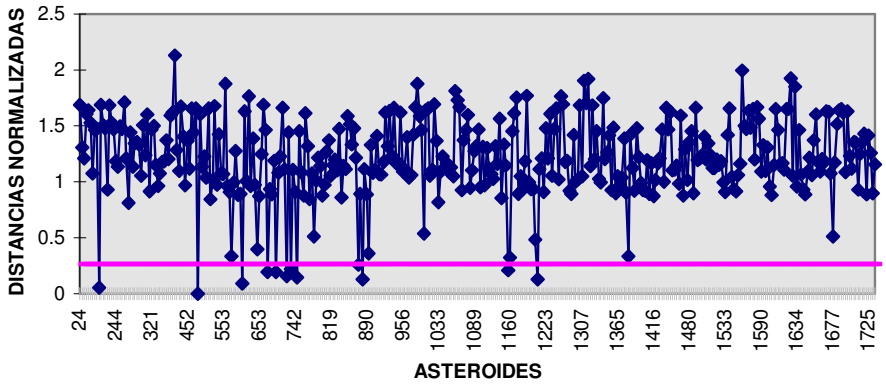
7.2.3.1. Familia María (20 elementos)

En forma representativa se ilustra con la familia MARÏA completa, pues la cantidad de espacio para imprimir todas las familias es muy grande y no aporta a la metodología.



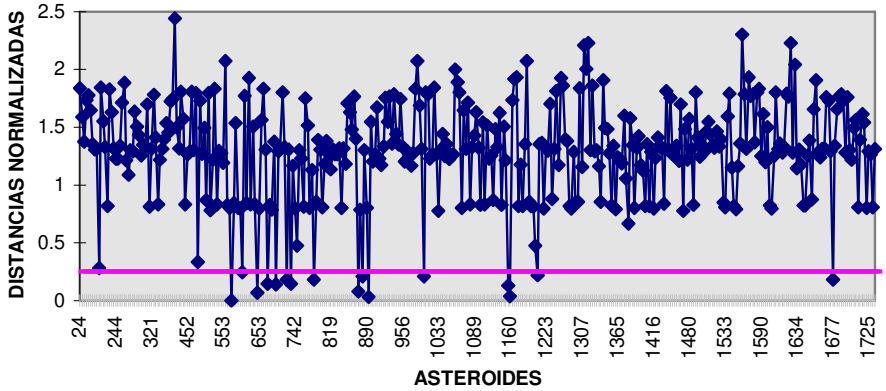
Invariantes:
Distancia Media 0.1261
Densidad 12.00
Dispersion 0.0568
Rango 0.2245

ESPECTRO DEL ASTEROIDE ROMA



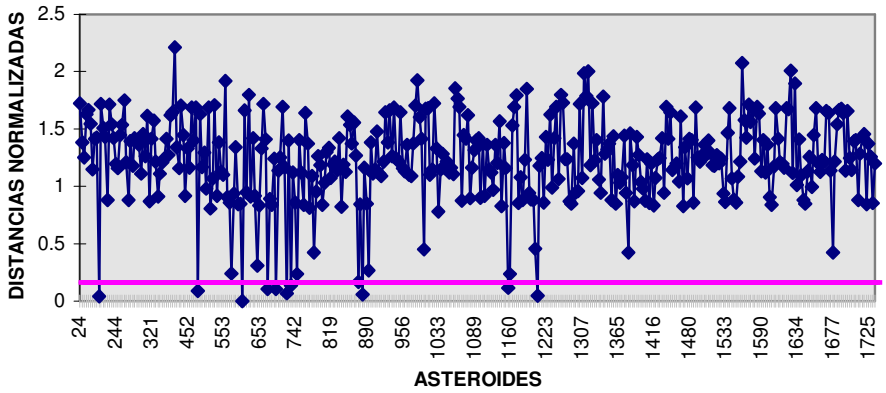
Invariantes:
Distancia Media 0.1613
Densidad 12.00
Dispersion 0.0593
Rango 0.2639

ESPECTRO DEL ASTEROIDE RENATE



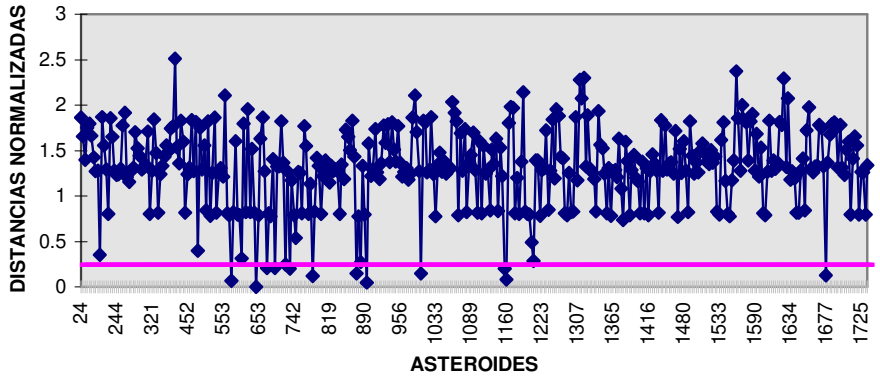
Invariantes:
Distancia Media 0.1394
Densidad 15.00
Dispersion 0.0642
Rango 0.2505

ESPECTRO DEL ASTEROIDE ELLY



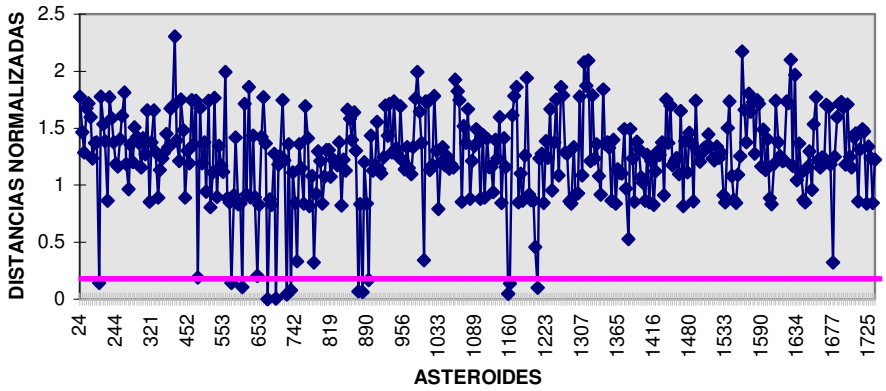
Invariantes:
Distancia Media 0.0933
Densidad 11.00
Dispersion 0.0411
Rango 0.1646

ESPECTRO DEL ASTEROIDE JUBILATRIX



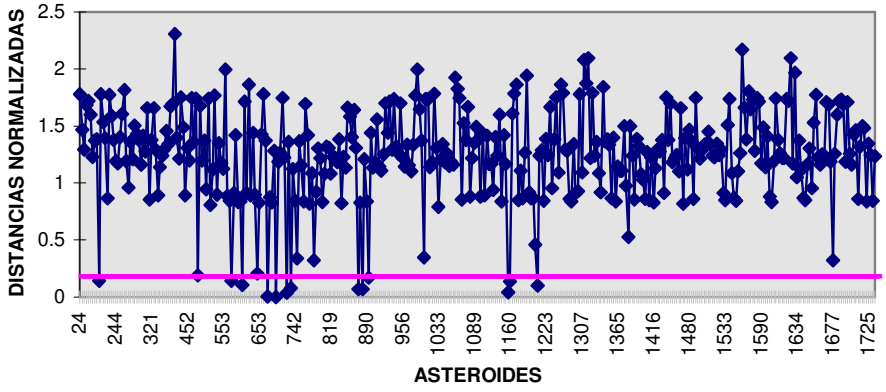
Invariantes:
Distancia Media 0.1393
Densidad 12.00
Dispersion 0.0589
Rango 0.2412

ESPECTRO DEL ASTEROIDE CRESCENTIA



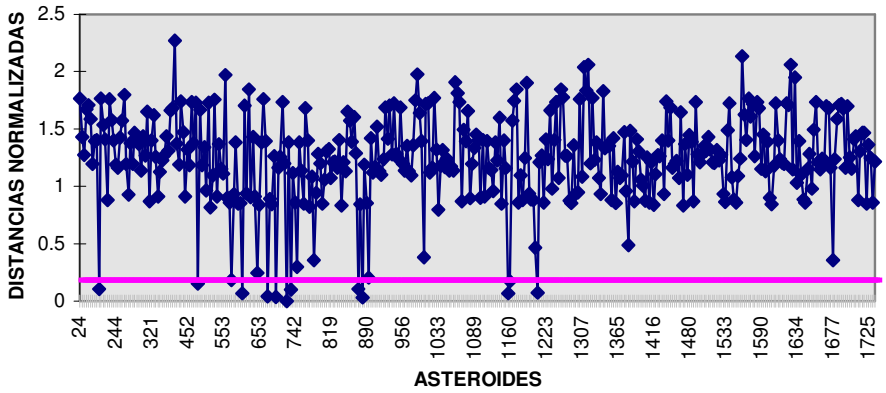
Invariantes:
Distancia Media 0.0926
Densidad 13.00
Dispersion 0.0491
Rango 0.1777

ESPECTRO DEL ASTEROIDE BELLA



Invariantes:
Distancia Media 0.0916
Densidad 13.00
Dispersion 0.0493
Rango 0.1771

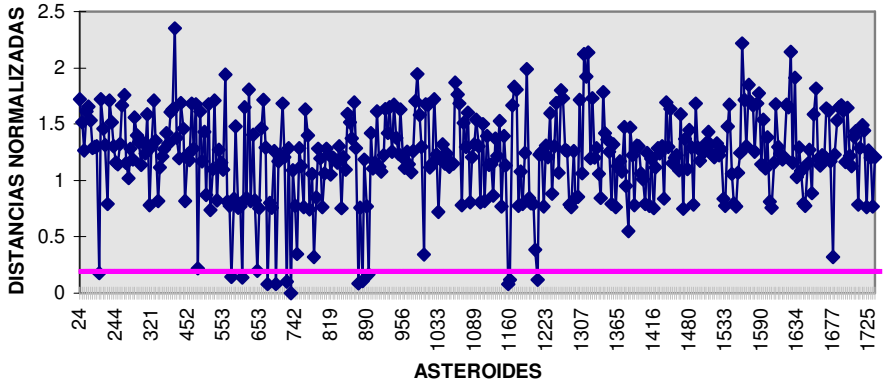
ESPECTRO DEL ASTEROIDE ULULA



Invariantes:

Distancia Media	0.0958
Densidad	13.00
Dispersion	0.0517
Rango	0.1855

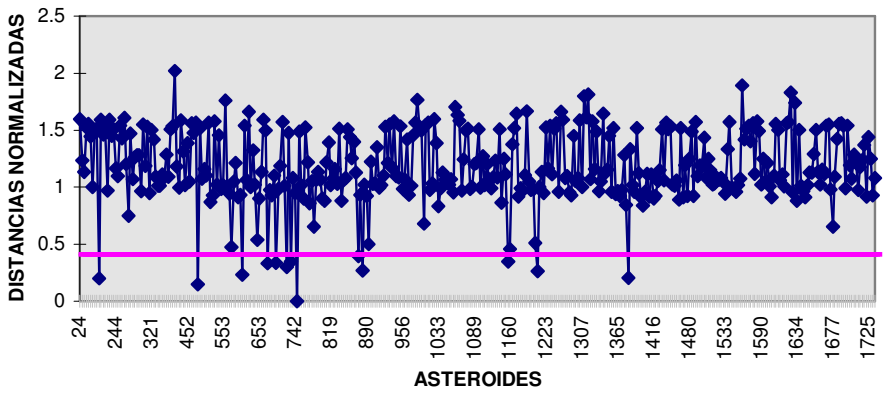
ESPECTRO DEL ASTEROIDE NIPPONIA



Invariantes:

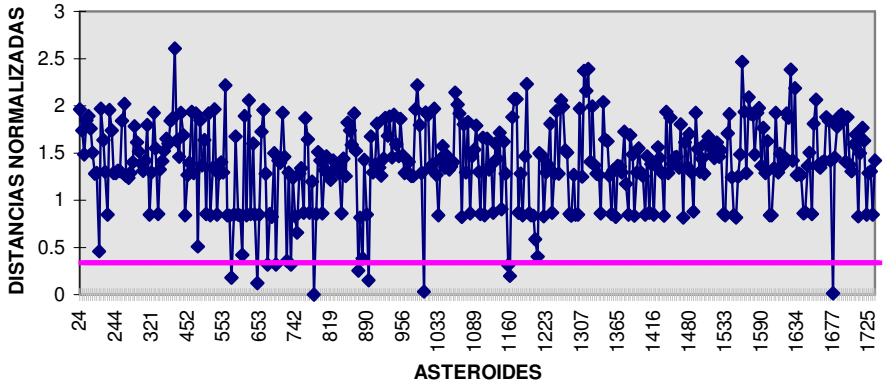
Distancia Media	0.1230
Densidad	14.00
Dispersion	0.0389
Rango	0.1904

ESPECTRO DEL ASTEROIDE FAINA



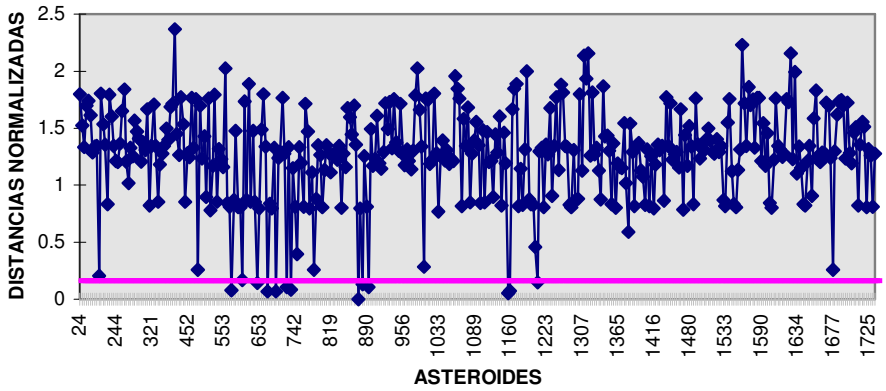
Invariantes:
Distancia Media 0.2821
Densidad 13.00
Dispersion 0.0748
Rango 0.4116

ESPECTRO DEL ASTEROIDE MOSKVA



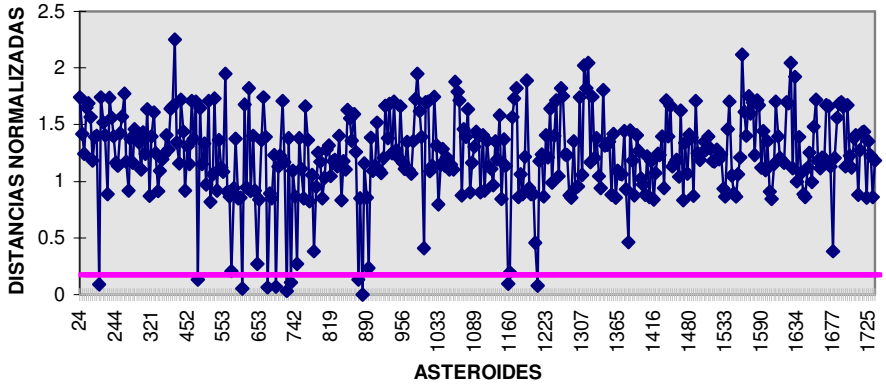
Invariantes:
Distancia Media 0.1583
Densidad 9.00
Dispersion 0.1019
Rango 0.3348

ESPECTRO DEL ASTEROIDE NYMPHE



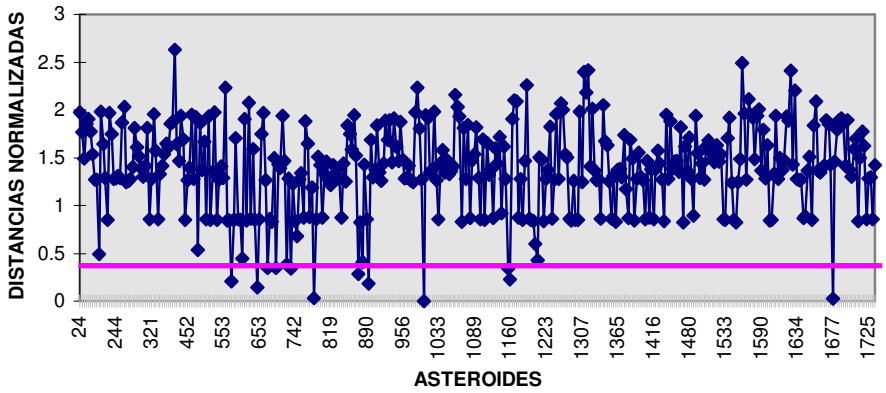
Invariantes:
Distancia Media 0.1026
Densidad 13.00
Dispersion 0.0366
Rango 0.1659

ESPECTRO DEL ASTEROIDE RICARDA



Invariantes:
Distancia Media 0.0961
Densidad 12.00
Dispersion 0.0457
Rango 0.1752

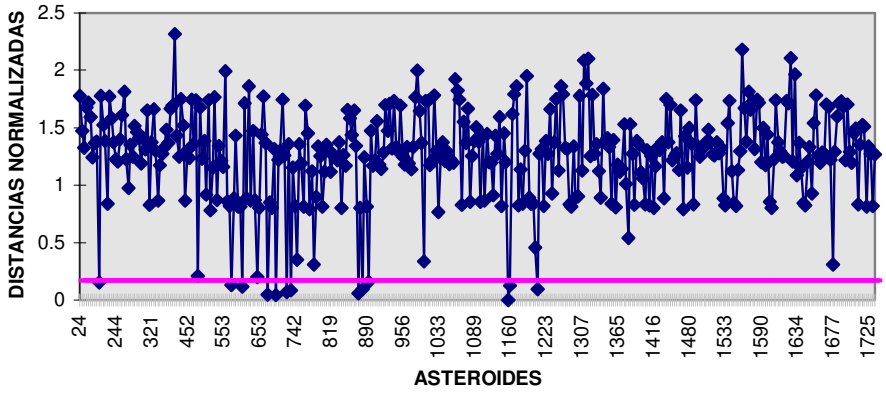
ESPECTRO DEL ASTEROIDE OTTHILD



Invariantes:

Distancia Media	0.1800
Densidad	9.00
Dispersion	0.1096
Rango	0.3698

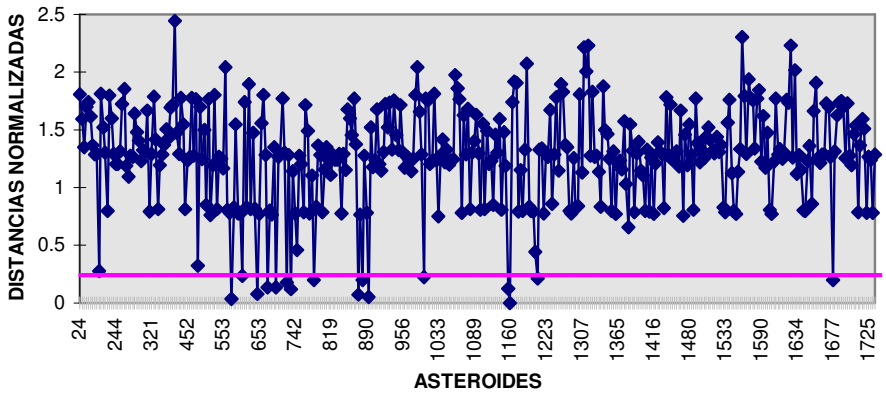
ESPECTRO DEL ASTEROIDE LUDA



Invariantes:

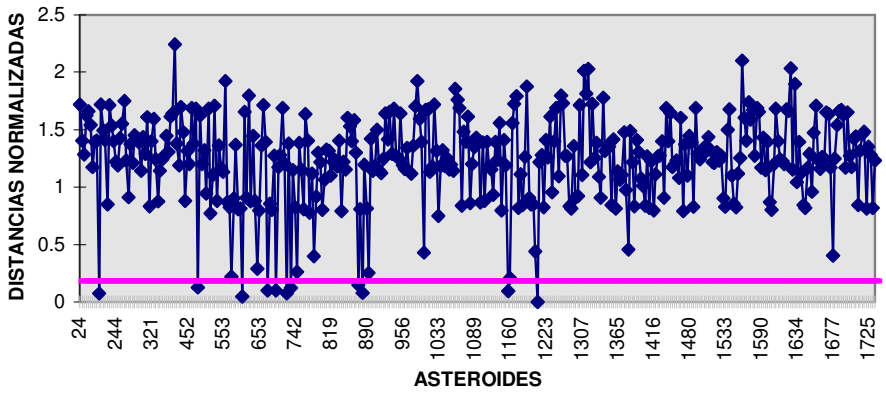
Distancia Media	0.0977
Densidad	13.00
Dispersion	0.0397
Rango	0.1666

ESPECTRO DEL ASTEROIDE ILLYRIA



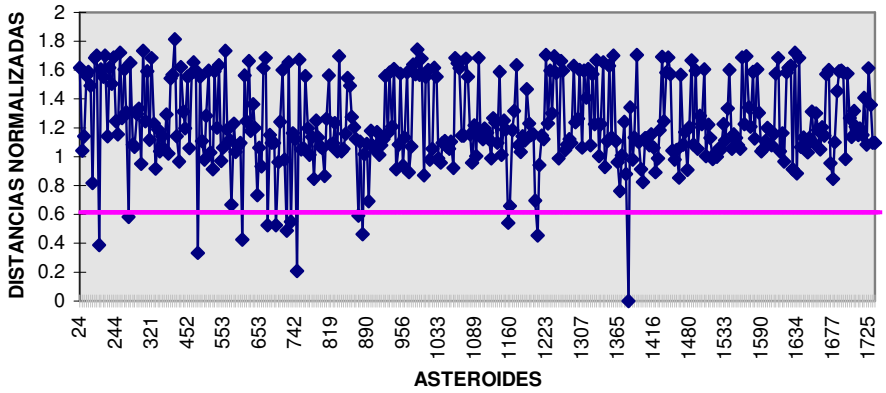
Invariantes:
Distancia Media 0.1338
Densidad 14.00
Dispersion 0.0606
Rango 0.2387

ESPECTRO DEL ASTEROIDE BOYER



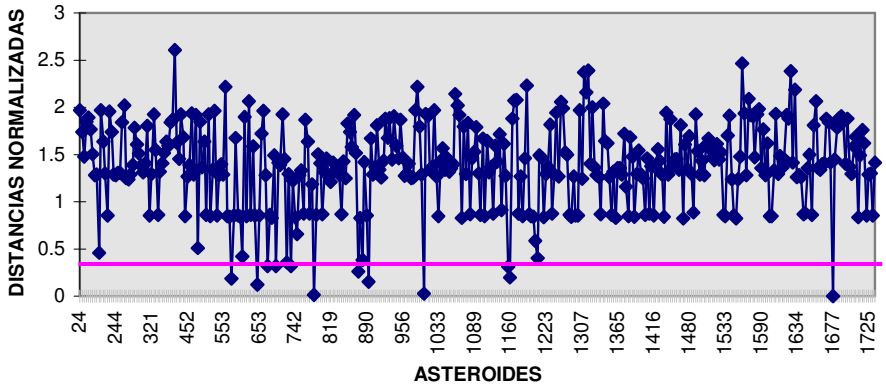
Invariantes:
Distancia Media 0.1064
Densidad 12.00
Dispersion 0.0446
Rango 0.1837

ESPECTRO DEL ASTEROIDE LOMONOWA

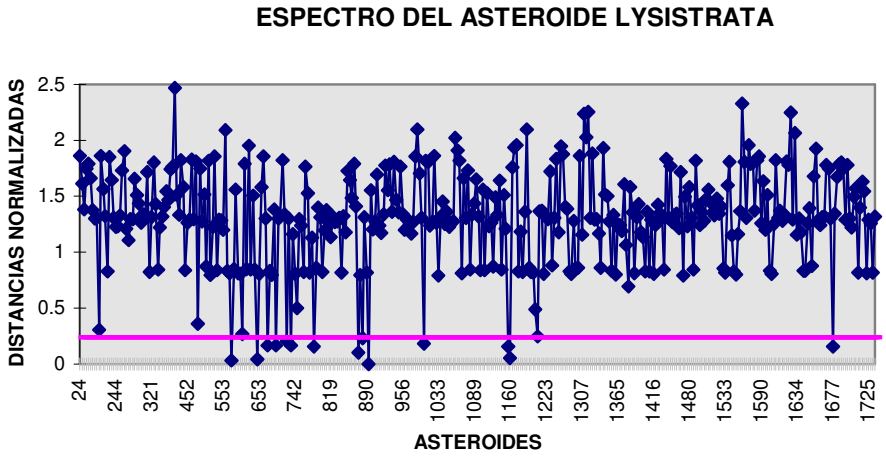


Invariantes:
Distancia Media 0.4350
Densidad 11.00
Dispersion 0.1030
Rango 0.6135

ESPECTRO DEL ASTEROIDE TYCHO BAHE



Invariantes:
Distancia Media 0.1586
Densidad 9.00
Dispersion 0.1035
Rango 0.3378



Invariantes:
Distancia Media 0.1321
Densidad 13.00
Dispersion 0.059
Rango 0.2343

7.2.4. Testeo usando Minería de Datos (Data Mining)

Un sistema de software fue construido para evaluar el algoritmo C4.5. Este sistema toma datos de entrenamiento como una entrada (INPUT) y le permite al usuario escoger si él quiere construir un árbol de decisión según C4.5. Si el usuario escoge C4.5, el árbol de decisión se genera, entonces se poda y las reglas de decisión se construyen.
Se evalúan los árboles de decisión y los conjuntos de reglas generados por C4.5, separados entre ellos.
Se usa el sistema para probar los algoritmos en dominios diferentes, principalmente Elita, la base de asteroides.

7.2.4.1 Cómputo de la Ganancia de Información

En los casos, en aquéllos que el conjunto T, contiene ejemplos que pertenecen a clases diferentes, se lleva a cabo una prueba con atributos diferentes y se consigue una partición según el "mejor" atributo. Para encontrar el "mejor" atributo, se usa la teoría de la información, que sostiene que la información se

maximiza cuando la entropía se minimiza. La entropía determina la aleatoriedad o desorden de un conjunto [Perichinsky, 2000 - 2007] (ver Capítulo **4**, de solución propuesta).

Se supone que se tienen ejemplos negativos y positivos. En este contexto la entropía del subconjunto Si, H (Si), puede calcularse como:

$$H(S_i) = -p_i^+ \log p_i^+ - p_i^- \log p_i^- \qquad \textbf{7.2.4.1.1.}$$

Donde p_i^+ es la probabilidad de que un ejemplo sea tomado en modo aleatorio (random) de S_i será positivo. Esa probabilidad puede ser calculada como

$$p_i^+ = \frac{n_i^+}{n_i^+ + n_i^-} \qquad \textbf{7.2.4.1.2.}$$

Siendo n_i^+ la cantidad de ejemplos positivos de S_i, y n_i^- la cantidad de ejemplos negativos.

La probabilidad p_i^- es calculada en forma análoga que para p_i^+, reemplazando la cantidad de ejemplos positivos por la cantidad de ejemplos negativos, y viceversa.

Generalizando la expresión ***(7.2.4.1.1.)*** para cualquier tipo de ejemplos, se obtiene la fórmula general de la entropía:

$$H(S_i) = \sum_{i=1}^{n} - p_i \log p_i \qquad \textbf{7.2.4.1.3.}$$

Para todo el cálculo relativo a la entropía, se define ***0log0*** igual a ***0***.

Si el atributo ***at*** divide al conjunto S en subconjuntos S_i, i = 1, 2, … , j , … ,n , entonces, la entropía total del sistema de subconjuntos será:

$$H(S, at) = \sum_{i=1}^{n} P(S_i) \cdot H(S_i) \qquad \textbf{7.2.4.1.4.}$$

Donde $H(S_i)$ is la entropís del subconjunto S_i y $P(S_i)$ es la probabilidad del hecho que un ejemplo pertenezca a S_i. Puede ser calculada, usando los tamaños relativos de los subconjuntos, como:

$$P(S_i) = \frac{|S_i|}{|S|} \qquad \textbf{7.2.4.1.5.}$$

La ganancia de información puede ser calculada como la disminución de la entropía. Así:

$$I(S, at) = H(S) - H(S, at) \qquad \textbf{7.2.4.1.6.}$$

Donde $H(S)$ es a priori el valor de la entropía, antes de lograr la subdivisión, y $H(S,at)$ es el valor de la entropía del sistema de los subconjuntos generado por la partición según ***at***.

El uso de la entropía para evaluar el mejor atributo no es el único método existente o usado en el Aprendizaje Automático (Automatic Learning). Sin embargo, es usado por Quinlan al desarrollar el ID3 y su teniendo el exitoso C4.5.

7.2.4.2. Datos Numéricos

Los árboles de decisión pueden generarse tanto como atributos discretos como atributos continuos. Cuando se trabaja con atributos discretos, la partición del conjunto de acuerdo al valor de un atributo, es simple.

Para resolver este problema, puede recurrirse al método binario. Este método consiste en ir formando dos rangos de valores de acuerdo al valor de un atributo que puede tomarse como simbólicos.

7.2.4.3. Resultados y Conclusiones

7.2.4.3.1. Resultados de C4.5

C4.5 con post-poda, resultan árboles menores y menos frondosos (espesos). Si se analizan los árboles obtenidos en el dominio, se ve que los porcentajes de error resultantes de C4.5 están entre un 3% y un 3,7%, puesto que C4.5 genera árboles menores y conjuntos de reglas menores. Derivados del hecho que cada hoja en un árbol generó las envolventes de una distribución de clases.

7.2.4.3.2. Procentaje de error

{ELITA} {[1]: C4.5- Árboles de Ganancia [2]: C4.5- Reglas de Ganancia [3]: C4.5-Proporción de Árboles de Ganancia [4]: C4.5- Proporción de Reglas de Árboles de Ganancia} < 3% [Tesis de grado, de la Ing. Magdalena Servente

(2002), Departamento de Computación, Facultad de Ingeniería, Universidad de Buenos Aires.] [Perichinsky, 2003a, 2003d].

Del análisis de éste valor se podría concluir que ningún método puede generar un modelo claramente superior para el dominio. Al contrario, se podría establecer que el porcentaje de error no parece depender del método usado, pero en el dominio analizado.

7.2.4.4. Espacio de Hipótesis

El espacio de hipótesis, para este algoritmo, está completo de acuerdo a los atributos disponibles. Porque cualquier prueba de valor puede representarse con un árbol de decisión, este algoritmo evita uno de los riesgos principales del método inductivo, que trabaja reduciendo los espacios para las hipótesis.

Un rasgo importante del algoritmo C4.5, es que usa todos los datos disponibles en cada paso para elegir el "mejor" atributo; ésta es una decisión que surge de métodos estadísticos. Este hecho favorece a éste algoritmo por encima de otros algoritmos porque analiza cómo los conjuntos de datos de INPUT toman la representación en los árboles de decisión en forma consistente.

Una vez que un atributo ha sido seleccionado como un nodo de decisión, el algoritmo no se remonta por encima de sus opciones. Ésta es la razón por qué este algoritmo puede converger a un máximo local [Michie, 1986, 1988] [Quinlan, 1986 - 1996] [Mitchell, 2000]. El algoritmo de C4.5 agrega un cierto grado de reconsideración de sus opciones en la post-poda de los árboles de decisión.

No obstante, se puede establecer que la muestra de los resultados, que la proporción de error depende del dominio de los datos. Para el estudio del futuro, pienso que se debe hacer un análisis del conjunto de datos de INPUT, con métodos de clustering numérico y escoger para el dominio, el método que mantenga un porcentaje de error bajo para bases de datos extensas como un método de robustez.

ALGORITMIA

El orden y la conexión de las ideas es lo mismo que el orden y la conexión de las cosas

Baruch (Benedict de) Spinoza

CCXLII

8. ALGORITMIA

with Ada…➔ …

procedure **unasa** is

procedure **impm** () is

```
subtype rangoM is integer range 1.. ξ;
subtype rangoE is integer range 1..4;
```

<CONJUNTO DE ARREGLOS>

< REGISTRO VISIÓN>

<ADMINISTRACION DE MEMORIA>

```
type floatarray is array (rangoM) of float;
type ptrfloatarray is access floatarray;
type matriz is array (rangoM) of ptrfloatarray;


type enteroarray is array (rangoM) of integer;
type ptrenteroarray is access enteroarray;
type matrize is array (rangoM) of ptrenteroarray;
```

<ASIGNACION DE ARCHIVOS>

<ASIGNACION REGISTROS INSTANCIAS DE VISIÓN (GET)>

<LOCALIZACIÓN E INICIALIZACION DE ARREGLOS>

```
procedure alocarmatriz(mtaxa: in out matriz) is
begin
  for i in rangoM loop
    mtaxa(i) := new floatarray;
  end loop;
end;

procedure alocarmatrize(mtaxones: in out matrize) is
begin
  for i in rangoM loop
```

```
    mtaxones(i) := new enteroarray;
   end loop;
  end;


  procedure borrarmatriz(mtaxa: in out matriz) is ➔ …

  procedure borrarmatrize(mtaxones: in out matrize) is ➔ …
```

<CARGA DE DOMINIOS, DEFINIR E INICIALIZAR VARIABLES>

<ARMADO DE LA VISIÓN>

<COMIENZA EL PROCEDIMIENTO>

```
begin

   for i in 1..ξ loop
     leeaster (vision);
     xnombre:=vision.nombre;
     ns(i):=xnombre;
     xnumero:=vision.numero;
     nus(i):=xnumero;
     xsemieje:=vision.semieje;
     m(i,1):=xsemieje;
     xexcentricidad:=vision.excentricidad;
     m(i,2):=xexcentricidad;
     xinclinacion:=vision.seno_inclinacion;
     m(i,3):= xinclinacion;
     .............
          ↓
     .............
   end loop;

close (f);
```

<CÁLCULO DE PROMEDIO DE LOS DOMINIOS>

```
   for j in 1..γ loop
     v1(j):= 0.0;
       for i in 1.. ξ loop
         v1(j):= v1(j) + m(i,j);
       end loop;
    end loop;

   for i in 1.. γ loop
      v1(i):= v1(i) / ξ.0;
   end loop;
```

<CÁLCULO DE SIMILITUDES EUCLÍDEAS>

```
for j in 1.. γ loop
  v2(j):= 0.0;
    for i in 1.. ξ loop
      v2(j):= v2(j) + (v1(j) - m(i,j))**2;
    end loop;
end loop;

for i in 1.. γ loop
    v2(i):= (v2(i)/( ξ - 1.0 )**0.5;        <VECTOR DE COVARIANCIA>
end loop;
```

<DOMINIOS DE DATOS ➔ VALORES>

<NORMALIZACIÓN DE VALORES DE ATRIBUTOS>

```
for j in 1.. γ loop
    for i in 1.. ξ loop
    p(i,j):= (m(i,j)-v1(j))/v2(j);
    end loop;
end loop;

for j in 1.. ξ loop
      vs(j):=0.0;
    for i in 1.. ξ loop
      r(i,j):= 0.0;
                if j < 5 then tax(i,j):= 0.0; end if;
    end loop;
end loop;

for i in 1.. ξ loop
    for j in 1.. ξ loop
     simil:=0.0;
            for k in 3 loop..

                simil:= simil + (p(i,k) - p(j,k))**2;

            end loop;

              r(i,j):= (simil/γ)**0.5;

              if disimil < r(i,j) then disimil:= r(i,j);
              end if;

--              if i>j then mt2(i,j):= r(i,j); else mt2(i,j):=0.0;
--              end if;
```

```
            mt2(i,j):= r(i,j);


   end loop;
end loop;
```

<IDENTIFICACION DE OTU´s AISLADOS>

↓

<AGRUPAMIENTO POR INVARIANTES>
<MATRIZ DE SIMILITUD>

```
 for i in 1.. ξ loop

    for lj in 1.. ξ loop
          if  (mt2(i,lj) > 0.0) and (mt2(i,lj) < 1.0) then
             distancia_media := distancia_media + mt2(i,lj);
             ene := ene + 1.0;
          end if;
     end loop;

        distancia_media := distancia_media / ene;
```

<ITERACIÓN ALREDEDOR DEL CENTROIDE>

```
       radio := distancia_media;
       ene := 0.0;
       distancia_media := 0.0;

    for lj in 1.. ξ loop
          if  (mt2(i,lj) > 0.0) and (mt2(i,lj) < radio) then
             distancia_media := distancia_media + mt2(i,lj);
             ene := ene + 1.0;
          end if;
     end loop;

        distancia_media := distancia_media / ene;


for lj in 1.. ξ loop
    if (mt2(i,lj) > 0.0) and  (mt2(i,lj) < distancia_media) then
          xsigma := xsigma + (distancia_media - mt2(i,lj))**2;
    end if;
 end loop;

   radio := (xsigma/(ene - 1.0))**0.5;
```

rango := xk*radio + distancia_media;

ene := 0.0;

for j in 1.. ξ loop

if (mt2(i,j) < distancia_media) then
ene := ene + 1.0;
r(i,j) := mt2(i,j);
end if;

end loop;

<INVARIANTES>

tax(i,1) := **distancia_media**;
tax(i,2) := **ene**; **<DENSIDAD>**
tax(i,3) := **radio**;
tax(i,4) := **rango**;

end loop;

<AGRUPAMIENTO POR INVARIANTES>

<ARMADO DE FAMILIAS>

<TAXONES>

<TRATAMIENTO DE FAMILIAS NO SOLAPADAS>

lim := 0;
jmx := 0;
jkx := **ξ - 1**;
ikx := **ξ**;

for i in 1..jkx loop
if (taxones(i)(i) > -1) then

ik := i + 1;
for imx in ik..ikx loop
for j in rangoM loop

<IDENTIFICACION DE OTU´s>

if ((i /= j) and (taxones(i)(j) = 1)) and ((imx /= j) and (taxones(imx)(j) = 1)) then

if (tax(i,1) > tax(imx,1)) then

<OTU´s AISLADOS>

```
                                    taxones(i)(j) := 0;
                                    taxones(imx)(j) := 1;
                                                    else
                                    taxones(i)(j) := 1;
                                    taxones(imx)(j) := 0;
                                end if;
            end if;
                                    end loop;

        end loop;

            if (taxones(i)(i) < 0) then
                    tax(i,1) := 0.0;
                    tax(i,2) := 0.0;
                    tax(i,3) := 0.0;
                    tax(i,4) := 0.0;
            end if;

    end if;

    end loop;
```

<ELIMINACION DE OTU´s EXTERNOS>

```
    for i in rangoM loop

    if (taxones(i)(i) > -1) then

        for j in rangoM loop

            if ((i /= j) and (taxones(i)(j) = 1)) then

                                    if (matnt(i)(j) > tax(i,1)) then

                                        taxones(i)(j) := 0;

                                    end if;

            end if;

        end loop;

    end if;

    end loop;


lim := 0;
jmx := 0;
jkx := ξ - 1;
```

```
 for i in 1..jkx loop

 if (taxones(i)(i) > -1) then

      for j in rangoM loop
          if ((i /= j) and (taxones(i)(j) = 1)) then
                          jmx := i + 1;
                                  for imx in jmx..ikx loop
                                    taxones(imx)(j) := 0;
                                  end loop;

          end if;


      end loop;

 end if;

 end loop;
```

<ITERACION ALREDEDOR DEL CENTROIDE>

```
lim := 0;
jmx := 0;
jkx := ξ - 1;

 for i in rangoM loop

        ene := 0.0;
        distancia_media := 0.0;
        xsigma := 0.0;
        radio := 0.0;
        rango := 0.0;

  if (taxones(i)(i) < 0) then
                        tax(i,1) := 0.0;
                        tax(i,2) := 0.0;
                        tax(i,3) := 0.0;
                        tax(i,4) := 0.0;
  end if;

      for j in rangoM loop
```

<DISTANCIA MEDIA>

```
          if ((i /= j) and (taxones(i)(j) = 1)) then
               distancia_media := distancia_media + matnt(i)(j);
               ene := ene + 1.0;
          end if;
```

```
 end loop;

        distancia_media := distancia_media / ene;
```

<VARIANZA y DISPERSIÓN>

```
    for j in rangoM loop

       if ((i /= j) and (taxones(i)(j) = 1)) then
            xsigma := xsigma + (distancia_media - matnt(i)(j))**2;
       end if;

    end loop;

        radio := (xsigma/(ene))**0.5;

        rango := xk*radio + distancia_media;

        simil := 2.0 * radio + distancia_media;

    for j in rangoM loop

       if ((i /= j) and (taxones(i)(j) = 1)) then
         if  (matnt(i)(j) > rango) then
            taxones(i)(j) := 0;
         end if;
       end if;

 end loop;
```

<INVARIANTES>

```
            tax(i,1) := distancia_media;
            tax(i,2) := ene;
            tax(i,3) := radio;
            tax(i,4) := rango;

 end loop;
```

<ARMADO DE FAMILIAS>

```
lim := 0;
jmx := 0;
jkx := ξ - 1;
ikx := ξ;

  for i in 1..jkx loop
```

```
 if (taxones(i)(i) > -1) then

                                        lix := 0;
                                        for j in rangoM loop
                                         if ((i /= j) and (taxones(i)(j) = 1)) then
                                         lix := lix + 1;
                                         end if;
                                        end loop;
                                         if lix /=0 then
                                         jmx := jmx + lix;
                                         lim := lim + 1;
                                         taxones(i)(i) := lim;
                                         fam(i) := lix;
                                               else
                                         taxones(i)(i) := -2;
                                         end if;

 end if;

 end loop;
```

<COMPLETA LOS CLUSTERS CON UN WHILE>
<ITERACION CON LAS DENSIDADES>

```
                  While tax(i,2) > -2    <loops de 1 a 11>

1
 for i in 1.. ξ loop

          simil := tax(i,1) + tax(i,3);
1a
                                  for lim in 1.. ξ loop
2
   for j in 1.. ξ loop

11
                       if (r(i,j) > (tax(i,1) + tax(i,3))) then
                                  r(i,j) := 0.0;
                       end if;
11

3
     if r(i,j) /= 0.0 then

          imx := j;

4
              for jmx in 1.. ξ loop
```

```
5
    if r(imx,jmx) /= 0.0 then
6
        if (r(imx,jmx) /= 0.0) then
7
            if (i /= jmx) then
8
                if (r(i,jmx) = 0.0) then
9
                    if (r(imx,jmx) < (tax(imx,1) + tax(imx,3))) then
                        r(i,jmx) := r(imx,jmx);
                    end if;
9
                end if;
8
            end if;
7
        end if;
6
    end if;
5
            end loop;
4
        simil := tax(i,1) + tax(i,3);
10
    for jlx in 1.. ξ loop
        r(imx,jlx) := 0.0;
    end loop;
10
    end if;
3
  end loop;
2
                            end loop;
```

<COMPLETA LOS CLUSTERS DEL WHILE>

1a

```
 end loop;
```

<DISTORSIÓN DEL ESPACIO>

```
      for lj in 1.. ξ loop
       if r(i,lj) /= 0.0 then
         li := i + 1;
           for ik in li.. ξ loop
               r(ik,lj) := 0.0;
           end loop;
        end if;
       end loop;

 end impm;


begin

impm();


end unasa;
```

CONCLUSIONES

El hombre sabio querrá estar siempre con alguien que sea mejor que el (Platon)

CLXXVI

9. CONCLUSIONES

9.1. APORTES ORIGINALES

- **Definición de un método numérico basado en invariantes para la identificación automática de pertenencia de OTU´s a una familia.**

- **Aplicación de la metodología a asteroides según las familias de Hirayama.**

- **Identificación, testeo y puesta a punto de las invariantes: centroide, varianza y radio del método definido.**

- **Comprobación del método para familias de asteroides en forma no arbitraria de identificación de elementos.**

- **Definición de medidas que permiten determinar la estabilidad de un familia por sensibilidad de atributos.**

- **Contrastación de las familias propuestas por Hirayama, por el método numérico descripto en la tesis (objetivo).**

- **Utilización original del método de superposición e interferencia de espectros para confirmar que los resultados obtenidos por el método numérico propuesto son congruentes (consistentes).**

- **Contrastación con Cúmulos, Nebulosas y Galaxias, comprobándose la separación de Galaxias gemelas a las cuales el método las separó en dos.**

9.2. FUTURAS LINEAS DE INVESTIGACION

- **Extensión del método propuesto a clasificación y búsqueda de eficiencia y performance en la algoritmia.**

- **Exploración de modelos conceptuales que justifiquen el ordenamiento de las familias entre si; en particular tecnologías emergentes.**

- **Avanzar en Inteligencia Artificial con las familias de Quinlan y similares, Redes Neuronales, tratando de concretar las pruebas en Back Propagation, que ya se comenzaron y en teoría de Onditas (Wavelet[1]) y Algoritmos Genéticos.**

[1] http://engineering.rowan.edu./~polikar/WAVELETS/WTpart(1 o 2).html ROBI POLIKAR

ANEXO I

MATRIZ DE DATOS

El valor del conocimiento está no en su acumulación, sino en su utilización

Green

CCLVII

10. ANEXO I

10.1. ELEMENTOS DE LA MATRIZ DE DATOS

[1] numero

[2] nombre

[3] semieje

[4] excentricidad

[5] inclinacion

[6] seno_inclinacion

[7] argumento_perihelio

[8] resonancia

[9] distancia_marte

[10] distancia_jupiter

10.2. MATRIZ DE DATOS

[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
8	Flora	2.201	0.141	5.88638	0.097	60	116.9	0.024	2.457
24	Themis	3.133	0.159	0.76258	0.02	155.3	315.1	0.794	1.409
25	Phocaea	2.4	0.183	21.57367	0.417	295.4	214.6	0.192	2.071
43	Ariadne	2.203	0.14	3.4662	0.071	260.7	261	0.046	2.473
62	Erato	3.122	0.146	2.22285	0.023	47.6	151.2	0.825	1.462
90	Antiope	3.148	0.15	2.23233	0.024	295.7	43.3	0.832	1.419
104	Klymene	3.149	0.141	2.8238	0.044	80.5	16	0.861	1.447
105	Artemis	2.374	0.168	21.48458	0.387	234.7	187	0.217	2.131
158	Koronis	2.869	0.045	1.00145	0.038	103.7	277.8	0.917	2.02
167	Urda	2.854	0.043	2.20436	0.037	236.7	197.2	0.909	2.042
170	Maria	2.554	0.099	14.42482	0.266	121.2	299.6	0.425	2.188
171	Ophelia	3.134	0.161	2.54263	0.024	160	101.8	0.787	1.401
184	Dejopeja	3.183	0.113	1.14847	0.038	187.4	304	0.979	1.499
208	Lacrimosa	2.893	0.045	1.75446	0.037	170.8	332.1	0.938	1.997
221	Eso	3.012	0.071	10.86469	0.174	318.9	147	0.957	1.796
222	Lucia	3.135	0.157	2.16047	0.019	247	57.7	0.801	1.412
223	Rosa	3.089	0.136	1.94484	0.027	124.4	10.8	0.831	1.525
243	Ida	2.862	0.045	1.14094	0.036	122.6	300.6	0.909	2.026
244	Sita	2.174	0.103	2.8407	0.06	39.8	212.5	0.096	2.574
254	Augusta	2.195	0.116	4.51629	0.07	231.1	22.9	0.088	2.531
263	Dresda	2.887	0.042	1.30803	0.037	23.3	245.1	0.943	2.01
265	Anna	2.419	0.175	25.62833	0.485	218.6	329.8	0.177	2.077
268	Adorea	3.097	0.17	2.43342	0.025	185.6	139.2	0.732	1.413
270	Anahita	2.198	0.092	2.36493	0.053	334.8	250.6	0.151	2.585
273	Atropos	2.395	0.149	20.41422	0.364	279.8	162.9	0.275	2.141
277	Elvira	2.886	0.051	1.15541	0.037	3.7	254	0.915	1.984
281	Lucretia	2.188	0.134	5.30895	0.084	75.2	26.6	0.026	2.484
290	Bruna	2.337	0.188	22.32099	0.406	127.8	12	0.145	2.128
291	Alice	2.222	0.141	1.8498	0.036	140.7	183.9	0.078	2.467
296	Phaetusa	2.229	0.123	1.74558	0.026	26.4	143.7	0.129	2.505
298	Baptistina	2.264	0.145	6.29281	0.107	147.4	2.6	0.094	2.412
311	Claudia	2.898	0.041	3.22719	0.037	180.2	73.1	0.954	2.005
315	Constantia	2.242	0.12	2.42378	0.044	331.4	178.4	0.149	2.501
316	Goberta	3.175	0.134	2.34017	0.024	96.4	149.1	0.906	1.442
320	Katharina	3.013	0.074	9.34721	0.175	10.9	226.6	0.949	1.787
321	Florentina	2.886	0.046	2.59711	0.038	120.5	14.3	0.93	2
323	Brucia	2.383	0.195	24.23379	0.438	32.5	101.5	0.137	2.068
326	Tamara	2.318	0.165	23.72419	0.412	256.5	28	0.177	2.211
339	Dorothea	3.012	0.067	9.94162	0.17	318.6	181.3	0.97	1.808

341 California	2.199	0.129	5.67202	0.092	318.2	25.5	0.053	2.489
352 Gisela	2.194	0.13	3.37728	0.07	54.4	247.6	0.054	2.499
364 Isara	2.221	0.154	6.00281	0.096	79	111.5	0.028	2.426
367 Amicitia	2.219	0.147	2.94411	0.04	144	86	0.061	2.456
376 Geometria	2.289	0.168	5.4295	0.111	241.7	298.6	0.066	2.337
379 Huenna	3.137	0.148	1.66536	0.032	348.1	216.8	0.832	1.44
383 Janina	3.135	0.148	2.65587	0.025	61.7	90.8	0.828	1.439
391 Ingeborg	2.32	0.255	23.17737	0.42	3.3	220	-0.039	1.984
422 Berolina	2.228	0.161	5.00321	0.082	346.4	1.6	0.027	2.406
428 Monachia	2.308	0.153	6.20334	0.104	44.7	9.3	0.126	2.357
431 Nephele	3.129	0.151	1.82699	0.013	321.5	148.3	0.815	1.437
440 Theodora	2.21	0.151	1.5976	0.039	124.3	278.9	0.039	2.451
450 Brigitta	3.015	0.065	10.16556	0.174	13.1	8.4	0.979	1.811
452 Hamiltonia	2.865	0.063	3.22435	0.036	168.3	90.8	0.862	1.973
453 Tea	2.183	0.136	5.55879	0.091	199.8	5	0.01	2.477
461 Saskia	3.112	0.159	1.43788	0.025	114.3	209.4	0.774	1.431
462 Eriphyla	2.874	0.05	3.19195	0.036	345.2	110.5	0.908	2.001
468 Lina	3.14	0.153	0.4442	0.022	348.3	298.9	0.819	1.42
472 Roma	2.542	0.103	15.80873	0.264	98	130.7	0.405	2.19
492 Gismonda	3.112	0.147	1.63363	0.023	329.2	0.7	0.813	1.469
496 Gryphia	2.199	0.128	3.78827	0.077	124.4	212.8	0.064	2.499
502 Sigune	2.384	0.173	24.99057	0.42	151.2	132.5	0.206	2.116
513 Centesima	3.014	0.056	9.7169	0.171	68	191.7	1.008	1.841
515 Athalia	3.12	0.154	2.02323	0.019	66.3	151.1	0.8	1.44
520 Franziska	3.006	0.082	10.98307	0.184	70	28.7	0.915	1.77
525 Adelaide	2.245	0.143	5.99052	0.117	122.3	208.6	0.074	2.426
526 Jena	3.121	0.162	2.1693	0.026	147.9	173	0.772	1.412
529 Preziosa	3.017	0.065	11.00623	0.174	55.9	62.5	0.98	1.809
534 Nassovia	2.884	0.053	3.27484	0.037	107.4	93.3	0.91	1.982
540 Rosamund	2.219	0.145	5.5739	0.109	168.2	207.9	0.038	2.441
553 Kundry	2.231	0.123	5.39162	0.083	91.1	72.9	0.116	2.49
555 Norma	3.169	0.189	2.64073	0.031	134.9	156.3	0.724	1.277
562 Salome	3.019	0.066	11.12101	0.177	310.8	68.1	0.98	1.805
573 Recha	3.014	0.073	9.84037	0.18	13	337.2	0.951	1.789
575 Renate	2.555	0.077	15.05039	0.262	320	346.5	0.489	2.248
579 Sidonia	3.013	0.062	11.00326	0.171	282.2	81.1	0.986	1.821
587 Hypsipyle	2.335	0.174	24.97898	0.427	154.3	322	0.162	2.173
590 Tomyris	3.001	0.078	11.15662	0.174	111.2	107.8	0.925	1.788
608 Adolfine	3.024	0.075	9.38495	0.186	1.8	292.2	0.955	1.775
616 Elly	2.553	0.096	14.98554	0.264	134.1	353.6	0.436	2.201
621 Werdandi	3.118	0.154	2.32152	0.026	113.7	39.4	0.796	1.442
633 Zelima	3.017	0.059	10.91094	0.177	316.9	152.9	0.999	1.827
637 Chrysothe	3.165	0.17	0.29074	0.025	172.4	289.6	0.783	1.338
639 Latona	3.016	0.068	8.56967	0.171	333.9	279.9	0.971	1.801
641 Agnes	2.22	0.131	1.71676	0.021	80.5	22.9	0.099	2.492
651 Antikleia	3.024	0.065	10.76947	0.177	41.9	33.1	0.986	1.801

652 Jubilatrix	2.555	0.072	15.77965	0.258	22.9	86.1	0.51	2.267
653 Berenike	3.014	0.079	11.29008	0.182	183.2	136.9	0.933	1.772
654 Zelinda	2.297	0.192	18.12805	0.332	133	275	0.077	2.134
656 Beagle	3.16	0.158	0.50817	0.025	161.5	257	0.82	1.382
658 Asteria	2.854	0.045	1.51477	0.037	85.6	320.1	0.902	2.034
660 Crescentia	2.535	0.088	15.23935	0.262	224.4	160.1	0.436	2.234
661 Cloelia	3.016	0.071	9.2591	0.173	166.6	330.6	0.96	1.792
669 Kypria	3.012	0.074	10.7778	0.185	257.6	177.4	0.946	1.79
685 Hermia	2.236	0.146	3.64478	0.079	306.9	238.1	0.071	2.435
695 Bella	2.538	0.088	13.86899	0.263	5	275.1	0.438	2.231
700 Auravictrix	2.229	0.148	6.79025	0.111	181.2	99.6	0.043	2.426
703 Noemi	2.175	0.117	2.45315	0.053	54	217.7	0.067	2.544
710 Gertrud	3.135	0.154	1.74782	0.021	227.7	183.5	0.81	1.421
711 Marmulla	2.237	0.152	6.09407	0.109	283.7	351.1	0.047	2.415
714 Ulula	2.535	0.091	14.27974	0.264	124.8	235.6	0.426	2.225
720 Bohlinia	2.887	0.051	2.36664	0.036	172.9	6.8	0.917	1.983
727 Nipponia	2.568	0.087	15.0327	0.25	76.3	136.2	0.494	2.227
736 Harvard	2.202	0.106	4.3724	0.073	337.6	144.2	0.118	2.546
742 Edisona	3.013	0.072	11.21177	0.181	340.9	61.5	0.953	1.793
751 Faina	2.552	0.114	15.60805	0.256	39.1	79.4	0.402	2.171
761 Brendelia	2.863	0.047	2.16915	0.037	285	353.4	0.907	2.021
763 Cupido	2.241	0.13	4.08296	0.084	33.2	284.9	0.112	2.467
766 Moguntia	3.021	0.081	10.10283	0.18	101.3	1.9	0.933	1.757
767 Bondia	3.117	0.15	2.42135	0.025	336.7	63.2	0.808	1.454
770 Bali	2.221	0.157	4.39143	0.067	81.5	38	0.031	2.425
775 Lumiere	3.012	0.086	9.28739	0.182	129.6	296	0.909	1.753
782 Montefiore	2.18	0.104	5.26081	0.083	152.4	83.8	0.089	2.558
787 Moskva	2.54	0.063	14.8439	0.264	288.6	187.9	0.502	2.289
798 Ruth	3.015	0.06	9.22919	0.171	223.7	220.8	0.996	1.828
800 Kressman	2.193	0.144	4.26818	0.076	303.7	316.6	0.018	2.466
802 Epyaxa	2.196	0.138	5.2074	0.087	135	2.2	0.03	2.472
807 Ceraskia	3.019	0.081	11.30494	0.18	136.4	136.6	0.932	1.762
809 Lundia	2.283	0.143	7.14404	0.121	353.6	160.7	0.113	2.396
810 Atossa	2.179	0.124	2.60873	0.047	355.3	166	0.061	2.53
811 Nauheima	2.897	0.062	3.13222	0.04	277.8	148.3	0.892	1.946
819 Barnardian	2.197	0.112	4.90112	0.087	252.8	326.3	0.092	2.53
823 Sisigambis	2.221	0.136	3.64507	0.078	127.3	253.4	0.077	2.469
825 Tanina	2.226	0.111	3.39981	0.051	189	109.2	0.146	2.528
831 Stateira	2.212	0.136	4.83436	0.091	65.4	184.4	0.056	2.468
832 Karin	2.864	0.044	1.00009	0.037	18.7	266.7	0.916	2.027
833 Monica	3.01	0.083	9.79663	0.179	36.9	346.8	0.917	1.763
836 Jole	2.191	0.142	4.83714	0.093	39.2	205.2	0.008	2.461
841 Arabella	2.255	0.109	3.79569	0.065	131.7	344.5	0.184	2.51
846 Lipperta	3.129	0.144	0.26516	0.027	36.8	278.4	0.836	1.459
848 Inna	3.106	0.138	1.04316	0.034	322.5	247.3	0.835	1.504
851 Zeissia	2.228	0.14	2.3883	0.04	150.7	158.9	0.086	2.463

852 Wladilena	2.363	0.196	23.03898	0.427	307.9	27.5	0.127	2.087
871 Amneris	2.222	0.147	4.24766	0.076	202.3	166.4	0.053	2.444
875 Nymphe	2.555	0.083	14.60103	0.263	302.2	199.8	0.47	2.23
876 Scott	3.011	0.073	11.35095	0.185	353.5	156.3	0.947	1.793
879 Ricarda	2.53	0.093	13.6828	0.26	24.5	270.4	0.422	2.23
883 Matterania	2.238	0.144	4.72188	0.095	323.3	280.3	0.072	2.433
890 Waltraut	3.023	0.077	10.86956	0.183	218.9	166.6	0.947	1.768
897 Lysistrata	2.544	0.075	14.31241	0.261	245.4	258.2	0.482	2.263
901 Brunsia	2.224	0.164	3.44693	0.077	331.5	262.7	0.017	2.403
905 Universita	2.216	0.122	5.32738	0.082	38.6	34	0.101	2.503
913 Otila	2.197	0.135	5.80681	0.091	262.3	98.4	0.037	2.477
914 Palisana	2.454	0.181	25.25333	0.456	284.6	251.2	0.214	2.019
915 Cosette	2.228	0.134	5.55312	0.094	70.9	1.5	0.083	2.464
929 Algunde	2.239	0.118	3.90743	0.08	227.3	233.5	0.139	2.497
935 Clivia	2.219	0.136	4.02951	0.072	65.8	335.9	0.076	2.473
936 Kunigunde	3.136	0.151	2.37134	0.029	302.5	35.9	0.819	1.429
937 Bethgea	2.231	0.166	3.69377	0.082	307.9	245	0.02	2.394
938 Chlosinde	3.161	0.149	2.66508	0.027	334.6	134.2	0.847	1.407
939 Isberga	2.247	0.129	2.59234	0.05	329.8	312.5	0.132	2.473
946 Poesia	3.122	0.149	1.43533	0.013	119.9	13	0.817	1.453
950 Ahrensa	2.371	0.172	23.48858	0.404	178.1	184.5	0.212	2.128
951 Gaspra	2.21	0.143	4.09747	0.084	40.4	254	0.042	2.458
954 Li	3.139	0.147	1.15938	0.024	297.5	224.7	0.836	1.439
956 Elisa	2.298	0.157	5.95278	0.115	310.6	201	0.1	2.353
960 Birgit	2.248	0.115	3.02143	0.068	335.3	249.7	0.162	2.501
962 Aslog	2.906	0.062	2.59717	0.035	9.9	171.5	0.9	1.935
963 Iduberga	2.248	0.152	7.98496	0.128	87.8	60.7	0.049	2.398
967 Helionape	2.226	0.118	5.41689	0.083	303.2	83.7	0.121	2.505
975 Perseveran	2.834	0.049	2.56616	0.038	141.4	13.1	0.874	2.045
981 Martina	3.1	0.165	2.06963	0.029	337	12.5	0.75	1.426
988 Appella	3.153	0.188	1.58078	0.024	15.2	352.3	0.714	1.298
991 McDonalda	3.145	0.137	2.09085	0.024	298.4	31.4	0.874	1.465
993 Moultona	2.861	0.046	1.76889	0.036	117.8	218.2	0.906	2.023
994 Otthild	2.53	0.061	15.37415	0.261	351.2	359.6	0.498	2.304
996 Hilaritas	3.093	0.161	0.66677	0.028	143	301.4	0.756	1.444
1003 Lilofee	3.15	0.151	1.83452	0.022	108	183.7	0.833	1.414
1016 Anitra	2.219	0.137	6.0457	0.103	85.9	2.2	0.061	2.462
1026 Ingrid	2.25	0.134	5.40195	0.082	308.7	108.1	0.116	2.452
1027 Aesculapia	3.161	0.159	1.25945	0.025	169	334.3	0.816	1.378
1029 La Plata	2.89	0.063	2.43846	0.039	178.6	2.1	0.882	1.948
1033 Simona	3.003	0.086	10.65418	0.188	57	195.3	0.898	1.762
1037 Davidweill	2.255	0.151	5.89328	0.11	20.2	207.2	0.071	2.404
1047 Geisha	2.241	0.157	5.66387	0.087	30.5	79.9	0.051	2.406
1052 Belgica	2.236	0.125	4.69424	0.072	57	105.4	0.122	2.485
1055 Tynka	2.198	0.145	5.27	0.089	319.6	154.9	0.018	2.455
1056 Azalea	2.23	0.128	5.42678	0.084	307.4	108.1	0.105	2.481

1058 Grubba	2.197	0.127	3.68672	0.078	307.2	224.9	0.06	2.501
1060 Magnolia	2.237	0.149	5.91617	0.12	293.4	224.3	0.047	2.415
1061 Paeonia	3.121	0.182	2.49148	0.024	36.9	88.1	0.709	1.354
1073 Gellivara	3.178	0.164	1.6197	0.026	320.7	354.2	0.81	1.343
1074 Beljawskya	3.155	0.151	0.82232	0.02	65.3	321.8	0.837	1.409
1075 Helina	3.014	0.069	11.51421	0.182	342.3	101.3	0.965	1.803
1078 Mentha	2.27	0.187	7.37355	0.118	141.1	93.3	-0.004	2.307
1079 Mimosa	2.874	0.047	1.18691	0.036	119.4	304	0.917	2.009
1082 Pirola	3.128	0.144	1.84789	0.025	323.7	192.8	0.837	1.461
1087 Arabis	3.015	0.071	10.06036	0.17	78.2	24.2	0.963	1.795
1088 Mitaka	2.202	0.154	7.65258	0.122	32.2	55.2	-0.024	2.414
1089 Tama	2.214	0.139	3.73009	0.053	87.4	72	0.069	2.474
1090 Sumida	2.36	0.196	21.52181	0.36	135.1	152.3	0.118	2.061
1100 Arnica	2.898	0.047	1.03746	0.037	292.6	290	0.938	1.987
1105 Fragaria	3.013	0.069	10.96578	0.171	323.8	119.4	0.965	1.802
1108 Demeter	2.428	0.163	24.92029	0.468	297	231.3	0.235	2.096
1112 Polonia	3.021	0.065	8.99225	0.177	39	300.8	0.985	1.807
1117 Reginita	2.248	0.161	4.3387	0.074	285.5	158.8	0.054	2.395
1120 Cannonia	2.216	0.121	4.04538	0.073	35.1	167.4	0.106	2.507
1123 Shapleya	2.225	0.14	6.41882	0.101	57.3	82.7	0.063	2.453
1129 Neujmina	3.022	0.059	8.61607	0.17	68.9	271.4	1.005	1.822
1130 Skuld	2.229	0.146	2.16242	0.05	326	226.3	0.071	2.448
1133 Lugduna	2.186	0.138	5.37617	0.083	20.2	59.1	0.016	2.478
1142 Aetolia	3.179	0.12	2.09698	0.026	223.8	172.8	0.954	1.482
1148 Rarahu	3.016	0.089	10.84498	0.175	304.8	151.4	0.905	1.74
1150 Achaia	2.191	0.147	2.38271	0.052	350.3	216.6	0.02	2.47
1153 Wallenber	2.196	0.109	3.33341	0.066	295.6	273.1	0.108	2.549
1158 Luda	2.564	0.087	14.8798	0.266	72.2	340.7	0.467	2.209
1160 Illyria	2.56	0.078	15.00308	0.256	20.7	0.1	0.503	2.25
1164 Kobolda	2.306	0.192	25.17298	0.426	148.4	160.4	0.095	2.163
1170 Siva	2.326	0.212	22.20613	0.409	53.9	354.8	0.071	2.087
1171 Rusthaweli	3.167	0.176	3.04739	0.039	56.2	138.3	0.765	1.318
1174 Marmara	3.022	0.079	10.1002	0.178	346.7	355.1	0.94	1.763
1185 Nikko	2.237	0.124	5.70294	0.088	97.3	71.9	0.12	2.482
1186 Turnera	3.021	0.071	10.7515	0.179	326.3	38.6	0.963	1.787
1188 Gothlandia	2.191	0.14	4.82647	0.079	28.8	358.1	0.022	2.474
1192 Prisma	2.365	0.224	23.85569	0.422	146.4	5	0.064	2.011
1199 Geldonia	3.019	0.063	8.77416	0.17	174.3	241.3	0.989	1.814
1207 Ostenia	3.021	0.07	10.37326	0.179	85.9	14.1	0.969	1.792
1210 Morosovia	3.011	0.069	11.26192	0.177	240.2	108.9	0.962	1.804
1214 Richilde	2.711	0.091	9.84984	0.189	303.4	284.6	0.631	2.05
1215 Boyer	2.579	0.094	15.89844	0.265	54.3	125.9	0.47	2.183
1216 Askania	2.232	0.16	7.59444	0.128	245.4	127.3	0.007	2.386
1219 Britta	2.213	0.138	4.42017	0.068	88.7	36.7	0.066	2.474
1220 Crocus	3.005	0.072	11.35919	0.178	112.9	116.1	0.945	1.8
1223 Neckar	2.869	0.043	2.55502	0.037	90.2	14.7	0.923	2.025

1225 Ariane	2.233	0.116	3.0805	0.049	129.9	0.9	0.147	2.515
1229 Tilia	3.215	0.135	0.98288	0.032	346.4	246.6	0.936	1.396
1234 Elyna	3.013	0.055	8.52827	0.168	50.7	302.1	1.01	1.844
1245 Calvinia	2.893	0.043	2.88282	0.042	349.7	175.3	0.943	2.003
1247 Memoria	3.138	0.16	1.76782	0.031	283.2	205.8	0.795	1.402
1249 Rutherford	2.224	0.128	4.87071	0.099	134	256.9	0.088	2.476
1253 Frisia	3.169	0.169	1.35218	0.024	40.3	341.9	0.789	1.335
1259 Ogyalla	3.1	0.166	2.38758	0.024	216	55.2	0.745	1.425
1270 Datura	2.235	0.155	5.98686	0.097	3.2	100.3	0.041	2.41
1274 Delportia	2.229	0.148	4.40424	0.084	192.5	317.3	0.056	2.434
1286 Banachiew	3.023	0.074	9.73659	0.178	278.2	207.7	0.956	1.777
1287 Lorcia	3.012	0.075	9.8216	0.18	140.4	209.9	0.943	1.785
1289 Kutaissi	2.86	0.052	1.60588	0.037	270.3	226.8	0.89	2.01
1291 Phryne	3.012	0.061	9.09953	0.171	309.5	222.3	0.99	1.827
1297 Quadea	3.021	0.058	9.00369	0.176	94.4	294.9	1.005	1.826
1302 Werra	3.122	0.162	2.59427	0.024	94.7	83	0.775	1.413
1307 Cimmeria	2.251	0.118	3.94427	0.082	103.7	236	0.154	2.49
1310 Villigera	2.393	0.236	21.06541	0.424	318.2	225.2 4	0.023	1.939
1318 Nerina	2.308	0.217	24.66142	0.422	190.8	354.4	0.033	2.101
1322 Coppernic	2.422	0.238	23.31458	0.423	270.7	249.6	0.057	1.91
1324 Knysna	2.185	0.136	4.51544	0.086	249.3	298.2	0.017	2.481
1331 Solvejg	3.104	0.171	3.08751	0.036	291.1	137.2	0.729	1.403
1335 Demoulina	2.241	0.116	2.54154	0.049	23.1	186.6	0.155	2.509
1336 Zeelandia	2.851	0.047	3.19592	0.036	278	98.4	0.893	2.031
1338 Duponta	2.264	0.126	4.82162	0.091	97.7	318.8	0.143	2.46
1339 Desagneau	3.021	0.063	8.67773	0.171	126.3	289.9	0.991	1.812
1340 Yvette	3.183	0.167	0.42545	0.028	206.8	295.1	0.805	1.33
1342 Brabantia	2.289	0.179	20.94754	0.382	180.5	308.3	0.128	2.202
1344 Caubeta	2.248	0.163	5.66018	0.089	181.3	58.3	0.043	2.385
1350 Rosselia	2.858	0.051	2.93271	0.039	23.9	160.4	0.889	2.013
1353 Maartje	3.012	0.073	9.17873	0.172	280.9	218.3	0.952	1.792
1363 Herberta	2.903	0.047	1.09227	0.034	291.5	247	0.943	1.982
1364 Safara	3.012	0.073	11.50047	0.183	253.3	60.8	0.949	1.79
1365 Henyey	2.249	0.141	5.07067	0.104	213.6	258.9	0.09	2.434
1367 Nongoma	2.344	0.152	22.45392	0.396	258.2	271.2	0.241	2.203
1370 Hella	2.251	0.128	4.8065	0.091	297.9	299.3	0.127	2.462
1376 Michelle	2.228	0.165	3.54532	0.063	313.2	177.2	0.025	2.403
1377 Roberbaux	2.26	0.125	6.01798	0.118	197	226.3	0.129	2.455
1379 Lomonoso	2.528	0.129	15.57892	0.27	180.2	172.1	0.309	2.128
1382 Gerti	2.22	0.147	1.5676	0.029	218.2	325.9	0.063	2.456
1387 Kama	2.258	0.155	5.52898	0.109	325.1	210.3	0.067	2.394
1388 Aphrodite	3.019	0.071	11.18666	0.182	284.6	50.1	0.961	1.79
1389 Onnie	2.866	0.043	2.03479	0.037	167.3	206.8	0.919	2.027
1396 Outeniqua	2.248	0.153	4.4996	0.081	246.9	349.4	0.07	2.411
1399 Teneriffa	2.216	0.136	6.50798	0.118	45.6	166.5	0.046	2.455
1405 Sibelius	2.252	0.139	7.03293	0.133	70.8	308.2	0.082	2.422

1410 Margret	3.02	0.073	10.3473	0.177	54.1	177.4	0.958	1.784
1412 Lagrula	2.215	0.138	4.71906	0.071	101.4	64.8	0.066	2.47
1413 Roucarie	3.022	0.079	10.20022	0.178	143.1	186.6	0.939	1.762
1415 Malautra	2.224	0.125	3.43151	0.064	188.5	317.4	0.11	2.496
1416 Renauxa	3.018	0.079	10.05337	0.184	75.3	346.8	0.936	1.768
1418 Fayeta	2.242	0.147	7.19712	0.124	313.7	350.3	0.054	2.414
1419 Danzig	2.293	0.165	5.71953	0.114	100.5	218.1	0.076	2.34
1422 Stromgreni	2.247	0.13	2.67349	0.055	23.2	213	0.13	2.471
1423 Jose	2.86	0.044	2.91235	0.037	25.9	39.7	0.91	2.03
1434 Margot	3.018	0.059	10.81956	0.179	259.9	158.9	0.999	1.827
1440 Rostia	3.153	0.159	2.29007	0.032	44.1	12.7	0.81	1.386
1442 Corvina	2.875	0.045	1.24615	0.038	330.7	248.1	0.923	2.014
1445 Konkolya	3.114	0.149	2.284	0.021	356	79.7	0.808	1.461
1446 Sillanpaa	2.246	0.136	5.26238	0.086	193.7	9.5	0.105	2.45
1449 Virtanen	2.223	0.151	6.63628	0.11	218.7	116	0.029	2.424
1451 Grano	2.203	0.146	5.10592	0.095	200.6	180.8	0.019	2.448
1455 Mitchella	2.247	0.147	7.75234	0.132	204.5	132.6	0.056	2.406
1462 Zamenhof	3.152	0.138	0.97457	0.025	203.8	323	0.874	1.452
1464 Armisticia	3.002	0.075	11.554	0.183	160.4	85.9	0.934	1.795
1472 Muonio	2.234	0.155	4.56939	0.069	11.9	40.9	0.054	2.421
1476 Cox	2.281	0.144	6.33258	0.113	313.7	324.2	0.114	2.4
1480 Aunus	2.202	0.165	4.863	0.075	135.8	61.1	-0.015	2.415
1482 Sebastiana	2.872	0.049	2.9745	0.035	236.4	57.1	0.909	2.005
1485 Isa	3.026	0.08	8.93898	0.175	326.6	295.3	0.942	1.756
1487 Boda	3.143	0.154	2.47148	0.023	195.2	97.8	0.817	1.414
1492 Oppolzer	2.173	0.156	6.05272	0.106	188.7	142.5	-0.063	2.422
1494 Savo	2.19	0.1	2.45129	0.052	43.5	202.1	0.122	2.574
1496 Turku	2.206	0.125	2.50479	0.051	277.3	283.7	0.09	2.512
1497 Tampere	2.895	0.057	1.06377	0.038	301.4	288.3	0.906	1.962
1500 Jyvaskyla	2.243	0.17	7.44653	0.125	52.6	13.4	0	2.36
1513 Matra	2.193	0.159	3.97302	0.067	157.5	145.1	-0.013	2.435
1514 Ricouxa	2.241	0.152	4.52996	0.075	321	155.9	0.065	2.421
1518 Rovaniemi	2.226	0.154	6.71859	0.112	85.6	21.7	0.024	2.414
1523 Pieksamaki	2.242	0.147	5.14793	0.096	153.9	319.5	0.071	2.425
1526 Mikkeli	2.315	0.172	6.21614	0.12	63.1	329.4	0.081	2.304
1527 Malmquist	2.227	0.144	5.19801	0.087	314.3	9.9	0.063	2.445
1530 Rantasepp	2.249	0.155	4.41945	0.094	20.4	281.8	0.06	2.401
1532 Inari	3.005	0.063	8.7964	0.166	131.4	325.9	0.977	1.827
1533 Saimaa	3.013	0.074	10.70339	0.177	170.5	162.9	0.949	1.789
1536 Pielinen	2.204	0.151	1.52547	0.035	16	213	0.034	2.453
1539 Borrelly	3.147	0.151	1.71775	0.021	33.3	188.6	0.831	1.417
1549 Mikko	2.231	0.118	5.54706	0.086	114.2	87.4	0.128	2.502
1552 Bessel	3.01	0.07	9.86029	0.173	62.6	3.5	0.958	1.801
1557 Roehla	3.01	0.07	10.3151	0.183	349.3	349.7	0.959	1.804
1562 Gondolatsoe	2.226	0.116	4.88315	0.082	187.8	136.3	0.126	2.509
1563 Noel	2.191	0.148	5.9892	0.096	160.2	53.1	-0.002	2.453

1565 Lemaitre	2.393	0.242	21.41571	0.435	18.6	267	0.013	1.933
1568 Aisleen	2.352	0.197	24.88923	0.43	0.6	143.1	0.114	2.098
1570 Brunonia	2.844	0.043	1.6556	0.036	94.9	223.7	0.898	2.05
1573 Vaisala	2.37	0.212	24.58003	0.422	14.7	204.9	0.085	2.032
1575 Winifred	2.375	0.196	24.77852	0.43	201	209.9	0.135	2.074
1576 Fabiola	3.135	0.152	0.93705	0.022	58.7	234.4	0.817	1.429
1577 Reiss	2.23	0.141	4.35529	0.071	46.1	131.6	0.08	2.455
1581 Abanderad	3.164	0.153	2.53619	0.025	188.9	111.9	0.837	1.392
1584 Fuji	2.376	0.195	26.6803	0.458	133.8	302.8	0.118	2.079
1590 Tsiolkovsk	2.23	0.134	4.3482	0.088	258	228.7	0.089	2.465
1591 Baize	2.393	0.186	24.77415	0.413	335.5	173 3	0.215	2.062
1601 Patry	2.234	0.118	4.94359	0.074	246.3	75.2	0.136	2.503
1602 Indiana	2.245	0.154	4.16344	0.061	150.3	74	0.071	2.418
1604 Tombaugh	3.024	0.066	9.40488	0.18	328.2	305.9	0.982	1.799
1605 Milankovit	3.014	0.077	10.56534	0.182	120.5	181.3	0.939	1.779
1608 Munoz	2.214	0.119	3.94862	0.066	302.6	348	0.109	2.515
1615 Bardwell	3.113	0.158	1.67708	0.025	49.7	197.8	0.779	1.434
1618 Dawn	2.869	0.046	3.22588	0.036	223.9	107	0.915	2.018
1619 Ueta	2.241	0.151	6.21591	0.095	45.3	61	0.061	2.417
1621 Druzhba	2.23	0.123	3.16833	0.062	83.1	192.1	0.122	2.496
1622 Chacornac	2.234	0.153	6.46684	0.113	238.6	357.1	0.038	2.411
1623 Vivian	3.133	0.155	2.49054	0.025	86.3	133.5	0.806	1.422
1624 Rabe	3.18	0.143	1.98281	0.021	159.2	170.8	0.879	1.407
1626 Sadeya	2.364	0.233	25.31097	0.446	77.9	283.7	0.016	1.998
1631 Kopff	2.235	0.152	7.49588	0.125	331.1	13.2	0.033	2.406
1633 Chimay	3.169	0.185	2.67461	0.029	178.2	124.9	0.734	1.288
1634 Ndola	2.246	0.136	7.59833	0.12	261.7	92.3	0.089	2.44
1635 Bohrmann	2.855	0.045	1.80951	0.037	277.8	217.8	0.903	2.033
1636 Porter	2.235	0.12	4.43229	0.082	70.5	176.5	0.129	2.495
1641 Tana	3.019	0.073	9.3455	0.174	315.6	326.7	0.959	1.785
1649 Fabre	3.021	0.081	10.81762	0.176	168.3	150.8	0.933	1.757
1651 Behrens	2.18	0.135	5.07001	0.098	156.1	193.5	0.008	2.478
1652 Herge	2.251	0.144	3.19513	0.069	244.5	252.1	0.098	2.432
1654 Bojeva	3.017	0.052	10.45753	0.176	354.6	19.8	1.023	1.851
1657 Roemera	2.349	0.189	23.41922	0.413	155.9	102.4	0.138	2.114
1660 Wood	2.395	0.212	20.54997	0.411	135.1	216.7	0.116	2.004
1661 Granule	2.184	0.12	3.03248	0.064	197.4	258.1	0.065	2.529
1663 Van den	2.24	0.131	5.36277	0.084	5.8	84.9	0.109	2.466
1666 Van Gent	2.185	0.122	2.68527	0.059	352.9	258.5	0.063	2.524
1667 Pels	2.19	0.13	4.61664	0.07	252.7	83.5	0.049	2.502
1669 Dagmar	3.14	0.15	0.94905	0.026	194.6	319.3	0.828	1.429
1674 Groenevel	3.187	0.139	2.67881	0.025	106.3	93.9	0.898	1.412
1675 Simonida	2.233	0.149	6.80303	0.113	100.6	24.6	0.045	2.421
1677 Tycho	2.532	0.063	14.82029	0.263	259.6	335.6	0.493	2.296
1682 Karel	2.239	0.14	4.03188	0.075	333.7	315.8	0.091	2.449
1684 Iguassu	3.092	0.149	3.66021	0.044	240	112	0.79	1.482

1686 De Sitter	3.163	0.148	0.62697	0.025	287.7	306.2	0.854	1.41
1687 Glarona	3.158	0.151	2.64173	0.025	60.9	91.6	0.839	1.406
1691 Oort	3.165	0.142	1.05404	0.025	48.7	233.4	0.874	1.426
1696 Nurmela	2.262	0.145	6.04318	0.1	175.6	15.2	0.094	2.416
1698 Christophe	3.155	0.15	1.52218	0.029	158.9	341.9	0.839	1.412
1699 Honkasalo	2.211	0.112	1.97285	0.046	319	266	0.127	2.535
1703 Barry	2.215	0.117	4.5183	0.071	319.7	118.6	0.114	2.518
1704 Wachmann	2.223	0.136	0.96795	0.031	170.7	256.5	0.091	2.479
1707 Chantal	2.219	0.163	4.04239	0.07	66.6	354.2	0.016	2.413
1711 Sandrine	3.015	0.069	11.09417	0.18	32.3	139.2	0.965	1.801
1713 Bancilhon	2.228	0.136	3.74963	0.055	309.2	57.8	0.092	2.469
1717 Arlon	2.195	0.18	6.19631	0.113	113.7	335.7	-0.084	2.366
1720 Niels	2.188	0.13	0.72783	0.015	101.3	180.2	0.062	2.516
1723 Klemola	3.013	0.077	10.92293	0.179	168.4	155.4	0.939	1.778
1725 CrAO	2.903	0.057	3.16647	0.037	334.2	131.4	0.913	1.953
1729 Beryl	2.23	0.092	2.44739	0.038	240.6	353.5	0.197	2.57
1732 Heike	3.012	0.076	10.78776	0.177	0.7	161.5	0.943	1.784
1733 Silke	2.193	0.136	4.4276	0.081	124.4	168.7	0.034	2.482

ANEXO II

REDES NEURONALES

Prefiero los errores del entusiasmo a la indiferencia de la sabiduría.

Anatole France

CCLXVIII

ANEXO II

11.1. REDES NEURONALES

11.1.1. INTRODUCCIÓN

Como su propio nombre lo indica, la idea original de este tipo de computación era modelar el comportamiento de las auténticas redes de neuronas que tiene el cerebro. Los modelos creados hasta ahora son modelos extremadamente simplificados desde el punto de vista neurofisiológico; lo que se busca primordialmente no es ya imitar a las neuronas a las neuronas auténticas, sino lograr una máquina de computación formada por la interconexión de muchos elementos simples de cálculo que posea ciertas propiedades deseables que sí se encuentran en un cerebro real.

Tal vez la propiedad más buscada y a su vez la que diferencia claramente a las redes de neuronas de otro tipo de redes de características similares, como los autómatas finitos, es la capacidad de *aprender de la experiencia* y de *generalizar* a partir de ella. Casi todos los modelos creados de redes de neuronas tienen estas capacidades que imitan el comportamiento de los seres humanos. Es más, esta imitación también incorpora algunos defectos del modo de aprendizaje humano, como es la memorización de los hechos (que impide sacar conclusiones generales) o el olvido. Sin embargo, el paralelismo entre un tipo de aprendizaje y otro terminará ahí, en su comportamiento similar, ya que, como se verá, los mecanismos de aprendizaje que se lleva a cabo en las redes de neuronas artificiales son artificios matemáticos demasiado complejos para que se lleven a cabo en un cerebro.

La otra gran propiedad que comparten estos modelos de computación es que se comportan como una caja negra. Resuelven el problema en curso, pero no es posible explicar fácilmente cómo lo hacen. Existen dos razones básicas para ello. La primera es que aunque los elementos que forman la red son simples, al interconectar varios cientos o miles de ellos entre sí lo que se tiene es un sistema muy complejo de analizar. La segunda es que la <simplicidad> de los elementos no es tal matemáticamente hablando. Normalmente se

emplean elementos con un comportamiento no lineal, término que en matemáticas es sinónimo de complejidad.

11.1.2. UN POCO DE HISTORIA

Aunque por el auge actual del campo no, lo parezca, la idea de imitar el comportamiento de las neuronas no es nada nueva. Ya en 1943 McCulloch y Pitts publicaron el primer tratamiento formal de una neurona artificial [McCulloch, 1943]. Desde ese año hasta el presente el campo de las redes neuronales ha sufrido varios vaivenes.

En 1949, Donald Hebb indica un mecanismo por el cual es posible explicar cómo un cerebro puede aprender de la experiencia [Hebb, 1949]. Lo interesante del mecanismo es que se podía aplicar a algunas redes neuronales artificiales muy simples.

En 1962, Rosenblatt inventa el *Perceptrón*, una red neuronal simple junto con un mecanismo de aprendizaje [Rosenblatt, 1962]. Lo más importante de su invención es que estaba acompañada de una prueba matemática fundamental: si la red neuronal es capaz de resolver un problema utilizando su mecanismo de aprendizaje aprenderá a resolverlo.

En 1969, Minsky y Papert demuestran [Minsky, 1969] que el Perceptrón, el modelo más avanzado hasta entonces, no puede aprender a resolver problemas muy complejos. Los ánimos se enfrían entre la comunidad científica, y el campo deja de tener atractivo durante la década de los años 70.

A pesar de que algunos científicos siguen trabajando en él como Grossberg, Fukushima y Kohonen, no es hasta 1982 que John Hopfield [Hopfield, 1982] vuelve a calentar el ambiente cuando encuentra que las matemáticas de un tipo de red neuronal especial son similares a las matemáticas utilizadas por lo físicos para modelizar unos sistemas magnéticos llamados <cristales de spin>. Esto moviliza a la comunidad de físicos hacia el campo de las redes de neuronas. El entusiasmo crece y llega a su cumbre cuando en 1986 Rumelhart, Hinton y Williams [Rumelhart, 1986] popularizan una red de neuronas, el Perceptrón Multicapa, que junto con un método de aprendizaje especial es capaz de resolver aquellos problemas que el Perceptrón no podía. Es más, se recupera un teorema debido a Kolmogorov que apoya la tesis de

que los Perceptrones Multicapa pueden resolver casi cualquier tarea. El campo se convierte en multidisciplinar, y todo el mundo aporta nuevas ideas: ingenieros, matemáticos, físicos, psicólogos, biólogos, etc.

Actualmente el campo se encuentra todavía en ebullición, celebrándose decenas de congresos sobre el tema. Existen muchas revistas científicas especializadas en el tema, con nombres como *Neural Computation* o *Neural Networks,* y multitud de libros.

11.1.3. LA NEURONA ARTIFICIAL

Nuestro cerebro se compone de cerca de 10^{11} **neuronas** o **células nerviosas** de muchos tipos. En la **figura 1** se puede ver una representación esquemática de una sola neurona. En el cuerpo de la célula o **soma** se localiza el núcleo de la célula. De este soma se extienden fibras en forma de árbol que se llaman dendritas. También del soma sale una única y muy larga fibra llamada **axón** que, finalmente, también se subdivide arborizándose en más fibras. Al final de éstas se encuentran las **uniones sinápticas**, o simplemente **sinapsis**, a otras neuronas. Estas uniones se pueden realizar en las dendritas o en el soma de la neurona receptora. Una neurona suele estar unida a miles de otras mediante estas sinapsis.

La transmisión de una señal de una célula a otra en la sinapsis es un proceso químico muy complejo. El proceso comienza cuando la neurona transmisora libera una serie de sustancias químicas, neurotransmisores, en la sinapsis. Su efecto es incrementar (si la neurona es excitadora) o decrementar (si es inhibidora) el potencial eléctrico del membrana de la neurona receptora. Si el potencial alcanza un nivel determinado o **umbral**, esta neurona envía un pulso o *potencial de acción* de una determinada fuerza y duración a lo largo del axón. Se dice entonces que la neurona se ha **activado.** Este pulso viaja por el axón hasta alcanzar las sinapsis con otras neuronas, donde el proceso se repite otra vez. Después de activarse, la neurona debe esperar un tiempo antes de poder volverlo a hacer. A este tiempo se le llama **período refractario.**

La cantidad en que varía el potencial eléctrico del cuerpo de una neurona receptora después de la liberación de neurotransmisores por parte de la neurona transmisora, depende de muchos factores como la geometría de la

sinapsis, el tipo de neurotransmisor liberado o su número. No todas las sinapsis son iguales, y, por tanto, no todas tienen el mismo efecto sobre la neurona receptora. Es decir, algunas son más <fuertes> que otras, su efecto es más importante (por ejemplo, porque han liberado más neurotransmisores). Otro factor que importa a la hora de activar una neurona es el número de conexiones que recibe. Lo normal es que una neurona reciba a la vez cientos y miles de señales procedentes de otras. Todas estas señales se suman en el soma.

Las teorías actuales sobre el **aprendizaje** nos dicen que su efecto en el cerebro no es mas que una modificación de las conexiones o sinapsis entre neuronas. Esta modificación, producida en el timo por medio de la experiencia del sujeto, se realiza incrementado o decrementado el número de neurotransmisores liberados en determinadas sinapsis. La variación que se produce en el comportamiento de las neuronas puede ser sustancial: una neurona cuyo potencial antes no lograba alcanzar el umbral, ahora sí lo hace y se activa. Y viceversa, otra que se activa, ahora puede que no lo haga. Esta teoría juega un gran papel en el uso de redes de neuronas artificiales, ya que, como se verá, el mecanismo de aprendizaje que se emplea es una imitación de lo que ocurre en realidad.

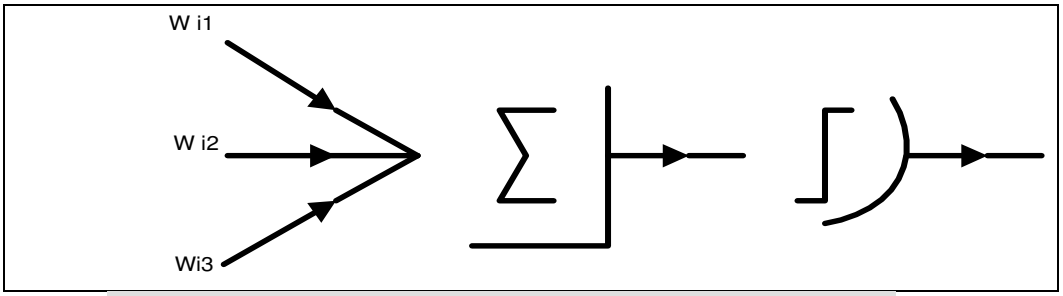


Figura 1. **Esquema de la neurona de McCulloch-Pitts**

En 1943, McCulloch y Pitts proponen (McCuloch, 1943) un modelo muy simple de neurona que se comportaba como una célula binaria de umbral. Específicamente, su modelo de neurona calculaba una suma ponderada de las entradas que recibía de otras neuronas, y producía un 0 o un 1, dependiendo de sí la suma superaba un determinado umbral o no. El modelo obedece a la siguiente ecuación, que está representada en la **figura 2:**

$$x_i = \Theta \left(\sum_{j=1}^{N} w_{ij} x_j + U_i \right)$$

Figura 2. **Ecuación que representa al Esquema de la neurona de McCulloch-Pitts**

donde x_i es o 1 ó 0, y representa es estado de la neurona I como activa o no activa respectivamente. Θ (n) es la **función escalón:**

$$\Theta(n) = \begin{cases} 1 & \text{si } n \geq 0 \\ 0 & \text{si } n < 0 \end{cases}$$

Los **pesos** w_{ij} (1) representa la fuerza de la sinapsis que conecta la neurona j con la i. La conexión puede ser positiva o negativa, emulando las sinapsis excitadoras e inhibitorias, respectivamente. Si no hay sinapsis entre la neurona i y la j, el peso de su conexión es 0. **n** es el número de neuronas que se conectan a la neurona i. El parámetro u_i representa el **umbral** o **sesgo** de la neurona.

En resumen, la ecuación anterior dice que si la suma de las entradas x_j que recibe la neurona j, ponderadas por los pesos w_{ij} supera el umbral, la neurona se activará. En caso contrario, su activación será 0. Repasando la descripción de la transmisión de señales en neuronas que se ha realizado antes se puede comprobar que este comportamiento es una muy buena aproximación de lo que ocurre en las neuronas reales.

A pesar de que el modelo es simple, la neurona de McCulloch-Pitts es un poderoso dispositivo de computación. McCulloch y Pitts probaron que un conjunto de tales neuronas, conectándolas debidamente y eligiendo correctamente los valores de los pesos, sería capaz de computar cualquier artefacto, como una Máquina de Turing.

La neurona de McCulloch-Pitts es el modelo de neurona artificial básico. El modelo más común actualmente es una derivación o generalización de él, que obedece a la ecuación:

$$x_i = f \left(\sum_{N} w_{ij} w_j + U_i \right)$$

j=1

11.1.4. LA RED DE NEURONAS ARTIFICIALES

En términos informáticos, se puede describir el cerebro como un sistema de computación paralela de cerca de 10^{11} procesadores. Con el modelo de neurona que se ha descrito antes, cada uno de estos procesadores tiene un programa muy simple: calcula una suma ponderada de los datos que recibe como entrada procedentes de otros procesadores, y produce un solo número, que es el resultado del cálculo de una función sobre esa suma. Este número, dato de salida, se envía a otros procesadores, que están continuamente realizando el mismo tipo de cálculo. Cada procesador utiliza su propio juego de pesos, y posiblemente su propia función de activación. Se puede pensar que los pesos y las funciones son ***datos locales*** de cada procesador.

Este tipo de computación tiene sus ventajas. La alta conectividad de esta red significa que un error en unos pocos procesadores o en los datos que se manejan probablemente no tenga una consecuencia sustancial en el resultado. La explicación es simple. Retomando la ecuación de una neurona se ve que la suma que se realiza se extiende a las **n** neuronas que se conectan a ella. Si el valor de n es muy grande, y los valores de todos los pesos w_{ij} se mantienen en el mismo rango, entonces si algún x_j tiene un valor erróneo, el valor final de la suma no va a cambiar substancialmente. Hay una manera elegante de llamar a esta propiedad: ***Robustez***. En el cerebro se están muriendo continuamente neuronas y no parece que afecte a su funcionamiento.

Por otro lado, la información se almacena en las sinapsis y en los pesos en forma *distribuida*. Es decir, la información sobre cómo resolver un problema se encuentra diseminada y almacenada entre todos los pesos y para resolverlo todos los pesos cuentan. Evidentemente esto es otra ventaja ya que al estar la información tan distribuida, si algún peso falla, la capacidad para resolver los problemas no habrá mermado más que en un infinitésimo. Pero esto último no es lo que se quiere resaltar aquí. Lo importante es la filosofía de diseño de estos sistemas, ya que en vez de asignar cada unidad de memoria a un problema diferente, como sería lo habitual, aquí cada unidad de memoria está

asignada a todas las tareas y debe guardar información sobre todas ellas a la vez.

Otra propiedad y ventaja muy importante es el ***procesamiento masivamente paralelo*** que se realiza. Las neuronas biológicas tienen un ciclo de reloj típico en milisegundos, mientras que sus equivalentes en silicio, los chips, pueden alcanzar velocidades de nanosegundos. Y sin embargo, el cerebro es capaz de realizar en segundos tareas de visión, control, etc, que ni una supercomputadora CRAY puede realizar en días. Esto es debido a una simple razón: en el cerebro todas las neuronas, billones de ellas, están operando simultáneamente, en paralelo. ¡Imagínese un sistema que tuviera una red de procesadores como el cerebro pero que operase a la velocidad de los chips!

La penúltima ventaja es la capacidad del cerebro de ***aprender de la experiencia***. Al contrario que otros sistemas de computación donde hay que especificar cada detalle del cálculo, en este caso sólo hay que enseñar al sistema cómo resolverlo a través de ejemplos. Piénsese en la diferencia de cómo se enseña a distinguir figuras a un humano, y la cantidad de trabajo que representaría programar una computadora para ello. Mientras que el humano tan sólo necesita que se le muestren ejemplos de figuras, a la computadora hay que programarlo de arriba a abajo empezando por indicarle dónde tiene que buscar en la imagen, y terminando por decirle qué características debe buscar para distinguir una figura de otra. Pero no es sólo el trabajo que se ahorra. También existen problemas donde es muy difícil, si no imposible, programar una solución en una computadora, simplemente porque no se sabe muy bien cómo lo hace un ser humano. El hombre sabe solucionar el problema, pero no sabe explicar cómo lo hace.

La última ventaja está relacionada con la anterior y se refiere a la capacidad de ***generalización*** . Esta capacidad se refiere a la manera en que, a partir de unos pocos ejemplos, se extrapola una solución general. Esta capacidad es lo contrario de la memorización. Se puede decir que es la consecuencia de un buen aprendizaje. Muchos sistemas aprenden, pero pocos generalizan tan bien como el cerebro.

Como se ve, el crear un sistema artificial que opere siguiendo las pautas anteriores sería muy deseable. La simulación de una neurona ya está lista

prácticamente. Realmente lo estaba desde hace 50 años. Pero lo que resta es lo más difícil: diseñar la red de neuronas y utilizarla para resolver problemas.

11.1.5. EL CLASIFICADOR BÁSICO: EL PERCEPTRÓN

El Perceptrón es uno de los primeros modelos de red neuronal artificial. Fue creado por Ronsenblatt en 1962 (Ronsenblatt, 1962) y es un modelo de red destinado a realizar clasificaciones. Las clasificaciones pueden ser unarias o múltiples. Si sólo se dice si los datos pertenecen a una sola clase o no (por ejemplo, es <animal> o no) es unaria. Si se especifica su pertenencia o no a varias clases es múltilpe (es <animal> y <mamífero>, pero no <peludo>, etc).

La ***retina*** está compuesta de células o unidades sensitivas, y representa la entrada al sistema. Sus unidades simplemente recogen, copian, los datos externos. Esta capa está conectada con la capa de unidades asociativas por medio de un conjunto de conexiones fijas. Las unidades asociativas a su vez están conectadas con las unidades de respuesta que van a calcular la salida de sistema. La función de activación que utilizan las neuronas de un Perceptrón suele ser la función de umbral, produciendo salidas binarias -1/+1.

El funcionamiento de Perceptrón es simple. Primero se introduce un patrón de entrada en el sistema. Esto se realiza copiando los datos x_i del patrón en las células de la retina. Los datos pueden ser reales o binarios (0/1 o -1/+1). A continuación se calcula la activación de cada una de las unidades asociativas, a_i , donde:

$$a_i = U\left(\sum_{j=0}^{M} w_{ij}\, x_i\right);\quad U(n) = \begin{cases} +1 \text{ si } n \geq 0 \\ -1 \text{ si } n < 0 \end{cases}$$

Cuando se han calculado se pasa entonces a la capa de respuesta o salida, y se hace lo mismo con sus neuronas.

$$I_i = U\left(\sum_{j=0}^{n} w_{ij}\, a_j\right)$$

La activación y_i de estas últimas serán los datos de salida de la red, y también serán datos binarios. Al ser un problema de clasificación, cada neurona de salida representará una clase. Si el patrón pertenece a una determinada clase, la neurona de salida correspondiente tendrá una activación de valor +1, si no será -1. Es importante observar que siempre hay que calcular las activaciones de las unidades asociativas antes que las unidades de respuesta.

11.1.6. EL ALGORITMO DE APRENDIZAJE Y LA SEPARABILIDAD LINEAL

Sea un Perceptrón sin capa de unidades de asociación, y con todas las neuronas de salida conectadas con todas las neuronas de entrada. Sean también sólo dos datos de entrada por patrón, es decir, que la retina esté compuesta de sólo dos neuronas, y que sólo se necesita una neurona de salida para ver si los datos se clasifican en una sola clase, es decir, el Perceptrón indicará si el patrón en curso pertenece o no a una clase.
La interpretación de lo que está calculando el Perceptrón es muy simple. Si se mira cómo se calcula el valor de la activación e la neurona de salida:

$$Y = U\,(X_1\, W_1 + X_2 + W_0)$$

$$U\,(a) = \begin{cases} +1 \text{ si } a \geq 0 \\ -1 \text{ si } a < 0 \end{cases}$$

se puede ver que es lo mismo que:

$$Y = \begin{cases} +1 \text{ si } X_2 \geq -\frac{W_1}{W_2} \times X_1 - W_0 \\ -1 \text{ si } X_2 \geq -\frac{W_1}{W_2} \times X_1 - W_0 \end{cases}$$

Esta ecuación nos dice que Y será +1 si el valor de X_2 es mayor que el valor de la recta

$$f(x) = -\frac{W_1}{W_2} \times x - W_0$$

Calculada en X_1 , y -1 si ocurre lo contrario. Es decir, la salida Y indica si el patrón representado por los datos X_1 y X_2 está por encima o por debajo de esa recta. Por tanto, lo que hace este Perceptrón es dividir los patrones de entrada en dos clases, los que están por encima de una *recta*, y los que caen por debajo. Si el problema hubiera tenido tres variables de entrada, el Perceptrón habría puesto *planos* para dividir los patrones. En caso de que hubiera más variables, el Perceptrón habría utilizado *hiperplanos*.

Este ejemplo también esclarece la importancia del sesgo, que está aquí representado por el valor de W_o Como se ve, el sesgo corresponde a la variable independiente de la ecuación de la recta o plano o hiperplano que esté utilizando el Perceptrón para separar los patrones. Y esta variable independiente es de vital importancia ya que es la encargada de desplazar la recta o plano, mientras que el resto de las variables, los otros pesos, se encargan de ajustar su inclinación.

Cuando se aplica el algoritmo de aprendizaje a un Perceptrón lo que se está haciendo en realidad es ir moviendo la recta (o plano o hiperplano) hasta que divide a los patrones en dos. Se pueden observar tres momentos en el aprendizaje. Para t = 0, que es el momento inicial, los pesos son aleatorios y la recta divide erróneamente a los patrones, que están representados por dos marcas diferentes, círculos y cruces, dependiendo de la clase a la que pertenezcan. Se puede verificar que para la iteración t = 10, los pesos se

modifican y la recta se mueve, clasificando correctamente más cruces. Finalmente, en la iteración 40, la recta ya divide correctamente a todos los patrones y el problema está resuelto.

Este algoritmo tiene dos problemas. Primero, el problema de la generalización. La recta encontrada funciona muy bien para los patrones de *entrenamiento*, es decir, para aquellos patrones con los que se entrado a la red y se han calculado los pesos. ¿Pero qué pasará con otros patrones nuevos? Pues que algunos <caerán> bien, en el lado correcto de la recta, y otros caerán en el incorrecto. Si el número de patrones que caen bien es muy grande, se puede entender que el Perceptrón ha *generalizado* bien, ya que su solución es muy general, resuelve casi todos los problemas. Si no es así, el Perceptrón debe volverse a entrenar con nuevos patrones, hasta que la recta encontrada separe bien a todos los patrones.

El segundo problema es más grave. El Perceptrón va ha resolver bien todos los problemas que sean *linealmente separables*, es decir, todos los problemas cuyos patrones se puedan separar con una recta (o un plano o un hiperplano si se trabaja con más de dos dimensiones). Éstos son los problemas que se han llamado *computables*.

11.1.7. EL PERCEPTRÓN MULTICAPA

Un Perceptrón sin capa de asociación es una máquina perfecta para resolver problemas lineales. Sin embargo, la capa de asociación le confiere una gran potencia ya que un problema no lineal para las células de salida lo puede convertir en lineal. Hay que hacer una pequeña salvedad: esto será así si las funciones de activación son no lineales, como la función de umbral o la sigmoide. En caso contrario es muy fácil demostrar matemáticamente que una red de N capas con neuronas que utilizan la función lineal es equivalente a un Perceptrón sin capa de asociación.

¿Qué ocurre si se ponen más capas como la de asociación utilizando funciones no lineales? Pues que la situación es cada vez mejor, pudiendo solucionar problemas más difíciles. A medida que se vayan poniendo más capas, los patrones se podrán separar utilizando regiones más complejas que un simple hiperplano. Al poner más capas el Perceptrón cambia de nombre y se le llama *Perceptrón Multicapa*. Este tipo de Perceptrón es potentísimo, pues puede resolver casi cualquier problema no lineal si tiene capas y células suficientes. Pero existe un problema, un gran problema. Como se puede visualizar, no sólo basta con emplear más capas, hay que saber también *cómo conectarlas y qué valores* debe tener en cada conexión. Por desgracia lo primero no será siempre fácil de saber. Lo segundo sí que tiene una respuesta, un algoritmo de aprendizaje que, al igual que hace el que se ha visto en este párrafo con los pesos que hay entre la capa de asociación y la de salida, calcule su valor adecuada y *automáticamente*. Este algoritmo existe y se verá en el párrafo dedicado en exclusiva a los Perceptrones Multicapa.

11.1.8. MEMORIAS ASOCIATIVAS: LA RED DE HOPFIELD

11.1.8.1. LAS MEMORIAS ASOCIATIVAS

Recordar es una operación muy común entre nosotros. Existen muchas maneras de hacerlo, y una de ellas es pensar en una pequeña clave que está *asociada* al dato que se quiere invocar. El cerebro reacciona a esa clave buscando en la memoria aquel dato al que corresponda la clave, y una vez encontrado, la memoria <devuelve> el dato completo. Es la misma operación que se realiza al manejar una base de datos y buscar la información de un ítem especificando sólo una aparte de él. Por ejemplo, si se trata de una base de datos bibliográfica, se puede encontrar toda la información acera de los libros del autor Javier Segovia simplemente especificando una pequeña clave como puede ser el apellido <Segovia>. Este tipo de memorias se llaman *memorias asociativas.*

Las memorias asociativas funcionan por claves. A partir de ellas la memoria <recuerda> el dato. Estas claves pueden ser cualquier cosa. Se pueden ver dos claves que van a permitir a una memoria recordar una determinada imagen. Estas claves son las figuras de la izquierda, y la imagen a recordar la de la derecha. En un caso la clave es un trozo de la imagen, y en otro es la misma imagen pero ruidosa. Este ejemplo también indica una propiedad interesante de algunas memorias asociativas: la clave puede no estar indicada a la hora de memorizar.

11.1.8.2. LA RED DE HOPFIELD

Una red **recurrente** es una red de neuronas artificiales donde las neuronas se pueden conectar formando *ciclos*. Es decir, si la neurona A se conecta a la neurona B, el ciclo se forma si B se conecta a A, directa o indirectamente, a través de otras neuronas. En una red recurrente también está permitido que una neurona esté conectada a sí misma.

Una red de *Hopfield* es una red recurrente que se caracteriza por lo siguiente:

a) **Sus neuronas se conectan todas con todas.**

b) **Las conexiones son simétricas, $W_{ij} = W_{ji}$**

c) **No existen conexiones de neuronas consigo mismas, $W_{ij} = 0$.**

d) **La función de activación es la de umbral en -1, +1 (también se puede emplear el intervalo 0, +1)**

La ecuación de activación es la habitual, que reproducimos aquí:

$$x_i = U\left(\sum_{j=1}^{n} w_{ij}\, x_j\right)$$

Aunque se pueden emplear sesgos, en el resto de este párrafo no lo haremos. Es por eso que en la fórmula anterior el sumatorio comienza en 1 y no en 0.

11.1.8.3. LOS ESTADOS ESTABLES DE UNA RED DE HOPFIELD: LAS MEMORIAS.

Si se observa la red de Hopfield y la ecuación de activación de sus neuronas y se piensa en su funcionamiento, se ve que existe un problema: ¿cómo calcular las activaciones de las neuronas? En una red como el Perceptrón el orden de cálculo era importante. Primero se calculaba la activación de las células de la capa de asociación y después las de salida. El orden dentro de una capa daba igual, ya que ninguna neurona estaba conectada a otra de la misma capa. Sin embargo, en una red de Hopfield todas las neuronas se ven por todas, por lo que el orden puede ser fundamental.

Hay dos modos básicos de especificar el cálculo:

- Modo asíncrono

- Modo síncrono

En el modo *asíncrono* las neuronas calculan su activación sin coordinarse unas con otras. Es decir, cada una calcula su activación <cuando quiere>. En la práctica esto se reduce a ir seleccionando de una en una neuronas al azar e ir actualizando su activación. En esta selección pueden ocurrir repeticiones.
En el modo *síncrono* se supone que todas las neuronas calculan sus activaciones a la vez. Realizar esto en una computadora secuencial requiere mantener dos copias de las activaciones de las neuronas: una antes desactualizarse y otra para almacenar sus nuevos valores.

11.1.9. APRENDIZAJE NO SUPERVISADO Y REDES CONSTRUCTIVAS: ART

11.1.9.1. APRENDIZAJE NO SUPERVISADO Y POR REFUERZO

Hasta ahora se han visto dos modelos de redes de neuronas que utilizan un algoritmo de *aprendizaje supervisado*. Aprendizaje supervisado quiere decir que alguien ha supervisado el aprendizaje y ha dicho *qué había que aprender*. En el caso del Perceptrón, el supervisor calculaba directamente los pesos para que la red hiciera lo debido.
Existen dos maneras más de trabajar con redes de neuronas. La primera es empleando *aprendizaje por refuerzo*. En este tipo de aprendizaje a la red de neurona no se le dice cuánto de bien o de mal está haciendo, sino sólo si lo está haciendo bien o mal. A diferencia del aprendizaje supervisado, también llamado *aprendizaje con profesor,* a este aprendizaje se le llama *aprendizaje con crítica,* ya que critíca pero no explica. Los modelos que trabajan con este tipo de aprendizaje no están muy extendido y por lo tanto no se verá ningún desarrollo de ellos.
La segunda manera es el *aprendizaje no supervisado* o sin profesor. En este caso la red aprende ella sola la tarea a realizar. Esto es conveniente ya que la pregunta siguiente sería ¿y cómo sabe ella lo que tiene que aprender? La explicación es sencilla, ya que las redes donde se utiliza este tipo de

aprendizaje sólo saben hacer una sola cosa, ya que las redes donde se utiliza este tipo de aprendizaje sólo saben hacer una sola cosa, tareas de clasificación: dado un conjunto de patrones o datos, la red los agrupa en clases siguiendo un determinado criterio. Este punto debe quedar muy claro, la red clasifica a su manera y si el usuario no está de acuerdo con ella la única opción que tiene es utilizar otro tipo de red. Las tareas de clasificación abarcan un gran espectro de problemas. Y los problemas a resolver pueden considerarse como problemas de clasificación si se los mira con el prisma adecuado. Por ejemplo, en el último punto del apartado anterior ya se vio cómo una memoria asociativa podía resolver funciones lógicas. En el caso de los clasificadores ocurre otro tanto.

11.1.9.2. REDES CON ARQUITECTURAS CONSTRUCTIVAS

Las redes que se han visto hasta ahora tenían una *arquitectura* o diseño (número de neuronas y forma de conexionarse) establecido *a priori*. Cuando se vean las redes multicapa se observará que ellas también necesitan saber de antemano cuántas células tiene que tener cada una de sus capas.

Éste es un gran problema, especialmente cuando se trabaja con las redes multicapa, ya que mientras que el diseño de la arquitectura es *esencial,* porque si se hace un mal diseño y no se dota a la misma de suficientes neuronas y conexiones no será capaz de resolver nada, no existe ninguna manera de saber cómo hacer este diseño.

Existen diferentes maneras de resolver este problema. Una de ellas es dotar a la red de más neuronas y conexiones de las necesarias y luego aplicar un algoritmo que las vaya eliminando, por desuso por ejemplo. Otra manera es partir de cero e ir añadiendo neuronas y conexiones a medida que la red las va necesitando. A esta última manera se la llama *arquitectura constructiva.*

11.1.9.3. EL MODELO ART COMO RED COMPETITIVA

El modelo ART, *Adaptative Resonance Theory* de Carpenter y Grossberg [Carpenter, 1987], es un modelo de redes de neuronas que es constructivo y que utiliza aprendizaje no supervisado. Es uno de los modelos más complejos que existen, y está descrito por una serie de ecuaciones diferenciales no lineales acopladas. Desde su primera versión, el ART1 para trabajar con patrones binarios, Carpenter y Grossberg y otros colaboradores han ido diseñando versiones más complejas: ART2 para patrones reales, ART3 para patrones temporales, ARTMAP para aprendizaje supervisado, etc. Aquí se verá una simplificación de su primera versión, el ART1, donde se habrán reducido al mínimo las matemáticas.

El ART se compone de dos capas, una capa de entrada y una capa de salida que tiene una función clasificatoria. En la capa de entrada se reciben los datos. Esta capa está conectada con la otra capa que intentará clasificar esos datos en alguna clase. La capa clasificatoria tiene tantas neuronas como clases posibles. Esta capa es una capa competitiva, donde *cada neurona representa una clase* y lucha por activarse mientras intenta desactivar al resto. Cuando sólo una de las neuronas quede activada habrá acabado la competición y la clase en la que quedarán clasificados los datos será aquella representada por la neurona ganadora.

La neurona **i** recibe conexiones de la capa de entrada. A su vez, la neurona **i** luchará compitiendo con las otras neuronas de su misma capa utilizando el conexionado. Por un lado, la neurona recibe una conexión positiva de sí misma, ya que se tiene que autofavorecer. Y por otro lado, debe inhibir al resto de sus competidoras por medio de conexiones negativas. Estas otras neuronas, a su vez, se conectarán a ella con conexiones inhibitorias tratando de impedir que ella sea la que se active. Este proceso de competición puede hacerse muy complicado o muy simple, dependiendo de la función de activación que se escoja. Si es una de las habituales, será simple y ni se simula, ya que se sabe que la neurona que reciba más estímulo procedente de las conexiones con la capa de entrada será la ganadora. Por lo tanto se mira desde el principio cuál de ellas es; y ya está.

Lo anterior es el funcionamiento básico de los modelos competitivos como el ART o los mapas de características de Kohonen (1989), otro modelo clásico de red de neuronas competitivo y de aprendizaje no supervisado. Sin embargo, el funcionamiento del ART es más complejo, ya que además de la conexión capa de entrada a capa de competición se incluye una realimentación entre las dos capas al conectar la capa de clasificación a la capa de entrada. El motivo de esta segunda conexión es comprobar el proceso como se verá en el siguiente apartado.

11.1.9.4. EL MODELO ART

Como se ha dicho, el conexiado entre una capa y otra es bidireccional, de capa de entrada a capa clasificatoria y de capa clasificatoria a capa de entrada. Este conexionado es complejo, no en la forma pero sí en ciertas restricciones que se ponen a los valores de los pesos. Antes de pasar a ver estas restricciones conviene describir cómo opera un ART de manera general.

Recuérdese que la tarea a realizar consiste en clasificar datos. El ART recibe el patrón de entrada en su capa inferior. Esta capa activa a las neuronas de la capa competitiva a través de sus conexiones. En esta capa se realiza la competición hasta que sólo una de sus neuronas permanece activa, que simbolizará a una clase candidata a la que pueden pertenecer los datos. Esta neurona genera lo que para ella es el <representante> o <prototipo> de esa clase. El paso siguiente es comparar los datos que en ese momento hay a la entrada. Llegado este punto pueden ocurrir tres cosas:

a) Que se asemejen bastante. En este caso ya se tienen clasificados los datos, la clase candidata era correcta. El ART redefine el representante de la clase seleccionada realizando una <mezcla> o <media> entre él y los datos nuevos.

b) Que no se asemejen bastante. Esto quiere decir que la clase candidata no era la correcta. En este caso, la neurona que ganó en

la competición se inhibe y el resto vuelve a repetir el proceso para buscar otra candidata.

c) Que no se asemejen y que no quede ninguna otra posibilidad en la capa de competición debido a que ya se han probado todas. Esto quiere decir que los datos no pertenecen a ninguna de las clases existentes en el modelo. Lo que hace entonces el ART es incorporar la clase nueva añadiendo otra neurona a la capa competitiva con un representante igual a los datos de entrada.

ANEXO III
CASOS DE USO:
BOTÁNICA

Los hechos no dejan de existir solo porque sean ignorados.

Tomas H. Huxley

CCLXXXVIII

ANEXO III

12. CASO DE USO: BOTÁNICA.

Se realiza una clasificación aplicando el Nuevo Criterio de la Tesis, a un conjunto de familias en Botánica, para corroborar, la clasificación realizada en "Introducción a la Teoría y Práctica de la Taxonomía Numérica", género Bulnesia y sus Especies (Zygophyllaceae) [Crisci, López Armengol, 1983, página 30], en base al método SAHN y deduciendo criterios con fenogramas y coinciden los clusters y familias, dando lugar a la primera contrastación del nuevo criterio [**el autor**].

12.1. ALGORITMIA

12.2. COMIENZA EL PROCEDIMIENTO

12.2.1. Cálculo de promedio de los dominios

12.2.2. Cálculo de similitudes euclideas

12.2.2. Cálculo del vector de covariancia

12.3. ELEMENTOS DE LA MATRIZ DE DATOS

12.3.1. DOMINIOS DE DATOS

Cada OTU tiene 43 ATRIBUTOS y un identificador 1 .. 8:

Caracteres de los OTU´s para asignar valor del dominio de los estados, según su codificación.

Son 43 atributos:

1. **Hábito**
2. **Longitud del internodio (en cm.)**
3. **Diámetro del internodio (en cm.)**
4. **Longitud de la hoja (en cm.)**
5. **Ancho de la hoja (en cm.)**
6. **Longitud del pecíolulo (en cm.)**
7. **Número de folíolos**
8. **Presencia peciólulos**
9. **Disposición de los folíolos en el caquis**
10. **Pubescencia**
11. **Longitud del folíolo (en mm.)**
12. **Ancho del folíolo (en mm.)**
13. **Número de nervaduras primarias del folíolo**
14. **Posición de los folíolos terminales**
15. **Presencia de mucrón en folíolos**
16. **Tipo de inflorescencia**
17. **Longitud del pedúnculo (en mm.)**
18. **Longitud del sépalo (en mm.)**
19. **Ancho del sépalo (en mm.)**
20. **Color de los pétalos**

21. **Longitud del pétalo (en mm.)**
22. **Ancho del pétalo (en mm.)**
23. **Número de nervaduras del pétalo (en mm.)**
24. **Tipo de estambres**
25. **Modificación de los estambres**
26. **Presencia de gran escama junto al estambre**
27. **Presencia de pelos en la base del filamento estaminal**
28. **Presencia de una escama suplementaria junto al estambre**
29. **Agrupación de los estambres**
30. **Longitud del filamento (en mm.)**
31. **Longitud de la antera (en mm.)**
32. **Longitud de la escama (en mm.)**
33. **Presencia de ápice laciniado en la escama estaminal**
34. **Número de carpelos**
35. **Curvatura del estilo**
36. **Número de óvulos por carpelo**
37. **Pubescencia del fruto**
38. **Longitud del fruto (en mm.)**
39. **Ancho del fruto (en mm.)**
40. **Desarrollo del carpóforo**
41. **Longitud del carpóforo (en mm.)**
42. **Forma de la semilla**
43. **Longitud de la semilla (en mm.)**

1	**B.arborea**	
2	**B.carrapo**	
3	**B.chilensis**	
4	**B.bonariensis**	**OCHO (8) OTU´s**
5	**B.retama**	
6	**B.foliosa**	
7	**B.schikendantzi**	
8	**B.sarmientoi**	

Matriz de datos

1 B.arborea 1 2.0000 34.9000 2.1000 84.9000 56.6000 7.7000 13.0000 2.0000 0.0000 2.0000 29.6000 8.6000 6.0000 0.0000 0.0000 2.0000 17.2000 7.1000 3.4000 2.0000 22.4000 17.3000 16.0000 0.0000 1.0000 2.0000 0.0000 0.0000 2.0000 10.2000 1.5000 5.2000 2.0000 2.0000 1.0000 2.0000 0.0000 45.7000 40.8000 2.0000 7.5000 1.0000 13.4000

2 B.carrapo 2 2.0000 35.7000 1.6000 96.6000 70.8000 9.0000 7.0000 2.0000 0.0000 2.0000 39.7000 16.4000 5.8000 1.0000 0.0000

2.0000 18.1000 6.4000 5.8000 2.0000 24.3000 18.5000 12.0000 0.0000 2.0000 2.0000 0.0000 0.0000 2.0000 9.9000 1.4000 5.3000 2.0000 2.0000 2.0000 2.0000 0.0000 56.2000 51.8000 2.0000 5.3000 1.0000 12.1000

3 B.chilensis 3 0.0000 23.5000 2.6000 13.5000 8.9000 1.8000 8.0000 2.0000 1.0000 0.0000 5.2000 2.4000 0.0000 2.0000 0.0000 1.0000 9.0000 7.1000 4.2000 2.0000 9.1000 5.7000 8.0000 0.0000 0.0000 0.0000 2.0000 0.0000 2.0000 7.0000 2.0000 4.4000 1.0000 1.0000 0.0000 7.0000 0.0000 13.4000 12.3000 2.0000 0.6000 2.0000 2.7000

4 B.bonariensis 4 0.0000 20.4000 1.3000 25.9000 17.8000 3.4000 14.0000 1.0000 0.0000 2.0000 8.9000 2.0000 1.0000 2.0000 0.0000 1.0000 11.5000 6.8000 3.9000 2.0000 17.8000 10.2000 10.0000 0.0000 1.0000 1.0000 0.0000 0.0000 1.0000 10.6000 1.6000 4.4000 0.0000 2.0000 2.0000 1.0000 0.0000 36.9000 32.6000 2.0000 4.8000 1.0000 10.8000

5 B.retama 5 1.0000 40.4000 2.0000 12.5000 11.3000 3.1000 5.0000 1.0000 1.0000 2.0000 6.6000 2.6000 2.0000 1.0000 1.0000 1.0000 10.0000 7.4000 4.4000 2.0000 7.7000 4.6000 7.0000 1.0000 0.0000 0.0000 1.0000 0.0000 1.0000 7.1000 2.2000 3.1000 1.0000 2.0000 1.0000 8.0000 0.0000 22.5000 18.8000 1.0000 0.8000 2.0000 11.0000

6 B.foliosa 6 0.0000 18.8000 1.3000 28.4000 25.0000 5.8000 4.0000 1.0000 1.0000 2.0000 13.6000 7.8000 3.0000 2.0000 0.0000 1.0000 13.2000 5.3000 3.0000 2.0000 8.5000 2.7000 6.0000 2.0000 0.0000 0.0000 0.0000 0.0000 0.0000 6.2000 1.6000 3.9000 1.0000 2.0000 1.0000 4.0000 1.0000 16.3000 13.3000 1.0000 0.7000 2.0000 4.9000

7 B.schikendantzi 7 0.0000 10.4000 1.5000 19.7000 11.6000 2.7000 10.0000 0.0000 0.0000 2.0000 5.7000 1.9000 1.0000 2.0000 0.0000 1.0000 10.4000 5.9000 3.1000 2.0000 9.3000 4.3000 5.0000 1.0000 0.0000 0.0000 0.0000 1.0000 0.0000 7.4000 1.7000 4.4000

2.0000 2.0000 1.0000 4.0000 1.0000 11.8000 12.9000 1.0000 0.4000 2.0000 5.3000

8 B.sarmientoi 8 2.0000 21.6000 1.4000 21.4000 27.1000 5.1000 2.0000 2.0000 2.0000 1.0000 16.8000 12.0000 5.0000 2.0000 0.0000 1.0000 3.9000 2.9000 2.3000 1.0000 11.5000 7.0000 6.0000 2.0000 0.0000 0.0000 0.0000 0.0000 0.0000 4.1000 1.1000 2.9000 2.0000 0.0000 0.0000 2.0000 0.0000 51.8000 47.7000 2.0000 5.2000 1.0000 13.5000

Matriz de Similitud

1. B.arborea	1	0.0000	0.6820	1.6727	1.1869	1.5795	1.5909	1.6991	1.6672
2. B.carrapo	2	0.6820	0.0000	1.8444	1.3074	1.7202	1.6585	1.8337	1.7309
3. B.chilensis	3	1.6727	1.8444	0.0000	1.2629	1.0168	1.1883	1.2208	1.5691
4. B.bonariensis	4	1.1869	1.3074	1.2629	0.0000	1.2132	1.1176	1.1271	1.5148
5. B.retama	5	1.5795	1.7202	1.0168	1.2132	0.0000	0.9979	1.1334	1.5256
6. B.foliosa	6	1.5909	1.6585	1.1883	1.1176	0.9979	0.0000	0.7346	1.2734
7. B.schikendantzi	7	1.6991	1.8337	1.2208	1.1271	1.1334	0.7346	0.0000	1.5519
8. B.sarmientoi	8	1.6672	1.7309	1.5691	1.5148	1.5256	1.2734	1.5519	0.0000

12.3.2. MAXIMA DISIMILITUD 1.84439

12.3.3. AGRUPAMIENTO POR INVARIANTES

1 B.arborea i 1

0.0000 1 B.arborea j 1

0.6820 2 B.carrapo j 2

Invariantes:

Distancia Media 1.0000

Densidad 2.00

Dispersion 1.0000

Rango 1.0000

2 B.carrapo i 2

0.6820 1 B.arborea j 1

0.0000 2 B.carrapo j 2

Invariantes:

Distancia Media 1.0000

Densidad 2.00

Dispersion 1.0000

Rango 1.0000

3 B.chilensis i 3

0.0000 3 B.chilensis j 3

Invariantes:

Distancia Media 1.0000

Densidad 1.00

Dispersion 0.0000

Rango 0.0000

4 B.bonariensis i 4

0.0000 4 B.bonariensis j 4

Invariantes:

Distancia Media 1.0000

Densidad 1.00

Dispersion 0.0000

Rango 0.0000

5 B.retama i 5

0.0000 5 B.retama j 5

0.9979 6 B.foliosa j 6

Invariantes:

Distancia Media 1.0000

Densidad 2.00

Dispersion 1.0000

Rango 1.0000

6 B.foliosa i 6

0.0000 6 B.foliosa j 6

0.7346 7 B.schikendantzi j 7

Invariantes:

Distancia Media 0.7346

Densidad 2.00

Dispersion 0.0000

Rango 0.7346

7 B.schikendantzi i 7

0.7346 6 B.foliosa j 6

0.0000 7 B.schikendantzi j 7

Invariantes:

Distancia Media 1.0000

Densidad 2.00

Dispersion 1.0000

Rango 1.0000

8 B.sarmientoi i 8

0.0000 8 B.sarmientoi j 8

Invariantes:

Distancia Media 1.0000

Densidad 1.00

Dispersion 0.0000

Rango 0.0000

12.3.4. TAXONES

1	B.arborea	0	1	0	0	0	0	0	0
2	B.carrapo	1	0	0	0	0	0	0	0
3	B.chilensis	0	0	1	0	0	0	0	0
4	B.bonariensis	0	0	0	1	0	0	0	0
5	B.retama	0	0	0	0	0	1	0	0
6	B.foliosa	0	0	0	0	0	0	1	0
7	B.schikendantzi	0	0	0	0	0	1	0	0
8	B.sarmientoi	0	0	0	0	0	0	0	1

12.3.5. FAMILIAS

1 B.arborea 2 B.carrapo

3 B.chilensis

4 B.bonariensis

5 B.retama 6 B.foliosa 7 B.schikendantzi

8 B.sarmientoi

12.3.6. CLUSTERING

1 B.arborea i 1

0.0000 1 B.arborea j 1

0.6820 2 B.carrapo j 2

Invariantes:

- **Distancia Media 1.0000**
- **Densidad 2.00**
- **Dispersion 1.0000**
- **Rango 1.0000**

2 B.carrapo i 2

0.0000 2 B.carrapo j 2

Invariantes:

- **Distancia Media 1.0000**
- **Densidad 1.00**
- **Dispersion 1.0000**
- **Rango 1.0000**

3 B.chilensis i 3

0.0000 3 B.chilensis j 3

Invariantes:

- **Distancia Media 1.0000**
- **Densidad 1.00**
- **Dispersion 0.0000**
- **Rango 0.0000**

4 B.bonariensis i 4

0.0000 4 B.bonariensis j 4

Invariantes:

- **Distancia Media 1.0000**
- **Densidad 1.00**
- **Dispersion 0.0000**
- **Rango 0.0000**

5 B.retama i 5

0.0000 5 B.retama j 5

0.7346 6 B.foliosa j 6

0.7346 7 B.schikendantzi j 7

Invariantes:

- **Distancia Media 1.0000**
- **Densidad 3.00**
- **Dispersion 1.0000**
- **Rango 1.0000**

6 B.foliosa i 6

0.0000 6 B.foliosa j 6

Invariantes:

- **Distancia Media 0.7346**
- **Densidad 1.00**
- **Dispersion 0.0000**
- **Rango 0.7346**

7 B.schikendantzi i 7

0.0000 7 B.schikendantzi j 7

Invariantes:

- **Distancia Media 1.0000**
- **Densidad 1.00**
- **Dispersion 1.0000**
- **Rango 1.0000**

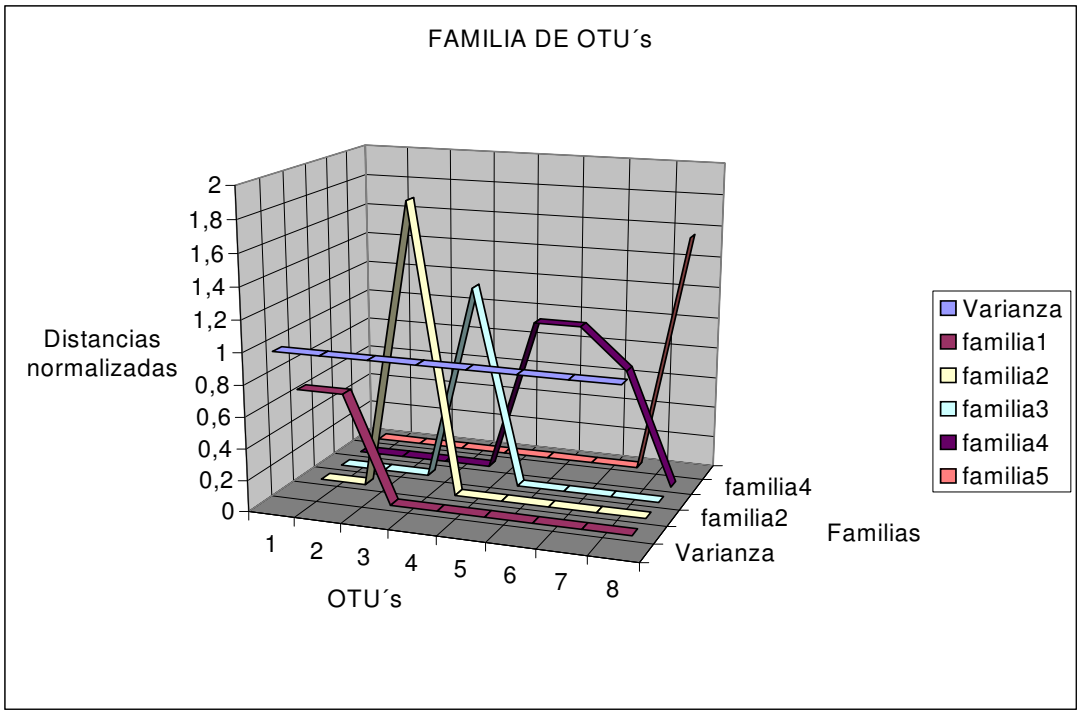
8 B.sarmientoi i 8

0.0000 8 B.sarmientoi j 8

Invariantes:

- **Distancia Media 1.0000**
- **Densidad 1.00**
- **Dispersion 0.0000**
- **Rango 0.0000**

12.4. Espectro de las familias.



BIBLIOGRAFÍA

ICCCVIII

BIBLIOGRAFÍA

Abramson, N., "Information Theory and Coding". McGraw Hill. Paraninfo. Madrid. 1966.

Acedo, C.F., A Plastino y A. N. Proto. Journal of Mathematical Sociology, 1997.

Acedo, C.F., y A.N. Proto. Proceedings of the Neurap97, Neurtal Networks and their applications. Maeseilles, March 12-14, 1997

Aho, A.V., Sethi, R., Ullman, J.D. "Compiladores: Principios, Técnicas y Herramientas". Addison Wesley Iberoamericana. 1990.

Aldenderfer, M., Blashfiel, R. "Cluster analysis". Beverly Hills: Sage Publications. 1984.

Alhassid, Y. and R.D. Levine, J. Chem. Phys. 67, (1977) 4321.

Aliaga, J. and A.N. Proto. Phys. Lett. A142 (1989) 63

Aliaga, J, J.L. Gruver, and A.N. Proto Condensed matter Theories Vol 8, San Juan, Pto. Rico L. Blum and B. Malik (Eds).1993 Plenum Press

Aliaga, J, G. Crespo, and A.N. Proto Phys. Rev. A August (1991)

Aliaga, J., G. Crespo and A.N. Proto. Phy. Rev. Lett.70 (1993) 434

Althusser, Louis. Pour Marx. Teoría semántica. F. Maspero. París. Francia. 1965.

Ames, J.S., "Some of the original articles by Fraunhofer and by Wollaston are collected in book form under the title Prismatic and Diffraction Spectra". Harper and Brothers. New York. 1898.

Anderberg, M. "Cluster analysis for applicatios". New York: Academic Press. 1973.

Anderson, J. R. "The place of cognitive architectures in a rational analysis". Proceedings of the Tenth Annual Cognitive Science Conference (pp. 1-10). Montreal, CA: Lawrence Erlbaum. 1988.

Anderson, J.R. “The place of rational analysis in a cognitive architecture”. In K. Van Lehn (Ed.), *Architectures for intelligence*. Hillsdale, NJ: Lawrence Erlbaum. 1998.

Arango J. Tormenta de Ideas. Colombia. Universidad EAFIT. 2002. Internet:

http://www.eafit.edu.co/tda/boletin/TORMENTA%20DE%20IDEAS.htm

Arnold, J.R., “Asteroids Families and Jet Streams”. The Astronomical Journal. 74: pp 1235-1242. 1969.

Bachman, C.V., “Bases de Donnes, un nouvel art de la Navigation”, 01-Informatique 79. 1974.

Bacon, F. El avance del conocimiento. Cambridge University Press. 1605.

Bacon, F. Novum Organum: Indicaciones relativas a la interpretación de la naturaleza. Cambridge University Press. 1620.

Bancilhon, F., “Object -Oriented Database Systems”. In Proceeding of ACM Symposium on Principles of Database Systems. Austin. 1988.

Barsalou, L. “The instability of graded structure: implications for the nature of concepts”. In U. Neisser (Ed.), Concepts and conceptual development: Ecological and intellectual factors in categorization. New York: Cambridge University Press. 1987.

Batini, C., Ceri,E., Navathe,S., “Conceptual Database Design”. Benjaming / Cumming Publishing Company. 1992.

Batory, D., Leung,T., Wise,T., “Implementation Concepts for an Extensible Data Model and Data Language”. ACM Transaction on Database Systems 13(3), pp 231-262. 1988.

Beckner, M. "Reduction, hierarchies, and organicism". En F.J. Ayala y Th. Dobzhansky, *Studies in the Philosophy of Biology*, Berkeley y Los Angeles, University of California Press, 163-177 1974.

Bertino, E., Martino,L., “Object-Oiented Database Management Systems: Concepts and Issues”. IEEE Computer Society, 24(4), pp 33-47 .1991.

Bertino, E., Martino,L., “Sistemas de Bases de Datos Orientados a Objetos. Conceptos y Arquitecturas”. Addison Wesley Iberoamericana. 1993.

Bertino, E. y Montesi,E. “Toward a logical-Objetc-Oriented Programmng Language for Databases.” In Proceeding of 3rd International Conference on Extended Database Technology. Viena. Austria. 1992.

Bertino, E. Damiani,M., Randi,P., “The ADKMS knowledge acquisition system”. In Proceeding of 2nd Far-East Workshop on Future Database Systems. Kyoto, Japón. Advanced Database Research and Development Series volume 3. World Scientific. 1992.

Berztiss, A.T. “Data Structures: Theory and Practice”. Computer Science and Applied Mathematics. Academic Press. New York. 1971.

Blockeel, H., De Raedt, L. *Top-Down Induction of Logical Decision Trees*. Katholieke Universiteit Leuven, Departament of Computer Science, Celestijnelaan, Bélgica. 1997.

Edward S. Blurock, *The ID3 Algorithm*, Research Institute for Symbolic Computation,www.risc.uni-linz.ac.at/people/bulrock/ANALYSIS/manual/document, Austria. 1996.

Borgida, A., Brahmand,R., McGuinness,D., Resnick,L., “CLASSIC: a structural data model of objects”. In Proceeding of ACM-SIGMOD International Conference on Management of Data. Portland. 1989

Breitl, R., “The GemStone data management system. In Objects Oriented Concepts”. Database and Applications. pp 283-308. Addison Wesley. 1989

Brito I. y Moreira, A. Integrating the NFR framework in a RE Model. In Early Aspects 2004: Aspects-Oriented Requirements Engineering and Architecture Design Workshop (AOSD). Lancaster. 2004.

Brooks, F. No Silver Bullet. Essence and Accident in Software Engineering. USA. IEEE Computer. 1987.

Bunge, M. “La Investigación Científica”. Ediciones Ariel. Barcelona. España. 1969.

Bunge, M. “Racionalidad y Realismo”. Ediciones Alianza. Madrid. España. 1983.

Bunge, M. “Las ciencias sociales en discusión.” Editorial Sudamericana. Buenos Aires. Argentina. 1999.

Cacase, F., Ceri,S., Crespi-Reghizzi, S., Tanca, L., Zicari,R., “Integrating object oriented data modeling with a rule-based programming paradigm”. In Proceeding of ACM-SIGMOD International Conference on Management of Data. Atlantic City. 1990.

Carusi, A., Massaro,E. Astronomy and Astrophycs. Supplements 34, p 81. 1978.

Carusi, A., Valsecchi, G.B. “On Asteroids Classifications in Families”. Astronomy and Astrophycs. pp 327-335. 1982.

Chen, P.P., “The Entity/Relationship Model: Toward a Unified View of Data”, CACM ,1,1. .1976.

Chen, P.P., “The Entity/Relationship Model: A Basis for the enterprise View of Data”, AFPS Conference Proceedings,Vol 46 .1977.

Chen, P.P., “Entity/Relationship Approach to Systems Analisys and Design”, North Holland Publish Company. 1980.

Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., Freeman, D. “Autoclass: A Bayesian classification system”. Proceedings of the Fifth International Conferencie on Machine Learning (pp. 54-64). Ann Arbor, MI: Morgan Kaufmann. 1988.

Codd E. F. “A Relational Model of Data for Large Shared Data Banks” CACM 13, 6 pp 377-387.1970.

Codd E. F. “Relational Completeness of Data Base Sublanguages”. Database Systems, Courant Computer Science Symposia Series 6, Englewood Cliffs, New Jersey, Prentice-Hall. 1972.

Codd E. F. “Extending the Data Base Relational Model to Capture More Meaning” ACM TODS 4, 4 pp 397-434. 1979.

Codd E. F. “Relational Data Base. A Practical Foundation for Productivity” CACM 25, 2. 1982.

Codd E. F. “How Relational is your Database Management System ?”. Computer World. 1985.

Codd E. F. “The Relational Model for Database Management: Version 2”. Addison Wesley. 1990.

Corter, J., Gluck, M., Bower, G. “Basic levels in hierarchically structured categories”. Proceedings of the Tenth Annual Cognitive Science Conferencie (pp. 118-124). Montreal, CA: Lawrence Erlbaum. 1988.

Cotrell G.W., Munro P. Y Zipser D. “Learning Internal Representations from Gray-Scale Images: An Example of Extensional Programming”. Ninth Annual Conference of the Cognitive Science Society, pp. 462-473. Hillsdale: Erlbaum. 1987.

Cramer, Harald. “Métodos Matemáticos de Estadística”. Ediciones Aguilar S.A. Madrid. Traducido por Cansado Enrique. 1958.

Crisci, J.V., Lopez Armengol, M.F. "Introducción a la Teoría y Práctica de la Taxonomía Numérica", Organización de los Estados Americanos. Programa Regional de Desarrollo Científico y Tecnológico.Washington D.C. 1983.

Cybenco, G. “Approximation by Superpositions of a Sigmoidal Function”. Mathematics of Control, Signals, and Systems 2, pp. 303-314. 1989.

Dahl, O.J., Nygaard, K., “SIMULA an Algol Based Simulation Language”. On ACM 9(9). 1966.

Date, C.J. “An Introduction to Datase Systems Vol. I”. 2ª Ed. Addison Wesley. 1981.

Date, C.J. “Relational Database: Selected Writings”. Addison Wesley. 1986.

Date, C.J. "Relational Database: Further Misconceptions #1". Info DB, spring, 1986.

Date, C.J. "A SQL Standard". Addison Wesley. 1987.

Date, C.J. "Where SQL Falls Short". Datamation pp 84-86. 1987.

Date, C.J. "An Introduction to Datase Systems Vol. I". 5ª Ed. Addison Wesley. 1990.

Date, C.J. "Date on Databases" On proceeding of the Codd & Date Relational Database Symposium". Madrid. 1992.

David, M. Ishikawa Fish Bone Diagram. 1998 Internet: http://www.mansci.uwaterloo.ca/~msci432/Notes/F_Fish_bone.htm

Davis y Bonnell, "Análisis de subtipos". 1990

de Miguel, A., Piatttini, M. "Concepción y Diseño de Bases de Datos". Addison Wesley. 1994.

de Miguel, A., Piatttini, M., Marcos, E. "Diseño de Bases de Datos Relacionales". Alfaomega - ra-ma. 2000.

Deum, Pierre, Ryle Gilbert y Einstein Albert, definieron las tesis de la Ciencia empírica en el Círculo de Viena (ver Popper, K. 1985).

Deux, O., "The Story of O_2". IEEE Transaction on Knowledge and Data Engineering 2(1), pp 91-108. 1990.

Domany, E., Hemmen, J. L., & Schulten, K. Model of Neural Networks. Springer-Verlag. 1991.

Duda, R., Hart, P. "Pattern classification and scene analysis". New York: John Wiley and Sons. 1973.

Elmasri, R., Navathe, S. "Fundamentals of Database Systems". The Benjamin/Cummings Publishing Company. 1989.

Erickson, G. and Ray Smith, C. (Eds.). Maximum-Entropy and Bayesian Methods. 1989.

Everitt, B. "Unresolved problems in cluster analysis". Biometrics, 35, 169-181. 1979.

Everitt, B. "Cluster analysis". London: Heinemann Educational. 1980.

Feigenbaum, E.A. "The simulation of verbal learning behavior". In E. A. Feigenbaum & J. Feldman (Eds.), Computers and thought. New York: McGraw-Hill. 1963.

Fenton, N.E., Pfleeger, Sh.L. "Software Metrics". PWS Publishing Company. 1997.

Feyerabend, Paul Karl. Tratado contra el método. Tecnos. Madrid. España. 1981.

Feynman, R.P., Leighton, R.B. & Sands, M. "Lectures on physics, Mainly Mechanics, Radiation and Heat". Fondo Educativo Interamericano. pp. 25-2 ff, 28-6 ff, 29-1 ff, 37-4. 1971.

Fikes, R., Kehler,T., "The role of frame-based representation on reasoning". ACM 28(9) pp 904-920. 1985.

Filman R., Elrad, T., Clarke, S. y Aksit, M. Aspect-Oriented Software Development. Addison Wesley, Boston. 2005.

Fisher, D. "A hierarchical conceptual clustering algorithm " (Technical Report 85-21). Irvine: University of California, Department of Information and Computer Science. 1984.

Fisher, D. "Knowledge acquisition via incremental conceptual clustering".(a) Machine Learning, 2, 139-172. 1987.

Fisher, D. "Knowledge acquisition via incremental conceptual clustering".(b) Doctoral dissertation, Department of Information & Computer Science, University of California, Irvine. 1987.

Fisher, D., Langley, P., "Methods of conceptual clustering and their relation to numerical taxonomy". In W. Gale (Ed.), Artificial intelligence and statistics. Reading MA: Addison Wesley.1986.

Fishman, D., "Overview of the IRIS DBMS. In Objects Oriented Concepts". Database and Applications. pp 219-250. Addison Wesley.1989.

Frank, N.H. "Introducción a la Mecánica y el Calor". Science Service. Washington. Editorial Atlante. 1949.

Fraunhofer, J. and Wollaston,W.H. “Prismatic and diffraction”. Annals der Physik. 56, 264, 1817. Translated and edited by J.S.Ames. Harper and Brothers. New York. 1898.

Freeman, J.A., Skapura, D.M. “Redes Neuronales. Algoritmos, aplicaciones y técnicas de programación”. Addison Wesley. Iberoamericana. 1991.

Fried, L.S., Holyoak, K. J. “Induction of category distributions: A framework for classification learning”. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10, 234-257. 1984.

Gallion, R., St Clair, D., Sabharwal, C., Bond, W.E. Dynamic ID3: A Symbolic Learning Algorithm for Many-Valued Attribute Domains. Engineering Education Center, University of Missouri-Rolla, St. Luis, EE.UU. 1993.

Gennari, J.H., Langley, P., Fisher, D. “Models of incremental concept formation” (a). Artificial Intelligence, 40, 11-61. 1989.

Gennari, J.H. “A Survey of Clustering Methods” (b). Reporte Técnico 89-38. Departamento de Informática y Ciencias de la Computación. Universidad de California., Irvine, CA 92717. 1989.

Gianella, Alicia E. Introducción a la epistemología y a la metodología de la ciencia. Editorial de la Universidad Nacional de La Plata. Buenos Aires. Argentina. 2000.

Gluck, M., Corter, J. “Information, uncertainty and the utility of categories”. Proceedings of the Seventh Annual Conference of the Cognitive Science Society (pp. 283-287). Irvine, CA: Lawrence Erlbaum. 1985.

Goldberg, A., Robson,D., “Smalltalk-80: The language and its Implementation”. Addison Wesley. 1983.

Gruver, J.L., J.Aliaga, H.A. Cerdeira and A.N. Proto, Phys.Rev. A 50 (1994) 5274

Gruver, J.L, J.Aliaga, H.A. Cerdeira and A.N. Proto, Phys. Lett.A 184 335 (1994).b

Gruver, J.L., J. Aliaga, H Cerdeira and A.N. Proto Phys.Lett..A 190 (1994) 363 c

Gruver, J.L., J. Aliaga, H. Cerdeira and A.N. Proto. Phys.Rev.E. 51 (1995) 6263

Hamming, R.W. "Coding and information theory". Englewood Clifs, NJ: Prentice Hall. 1980.

Hand, D.J. *"Discrimination and classification".* New York: John Wiley & Sons. 1981.

Hanson, S.J., Bauer, M. "Conceptual clustering, categorization, and polymorphy". Machine Learning, 3, 343-372. 1989.

Hebb, D.O. The organization of behaviour. New York; Wiley. 1949.

Hempel, Carl Gustav. La explicación científica. Estudios sobre la filosofía de la Ciencia.Ediciones Paidós Ibérica. Barcelona. España. 1996.

Hertz, J., Krogh, A. & Palmer, R.G. Introduction to the Theory of Neuronal Computation. Addison Wesley. 1991.

Hetcht, E. and Zajac, A., "Optica". Fondo Educativo Interamericano. pp. 5-11-206-207-293-297-459-534. 1977.

Hirayama, K. "Groups of Asteroids Probably Common Origin". Proceeding of Physics-Mathematics Society. Japan II: 9. pp 354-351. 1918a.

Hirayama, K. "Groups of Asteroids Probably Common Origin". The Astronomical Journal: 31, pp 185-188. 1918b.

Hirayama, K. "Present State of the Families of Asteroids". Proceeding of Physics-Mathematics Society. Japan II:9. pp 482-485. 1933.

Hopfield, J.J., "Neural Networks and Physical Systems with Emergent Colective Computational Abilities". Proceedings of the National Academy of Sciences, USA 79, pp. 2.554-2.558. 1982.

Hopfield, J.J. y Ank, D.W "Computing with Neural Circuits: A Model". *Science* 233, pp. 625-663. 1986.

Hunt, E.B., Marin, J., Stone, P.J. *Experiments in Induction*. New York: Academic Press, EE.UU. 1966.

Hunt, E.B. *Artificial Intelligence.* New York: Academic Press, EE.UU. 1975.

IEEE Std.1471. Recommended Practice for Architectural Description of Software-Intensive Systems. 2000.

Imre Lakatos. El falsacionismo sofisticado. Rodolfo Gaeta y Susana Lucero. Editorial Universitaria de Buenos Aires. Argentina. 1999.

Isasi, P., Martínez, P., Borrajo, D. "Lenguajes, Gramáticas y Autómatas". Addison Wesley Iberoamericana. 1997.

ISHIKAWA, K. Ishikawa Diagram. 1969. Internet: http://imedia.vuse.vanderbilt.edu/mt322/library2/ishikawa.htm

ISO/IEC: FCD 9126-1. Information Technology - Software Engineering Product Quality. Part 1: Quality Model. 2001.

Jacobson, I. Object Oriented Software Engineering. A Use Case Driven Approach. Addison Wesley. 1992.

Jacobson I., Booch G y Rumbaugh J. El Lenguaje de Modelado Unificado. Segunda Edición. Madrid: Addison Wesley. 2000.

Jaynes, E.T., Phys. Rev 106 (1957) 620; 108 (1957)171.

Jaynes, E.T. "Bayesian methods: General background". In J. H. Justice (Ed.), Maximun entropy and Bayesian methods in applied statistics (pp. 1-25). Cambridge, MA: Cambridge University Press. 1986.

Jimenez Rey, E.; Grossi, M., Fernandez, V. "Review of Numerical Taxonomics Methods", State of the art Technical Report. Computer Science Department. School of Engineering. University of Buenos Aires. 1996.

Kant, Immanuel. Crítica de la razón pura. Editorial Losada. Buenos Aires. Argentina. 1973.

Kim, W., Lochovsky,F.,"Objects Oriented Concepts. Databases and Applications. " Addison Wesley. 1989.

Kim, W., "Introduction to Objects Oriented Databases". Cambridge. The MIT Press. Addison Wesley. 1990.

Kim, W., Gallou,N., Garza,JF., Woelk,D., "A distibuted Objects Oriented Database System Suportting Shared and Private Databases". ACM Transactions on Information Systems. 9(1), pp 31-51. 1991.

Kirchhoff, G.R., Bunsen,R., Ångström,A.J. "Chemical analysis with the spectrometer". Annals der Physik. , 110, 160, 1860.

Kitchenham, B., Pickard,L., Pfleeger, S.L. "Case studies for method and tool evaluation". IEEE Software, 12(4) pp 52-62. 1995.

Klimovsky, Gregorio. Las desventuras del conocimiento científico. A-Z. Buenos Aires. Argentina. 1994.

Knêzevîc, Z., Milani, A. "Asteroids Proper Elements from an Analytical Second Order Theory". Astronomy and Astrophycs. pp 1073. 1990.

Knêzevîc, Z., Milani, A., Farinella, P., Froechle,Ch., Froechle, Cl, "Asteroids Family Identifications and Proper Elements". Icarus, 93, 316. 1991.

Kohonen, T. "Self - Organization and Associative Memory". Berlín: Springer - Verlag. 1989.

Korab, H. *Rule Induction: Decision Trees and Rules*, http://www.ncsa.uiuc.edu/News/Access/Stories/97Stories/KUFRIN.html. 1997.

Kotler, P. Dirección de la Mercadotecnia. Séptima Edición. Prentice Hall. 1993.

Kotonya, G. y Sommerville, I. Requirements Engineering. Processes and techniques. Wiley. 1998.

Kowalski, A.M. , A.N. Proto, and A. Plastino Phys. Lett. A 187 (1994) 220.

Kuhn, Thomas S. La estructura de las revoluciones científicas. Editorial del Fondo de la Cultura Económica. Mexico. 1980.

Langley, P. "Machine learning as an experimental science". Machine Learning, 3, 5-8. 1988.

Lebowitz, M. "Categorizing numeric information for generalization". Cognitive Science, 9, 285-309. 1985.

Lebowits, M. “Concept learning in a rich input domain: Generalization based memory”. In R.S. Michalski, J.G. Carbonell, & T.M. Mitchell (Eds.), Machine learning: An artificial intelligence approach (Vol. 2). Los Altos, CA: Morgan Kaufmann. 1986.

Lippmann, R.P., “An introduction to Computing with Neural Nets”. IEEE ASSP Magazine, abril 1987, pp. 4-22. 1987.

Luque Ruiz, Irene, Gómez-Nieto, Miguel Ángel, López Espinosa, Enrique, Cerruela Gracía, Gonzalo. “Bases de Datos desde Chen hasta Codd con ORACLE”. Alfaomega - ra-ma. 2002.

Malinowski, Bronislaw. Los argonautas del Pacífico occidental. Editorial Planeta. Barcelona. España. 1986.

McCulloch, W.S., Pitts,W., ”A logical calculus of ideas immanent in nervous activity. Bulletin of Mathematical Biophysics”, 5:115-133, 1943.

McClelland, J., Rumelhart, D., “Explorations in Parallel Distributed Processing”, Vol.: 1 and 2. MIT Press, Cambridge, MA. 1985.

MacNaughton-Smith, P., Williams, W.T., Dale, M.B., Mockett L.G. “Dissimilarity analysis”. Nature, 202, 1034-1035. 1964.

Martin, J. “Organización de las Bases de Datos”,. Prentice-Hall Latinoamericana. México. 1977.

Mayr, E. Toward a New Philosophy of Biology. Cambridge (Mass.), Harvard University Press. 1988. Teoría sintética de la evolución de la filosofía de la biología. Facultad de Filosofía. Universidad Complutense de Madrid. España. 2003.

Mervis, C., Rosch, E. “Categorization of natural objects”. Annual Review of Psychology, 32,89-115. 1981.

Meyer, B., “Object-Oriented Software Construction.” Prentice Hall. 1988.

Michalski, R.S., Stepp, R. “Learning from observation: Conceptual clustering”.(a) In R.S. Michalski, J.G. Carbonell, & T. M. Mitchell (Eds.), Machine learning: An artificial intelligence approach. Los Altos, CA: Morgan Kaufmann. 1983.

Michalski, R.S., Stepp, R. "Automated construction of classifications: Conceptual clustering versus numerical taxonomy".(b) IEEE Transactions on Pattern Analysis and Machine Intelligence, 5, 396-409. 1983.

Michie, D. On Machine Intelligence (2nd ed.), Ellis Horwood, Chichester, Reino Unido. 1986.

Michie, D. Machine Learning in the next five years, EWSL-88, 3rd European Working Session on Leaming, Pitman, Glasgow, Londres, Reino Unido. 1988.

Milani, A. "Asteroids Family Identifications and Proper Elements". Celestial Mechanics. 57,59. 1993.

Mitchell, T. *Machine Learning*. MCB/McGraw-Hill, Carnegie Mellon University, EE.UU. 1997.

Mitchell, T. *Decision Trees*. Cornell University, www.cs.cornell.edu/courses/ c5478/ 2000SP, EE.UU. 2000a.

Mitchell, T. *Decision Trees 2*. Cornell University, www.cs.cornell.edu/ courses/c5478/2000SP, EE.UU. 2000b.

Moon, D.A., "The common LISP Object-Oriented programming language standard. In Object-Oriented Concepts. Databases qand Applicatios. pp 49-78. Addison Wesley. 1989.

Muller, B. & Reinhart, J. Neural Networks. Springer-Verlag. 1991.

Nagel, Ernest, "La estructura de la ciencia". Paidós, Buenos Aires. Argentina. 1968.

Newton, Sir I., "Phenomena of colors". Transactions of the Royal Society. London. 1672.

Parsons, Talcott. La estructura de la acción social (1937), El sistema social (1951) y Sociedades: perspectivas evolucionistas y comparativas (1966). Harvard University Press. USA.

Perichinsky, G. “Multiple states of multiple state automata to key fast validation”. 11th. International Symposium Computer at University. Catvat. Zagreb. Yugoslavia. 1989.a.

Perichinsky, G., Servetto, A., Crocco, E. “Relational Data Bases Structured on Dynamic Domains of Attributes".18th Sessions. Operative Research and Informatic’s Argentine Society.SADIO. 1989.b.

Perichinsky, G., Servetto, A. "Dynamically Integrated Independent Domains on Data Bases”.19th Sessions Operations Research and Informatic’s Argentine Society.SADIO. 1990.

Perichinsky, G., Servetto, A., Crocco, E. J.Beltramone, A.Gomez Cataldi “Data Base Model Manager Structured on Independent Domains". Facultad de Ciencias Exactas. Universidad Nacional de La Plata. 1991.

Perichinsky, G. “Presidente de la Comisión de Ciencia y Técnica. Redactor, Junto con los Secretarios de la Comisión de Ciencia y Técnica, de las conclusiones y expositor en el plenario. Congreso Nacional Interdisciplinario Diagnósticos y Perspectivas Profesionales y Científicas con Miras al Siglo XXI”. La Plata. Buenos Aires. Con auspicio Académico de la Universidad Nacional de La Plata. 1991.

Perichinsky, G., Servetto, A., Crocco, E. “Data Base Model Manager Structured on Independent Domains". Facultad de Ciencias Exactas. Universidad Nacional de La Plata. 1992.

Perichinsky, G. “Bases de Datos con Dominios Dinámicos y Tablas Virtuales”. I Congreso Internacional de Ingeniería Informática. Buenos Aires. 1994.

Perichinsky, G. “Investigation, Education and Projection in Computer Science.” Proceedings of the First International Congress of Computer Engineering. Pages 306-314. Faculty of Engineering. University of Buenos Aires. Argentina. 1995.

Perichinsky, G., Feldgen, M., Clúa, O. “Bases de Datos Dinámicas de Tablas Virtuales”. II Congreso Internacional de Ingeniería Informática. Buenos Aires. 1995.

Perichinsky, G., Feldgen, M., Clúa, O. “Conceptual Contrast of Dynamic Data Bases with the Relational Model” in Proceedings International Association of Science and Technology for Development. 14ª Applied Informatics Conference. Innsbruck, Austria. 1996.a.

Perichinsky, G., Jimenez Rey, E.; Grossi, M. “Representación Taxonómica en Base de Datos.” Resúmenes de las III Jornadas de Informática e Investigación Operativa. Página 35. Facultad de Ingeniería, Universidad de la República. Montevideo. Uruguay. 1996.b.

Perichinsky, G., Feldgen, M., Clúa, O. "Dynamic Data Bases and Taxonomy" in Proceedings International Association of Science and Technology for Development. Applied Informatics Conference. Innsbruck. Austria. 1997.

Perichinsky, G., Jimenez Rey, E.; Grossi, M. “Domain Sandardization of Oerational Taxonomic Units (OTU´s) on Dynamic Data Bases” in Proceedings International Association of Science and Technology for Development. Applied Informatics Conference. Garmisch-Partenkirchen. Germany. 1998.a.

Perichinsky, G., Jimenez Rey, E.; Grossi, M. “Spectra of Objects of Taxonomic Evidence on the Dynamic Data Bases” in Proceedings International Association of Science and Technology for Development. Applied Informatics Conference. Garmisch-Partenkirchen.Germany. 1998.b.

Perichinsky, G., Clúa, O., Feldgen, M. “Managing multiple user interfaces with agents, object oriented data bases and the Amsterdam hypermedia model” in Proceedings International Association of Science and Technology for Development. Applied Informatics Conference. Garmisch-Partenkirchen. Germany. 1998.c.

Perichinsky, G., Jimenez Rey, E. y Grossi, M.D. Application of Dynamic Data Bases in Taxonomy Astronomic. Proceedings of the XVII

International Conference on Applied Informatics. Pages 120-126 .Innsbruck. Austria. 1999a.

Perichinsky, G., Orellana, R.B., Plastino, A.L., Jimenez Rey, E. y Grossi, M. D. Espectros de evidencia taxonomica en bases de datos. Aplicación cuerpos celestes. familias de asteroides. Proceedings del Quinto Congreso Internacional de Ingeniería Informática. Páginas 301-313. Editado por la Facultad de Ingeniería de la Universidad de Buenos Aires. 1999b.

Perichinsky, G., Orellana, R., Plastino, A.L., Jimenez Rey, E. y Grossi, M.D. Spectra of Taxonomic Evidence in Databases.II. Application in Celestial Bodies. Asteroids families. Proceedings of the XVIII International Conference on Applied Informatics. (Paper 307-7-1).Innsbruck. 2000a.

Perichinsky, G., García Martínez, R. y Proto, A., M.D. Knowledge Discovery Based on Computational Taxonomy and Intelligent Data Mining. VI Congreso Argentino de Ciencias de la Computación, CACIC, CD. Universidad Nacional San Juan Bosco. Sede Ushuaia. Argentina. 2000b.

Perichinsky, G., Servetto, A., Grossi, M. D., García Martínez, R. y Proto, A. Supervised and non-supervised intelligent knowledge discovery, Database and Taxonomy. Workshop de Investigadores en Ciencias de la Computación. Red de Universidades Nacionales con Carreras de Informática. Universidad Nacional de San Luís. 2001a.

Perichinsky, G., Insfran, J., Jaime, E., Sakuda, G., Martín, G., García-Martínez, R. Comparación de Alternativas de Mejoramiento de Compresión de Archivos Utilizando Algoritmos Genéticos. Anales del VII Congreso Internacional de Ingeniería Informática. Facultad de Ingeniería. UBA. ISBN 987-98197-0-5. 2001b.

Perichinsky, G., Orellana, R., Plastino, A.L. Spectra of Taxonomic Evidence in Databases.III. Application in Celestial Bodies. Asteroids

families. Pag. 212-226. International Association for (ACIS) Conference on Computer Science, Software Engineering, Information Technology, e-Business, and Applications. Institute (SEITI), Central Michigan University. Foz do Iguazú. Brazil. 2002.

Perichinsky, G., Servente, M., Servetto, A., García Martínez, R., Orellana, R., Plastino, A.L. Taxonomic Evidence Applying Algorithms of Intelligent Data Mining. Asteroids families. (pp 308-315). International Association for (ACIS) Conference on Computer Science, Software Engineering, Information Technology, e-Business, and Applications. Institute (SEITI), Central Michigan University. Río de Janeiro. 2003a.

Perichinsky, G., Yolis, E., Britos, P., Sicre, J., Servetto, A. C., García Martínez, R. Algoritmos genéticos aplicados a la categorización automática de documentos (pp. 1468-1479). IX Congreso Argentino de Ciencias de la Computación, CACIC, Anales del Congreso. Universidad de La Plata. Buenos Aires. 2003b.

Perichinsky, G., Felgaer, P., Britos, P., Sicre, J., Servetto, A. C., García Martínez, R. Optimización de redes bayesianas basado en técnicas de aprendizaje por inducción (pp. 1687-1687). IX Congreso Argentino de Ciencias de la Computación, CACIC. Anales del Congreso. Universidad de La Plata. Buenos Aires. 2003c.

Perichinsky, G., Servente, M., Servetto, A. C., Garcia Martinez, R., Orellana, R. B., Plastino, A. L. Taxonomic Evidence and Robustness of the Classification Applying Intelligent Data Mining (pp. 1797-1808). IX Congreso Argentino de Ciencias de la Computación, CACIC. Proceeding del Congreso. Universidad de La Plata. Buenos Aires. 2003d.

Perichinsky, G., Fiszelew, A., Britos, P. y García-Martínez, R. Automatic Generation of Neural Networks based on Genetic Algorithms.

Revista de Sistemas de Informação. ISSN: 1677-3071. Año II. Volumen 2. Nº 1. (pp. 1-9). 2003e.

Perichinsky, G. Valores Epistemológicos de la Investigación Científica y su Relación con la Informática. RIEMA. Revista de Informática Educativa y Medios Audiovisuales. ISSN: 1677-8338. Año I. Volumen 2. N° 1 (pp. 12-26). 2004.

Perichinsky Gregorio, Jiménez Rey Elizabeth Miriam, Grossi María Delia, Vallejos Félix Anibal, Servetto Arturo Carlos, Orellana Rosa Beatriz, Plastino Angel Luis. Taxonomic Evidence of Classification. Applying Intelligent Information Algorithm and the Principle of Maximum Entropy: the Case Study of Asteroids Families. Electronic Magazine of Systems of Information, RESI. ISSN 1677-3071. Edición 6 – Año IV – Volumen IV – Número 2. Departamento de Informática y Estadística. Universidad Federal de Santa Catarina. Brasil. 2005.

Perichinsky, G. Consideraciones sobre el mejoramiento de la capacidad de investigación en unidades académicas universitarias impacto en la excelencia académica. V Coloquio en Gestión Universitaria. Asociación de especialistas en Gestión de la Educación Superior. Universidad de Santa Catarina (Brasil), Universidad Nacional de Mar del Plata. Agencia Nacional de Promoción Científica y Tecnológica de la República Argentina y el Instituto de Estudios Superiores de América Latina y el Caribe (IESALC-UNESCO). Mar del Plata. Argentina. 2005.

Perichinsky Gregorio, Jiménez Rey Elizabeth Miriam, Grossi María Delia, Vallejos Félix Anibal, Servetto Arturo Carlos, Orellana Rosa Beatriz, Plastino Angel Luis. Taxonomic Evidence Applying Intelligent Information Algorithm and the Principle of Maximum Entropy: the Case Study of Asteroids Families. Electronic Magazine of Systems of Information, RESI. ISSN 1677-3071. Edición 6 – Año IV – Volumen IV – Número 2. Departamento

de Informática y Estadística. Universidad Federal de Santa Catarina. Brasil. 2006.

Perichinsky, G. “Epistemología como La Investigación en Ciencias” en Epistemología y Metodología de la Investigación psicológica. Facultad de Psicología. Universidad de La Plata. Buenos Aires. Argentina. 2007.

Perichinsky Gregorio, Jiménez Rey Elizabeth Miriam, Grossi María Delia, Vallejos Félix Anibal, Servetto Arturo Carlos, Orellana Rosa Beatriz, Plastino Angel Luis. Taxonomic Evidence of Classification. Applying Intelligent Data Mining. Galactic and Globular clusters. Annals of the Faculty of Engineering Hunedoara - Journal of Engineering. Tome V. Fascicule 2. ISSN 1584 – 2665. University “Politechnica” Timisoara Faculty of Engineering – Hunedoara. 2007.

Poincaré, Henri. Ciencia y método. Académica Científica de Francia. París. 1908.

Popper, Karl. The logic of Scientific Discovery. Science Editions. New Cork. USA. 1961.

Popper, Karl. Conjetures and Refutations. Routledge and Kegan. London. U.K. 1963.

Popper, Karl. Conocimiento objetivo. Tecnos. Madrid. España. 1974.

Popper, Karl. Realismo y el objetivo de la ciencia. Tecnos. Madrid. España. 1985.

Pressman, R. INGENIERÍA DEL SOFTWARE: Un enfoque práctico. Quinta edición. McGraw-Hill. 2002.

Proto, A.N., "Maximun Entropy Principle and Quantum Mechanics" . Condensed Matter Theories, Vol. 5 Valdir Aguilera-Casaca Ed. Plenun Press. 1989.

Quine, William van Orman. Los métodos de la lógica. Ariel. Barcelona. España. 1975.

Quine, William van Orman. Theories and things. Cambridge. Harvard University Press. USA. 1981.

Quinlan, J. “Introduction of Decision Trees”. Machine Learning. Vol.1. N° 1. Pp. 81-106. 1986.

Quinlan, J.R., Cameron-Jones, R.M. *Oversearching and Layered Search in Empirical Learning*. Basser Departament of Computer Science, University of Science, Australia. 1995.

Quinlan, J.R. *Generating Production Rules from Decision trees*. Proceeding of the Tenth International Joint Conference on Artificial Intelligence, páginas. 304-307. San Mateo, CA., Morgan Kaufmann, EE.UU. 1987.

Quinlan, J.R. *Decision trees and multi-valued attributes*. En J.E. Hayes, D. Michie, and J. Richards (eds.), Machine Intelligence, Volumen II, páginas. 305-318.Oxford University Press, Oxford, Reino Unido. 1988b.

Quinlan, J.R. *Unknown Attribute Values in Induction*. Basser Departament of Computer Science, University of Science, Australia. 1989.

Quinlan, J. R. *Learning Logic Definitions from Relations*. En Machine Leaming, Vol 5, páginas 239-266. Oxford University Press, Oxford, Reino Unido. 1990.

Quinlan, J.R. *The Effect of Noise on Concept Learning*, En R. S. Michalski, J. G. Carbonell, & T. M. Mitchells (Eds.) Machine Learning, The Artificial Intelligence Approach. Morgan Kaufmann, Vol. I, Capítulo 6, páginas149-167. San Mateo, CA: Morgan Kaufmann, EE.UU. 1993a.

Quinlan, J.R. *Learning Efficient Classification Procedures and Their Application to Chess Games*, En R. S. Michalski, J. G. Carbonell, & T. M. Mitchells (Eds.) Machine Learning, The Artificial Intelligence Approach. Morgan Kaufmann,Vol. II, Capítulo 15, páginas 463-482, EE.UU. 1993b.

Quinlan, J.R. *Combining Instance-Based and Model-Based Learning*. Basser Departament of Computer Science, University of Science, Australia. 1993c.

Quinlan, J.R. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, San Mateo, California, EE.UU. 1993d

Quinlan, J.R. *MDL and Categorical Theories*. Basser Departament of Computer Science, University of Science, Australia. 1995.

Quinlan, J.R. *Improved Use of Continuous Attributes in C4.5*. Basser Departament of Computer Science, University of Science, Australia. 1996a.

Quinlan, J.R. *Learning First-Order Definitions of Functions*. Basser Departament of Computer Science, University of Science, Australia. 1996b.

Robertson, S. y Robertson, J. Mastering the Requeriments Process. Pearson. 1999.

Rohlf, F.J., Sokal, R.R. "Coefficients of Correlation and Distance in Numerical Taxonomy". Kansas University. Sci. Bull. 45,3 1965

Roth, M.A., Korth, H.F., Siberschatz, A., “Extended Algebra and Calculus for Nested Relational Database”. ACM Transaction on Database Systems. 13(4), pp 389-417. 1988.

Romesburg, H.C. “Cluster analysis for researchers”. Belmont, C.A: Lifetime Learning Publications.1984.

Samaja, Juan. Epistemología y Metodología. Editorial Universitaria de Buenos Aires. Argentina. 1993.

Sawyer, R.A. “Experimental Spectroscopy”. Dover Publication. New York. 1963.

SEI “Rationale for SQL Ada module Description language (SAMeDL)” Ver. 2.0 CMU/SEI-92-TR-16, oct 1992.

Schlimmer, J.C., Granger, J.H. “Beyond incremental processing: Tracking concept drift”. Proceedings of the Fifth National Conference on

Artificial Intelligence (pp. 502-507). Philadelphia, PA: Morgan Kaufmann. 1986.

Schwarz, “Extensibility in the Starburst database system”. In Proceeding of the International Worksh on Object-Oriented Database systems. Pacific Grove, CA. 1986.

Shan y Shixuan, “Normal Entity-Relationship Model. A new Method for Enterprise Schema Design”. IEEE. 1984

Shannon, C.E., Bell. Sys. Tech.J.27 (1948)

Shannon, C.E., The mathematical theory of communication. University of Illinois Press, Urbana, IL. 1963.

Sheck, H.J., Scholl, M.H. “The Relational Model with Relational-values attributes”. Information Systems. 11(2), pp 137-146. 1986

Silberschatz, A., Korth, H.F., Sudarshan, S. “Data Bases Design”. Addison Wesley. 1993.

Silberschatz, A., Korth, H.F. Sudarshan, S. “Data Bases Concepts”. McGraw Hill. 1997.

Simpson, G.G. "Principles of Animal Taxonomy". Columbia University Press. New York. 1961.

Smith, P.D., Barnes, G.M., “Files and Databases: An Introduction”. Addison Wesley. 1987.

Smith, E.E., Medin, D.L. “Categories and concepts”. Cambridge, MA: Harvard University Press. 1981.

Sokal, R.R., Sneath, P.H.A. "Numerical Taxonomy". W.H. Freeman and Company. 1973.

Sommerville I y Sawyer, P. Requirements Engineering. A Good Practice Guide. John Wiley and Sons, New Cork 1997.

Staugaard, Andrew C. Jr. "Técnicas ESTRUCTURADAS Orientadas a Objetos. Una introducción utilizando C++." 2da. Ed. Prentice-Hall. 1998.

Stepp, R. “Concepts in conceptual clustering”. Proceedings of the Tenth International Joint Conference on Artificial Intelligence (pp. 211-213). Milan, Italy: Morgan Kaufmann. 1987.

Stonebraker, M., Wong,E., Kreps,P., Held,G., “The Design and Implementation of INGRES”. ACM Transaction on Database Systems, 1(3), pp 189-222. 1976.

Stonebraker, M., Rowe,L., Hirohama,M., “The Implementation of Posgres”. IEEE Transaction on Knowledge and Data Engineering”, 2(1) , pp 125-142. 1990.

Stonebraker, M., “The INGRES Papers: The Anatomy of a Relational Database Management System”. Addison Wesley. 1990.

Stroustrup, B., “The C++ Programming Language”. Addison Wesley. 1986.

Teorey, T.J., Fry,J.P., “Design of Database Structures”. Englewood Cliffs. 1982.

Ullman, J - “Principles of Database and Knowledgebase Systems”. Vol. I Computer Science Press. 1990.

van Wijngaarden, A., “Revised Report on the Algorithmic Language Algol 68”. Springer-Verlag. 1976.

Weidlich, W Phys. Rep. 204 (1991)1.

Wiederhold, G., **“**Data Base Design”. McGraw-Hill Book Company.1983. Edición 1985.

Wiegers K. Software Requirements: Practical techniques for gathering and managing requirements throughout the product development cycle. Microsoft Press, Washington, USA, pp 12-14. 2003.

Williams, J.G. “Proper Elements and family membership of asteroids”,. Icarus: 72, pp 276-303. 1987.

Williams, J.G. Hierath, J. “Palomar-Leiden minor planets. Proper Elements, Frequency distributions, Belt boundaries and Family memberships”. In Asteroids, pp 1040-1063. 1979.

Williams, J.G. “Asteroids Family Identifications and Proper Elements”. Jet Propulsion Laboratory. Palomar-Leiden minor planets. Asteroids Edition. University of Arizona Press. p 1034. 1989.

Williams, J.G. “Asteroids Family Identifications and Proper Elements”.Icarus. 96,251. 1992a.

Williams, J.G. “Asteroids Family Identifications and Proper Elements”.Icarus. Sbm. 1992b.

Witten, I.H., Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Diego, EE.UU. 2000.

Yourdon, E. Análisis Estructurado Moderno. Primera edición. Prentice Hall. 1993.

Zappala, V. Cellino, A., Farinella, P., Knêzevîc, Z. **"**Asteroid Families. I. Identification by Hierarchical Clustering and Reliability Assessment " The Astronomical Journal, 100, 2030. 1990.

Zappala, V. Cellino, A., Farinella, P., Milani, A., "Asteroid Families. II. Extension to Unnumbered Multiopposition Asteroids" The Astronomical Journal, 107, 772. 1994.