Environment based clustering. A new approach

Lic. Lanzarini Laura
Full-Time Co-Chair Professor, Fac. of Computer Sciences, UNLP.
e-mail: laural@info.unlp.edu.ar
Ing. De Giusti Armando
Director of the LIDI. Principal Researcher CONICET. Full-Time Chair Prof., Fac of Computer Sciences, UNLP.
e-mail: degiusti@info.unlp.edu.ar
LIDI - Laboratorio de Investigación y Desarrollo en Informática
Facultad de Informática, Universidad Nacional de La Plata
Calle 50 y 115 - 1er piso - 1900 La Plata - Buenos Aires - Argentina.
Tel / Fax: 54 - 21 - 22 7707
E-mail: lidi @ info.unlp.edu.ar

Abstract

There is a variety of problems that require an automatic classification of a set of data. In this sense, clustering techniques have been widely applied, since they are known for forming classes or groups using a predefined similarity measure.

This paper defines a new method which, as opposed to the solutions found so far, does not require any previous information about the data to be classified.

The performance of this new proposal has been compared with a winner-take-all type method (WTA), which is widely used in clustering processes, and with the CDL method (Torbjorn, 1998), with satisfactory results.

Keywords: Clustering Techniques – Image Segmentation – Classification.

1 Introduction

Clustering techniques, as their name suggests, are characterized because they input group objects using some similarity measure. The result of this process is the creation of *classes or groups*.

The elements to be grouped are represented by their corresponding characteristic vectors, and it is assumed that the elements belonging to the same class present similar values for a given similarity measure.

When looking for a non-supervised pattern classification, clustering techniques become an ideal solution. Applications from different areas, such as sonar, radar or medical diagnosis, require the classification of a situation.

Due to this wide range of applications, non-supervised patterns classification has been studied in detail (Maravall, 1994), (Gonzalez, 1992), (Jain, 1998).

In the particular case of clustering techniques, the algorithms can be separated in two classes: those that use an only representative or descriptor for each class, and those that used several descriptors.

The former produce good results when applied to problems where the classes present very little dispersion since with only one representative only what belongs to the surrounding hypersphere can be recognized.

Variations of these methods use hypercubes and produce a similar effect (Meneganti, 1998). In particular, the methods proposed by Simpson (Simpson, 1992), (Simpson, 1993)

present alternatives to classify data in a few runs, but their result depends on the order of the input patterns.

On the other hand, those using several representatives, such as (Torbjorn, 1998), solve the class dispersion problem but require the setting of initial similarity parameters, which depend on the problem, allowing to establish a relationship between those representatives.

This paper defines a new clustering method belonging to the second group with the aim of improving the previous proposals in order to achieve an automatic classification that does not require initial parameters nor is dependent on the order of the analysis of the data.

2 New Technique Proposed

Step 1: Analysis of the environment of each pattern to classify

The process starts with an analysis of the input data or patterns.

Since the purpose is to relate them, their corresponding environments will be analyzed (see section 3). This analysis will allow to obtain two values P_i for each pattern:

1. *DistMAX:* every pattern P_j , with $j \neq i$, that is within a distance shorter than that value, will be considered to be similar to P_i and will therefore have a tendency to belong to the same class.
2. *DistMIN*: if the distance between P_i and P_j is shorter than this value, P_i and P_j will be considered to be very similar, and therefore it will be enough to use only one descriptor to represent both.

Step 2: Initial classes

Initially there will be no class assigned.

Class formation:

From this point on, the next iterative process will allow to relate the patterns by creating the corresponding classes:

Step 3: Distribution of the patterns among the existing classes

Let $C = \{C_1, ..., C_k\}$ be the set of classes created so far.

Let $P = \{ P_1, ..., P_n \}$ be the set of patterns to classify.

Each class C_l will be represented by a set of prototypes:

$Prot_l = \{Prot_{l1}, ..., Prot_{ls}\}$. with $l=1..k$

Note that the amount of prototypes varies with the class.

Each pattern not yet classified will analyze its distance with the prototypes of each class in the following way:

If $dist(Prot_{ji}, P_t) < DistMAX_{Class\ j}$, the pattern P_t will belong to class j, where

$DistMAX_{Class\ j}$,= average($DistMAX_{Protji}$) with $i=1..s$

s = number of prototypes in class j

If P_t turns out to belong to several classes, these will be all joined into only one class.

If P_t turns out to belong to an only class, it will be necessary to analyze if there is some new information to contribute to the class; that is, if it can be a new prototype.

To do so it must be true that $dist(Prot_{ji}, P_t) > DistMIN_{Clase\ j}$

If, on the contrary, P_t did not belong to any of the existing classes, a new class with this P_t as the only prototype will be created.

The values $DistMAX_{Clase\ j}$ and $DistMIN_{Clase\ j}$ will be obtained from the average of the values of the prototypes of $Class_j$.

Step 4: Deletion of small classes

All classes with less patterns than the 0.5% of the total of patterns to classify will be deleted.

Repeat steps 3 and 4 until 90% of the input patterns are classified, or until the number of patterns per class is constant.

Step 5: Joining of near neighbors

Each pattern and its closest neighbor will be analyzed. If they belong to different classes buth the distance between them is shorter than the DistMAX of any of the two classes, the classes will be joined.

3 Initial analysis of input patterns

a) Representation of the input patterns

It is important to bear in mind that, depending on the characterization used and the problem involved, patterns could be repeated; therefore, not only pattern characteristics but also their cardinality will be taken into account for each pattern.

b) Distance estimation

This is one of the most important steps to achieve a correct result.

For each pattern, two distance values are required:

— Distance to its nearest neighbor:

For pattern P_i, it will be denoted as $DistMIN_{Pi}$.

$$DistMIN_{Pi} = \min(\ dist(\ P_i, P_j\ )\ ) \qquad \text{with } j \neq i$$

— Distance between patterns of a same class

P_i, it will accept as members of its class those patterns that fulfill the following condition $dist(\ P_i, P_j\ ) \leq DistMAX_{Pi}$ with $j \neq i$

In order to determine this threshold value, the three shortest distances will be considered, and for each of them the number of patterns (multiplied by their cardinality) will be registered. Be *TotPatrones* the sum of the patterns found at these three distances (see Fig. 1).

$DistMAX_{Pi}$ will be the distance that allows to include 50% of *TotPatrones*.

Thus, $DistMAX_{Pi}$ will be for P_i a measure of proximity.

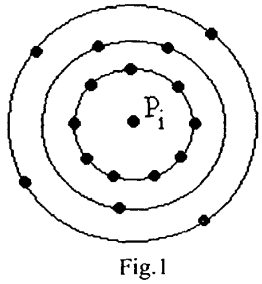

Fig.1
P_i will have 9 neighbors at a distance D1, 3 at a distance D2 and 4 at a distance D3, with D1<D2<D3.

In the example of Fig. 1, $DistMAX_{Pi} = D1$ and every pattern at a distance which is less than or equal to D1 will belong, for P_i, to its class.

By taking three distances greater than zero, the central pattern is forced to have many representatives in order to be isolated, otherwise it will have at least one neighbor to which it will have to be joined.

4 Implementation aspects

As it can be seen, classes are formed around prototypes. Even though it is not necessary to have the initial classes, the classification process can be sped up by chunking the characteristics space into equal sectors and by selecting a pattern from each of them.

Then each pattern will be considered the first prototype of a new class.

On the other hand, each prototype added to a class not only contributes to it with its characteristics but also with its similarity values DistMAX and DistMIN.

The admission of a pattern as a member or new prototype of class j will be given by $DistMAX_{Class\ j}$ and $DistMIN_{Class\ j}$ respectively. each of them is obtained as an average of the distance values of the prototypes forming it so far. In order to make the method independent from pattern insertion order, step 5 is used.

5 Results obtained

The method proposed was applied to two different types of problem.

Case 1: The idea is to classify the black pixels of the binary image of Fig. 2 characterized by their position. This generates a set of patterns with no repetition where each class presents a high dispersion.

The method with only one descriptor used was that of K-Means (Maravall, 1996) with k=7. Fig. 3 shows that it is incapable of finding the correct answer even though tha number of classes to create is indicated *a priori*.

Fig. 4 shows the classification carried out by means of the CDL method (Torbjorn, 1998) and the one proposed in this paper.

For CDL, the following initial parameters were assigned: $\xi = 0.025$, $\xi_L = 0.035$, $\eta_{min}=0.5\%$, $\eta_{max}=60\%$.

The method proposed here (EBC) achieves a correct classification with no previous information at all.

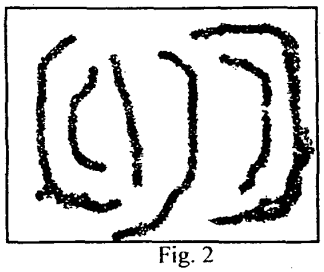

Fig. 2
Original Image

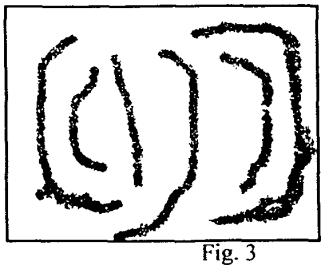

Fig. 3
k-Means with 7 classes

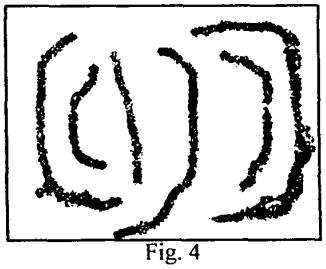

Fig. 4
CDL and EBC

Case 2: Analysis of a histologic sample (liver tissue).
The idea is to group the pixels of the image (Fig. 5) using only their color as characteristic.

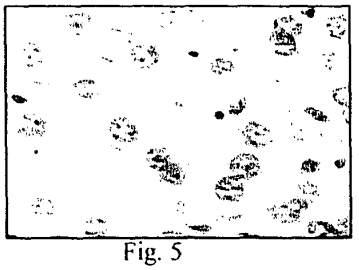

Fig. 5
Original Image

Figures 7 and 8 show the application of the k-means method using two different values for the number of classes. Since the dispersion of the classes to create is lower than in the previous example, the result obtained is only affected by the value of k. It can be seen that the fourth class of the k-means application with four classes (Fig. 6) is separated in two (Figs. 7 and 8) when k increases, which shows its high dependence on the parameter.

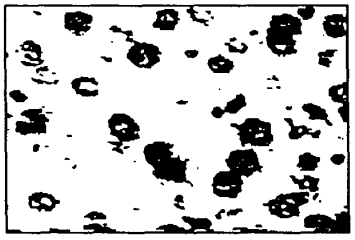

Fig. 6
Class Nr. 4 of K-Means with 4 classes (k=4)

Fig. 7
Class Nr. 4 of K-Means with 5 classes (k=5)

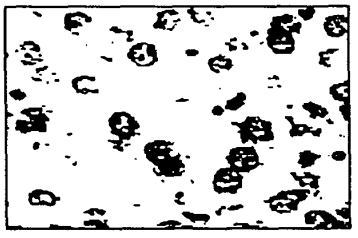

Fig. 8
Class Nr. 5 of K-Means with 5 classes (k=5)

The application of the CDL method implies the determination of the value of the input parameters. The values used in Fig. 9 were obtained as intermediate results of the EBC method. Fig. 10 is the consequence of the application of CDL with incorrect parameters.

Fig. 9
ξ=0.009,ξ_L=0.03, η_{min}=0.5,η_{max}=60

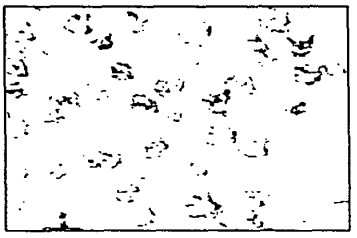

Fig. 10
ξ=0.01, ξ_L=0.02, η_{min}=0.5, η_{max}=60

Finally, the EBC method allows us to get to the solution in Fig. 11 with no need for additional information.

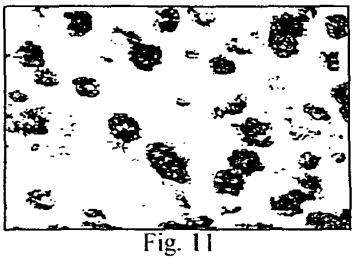

Fig. 11

Conclusions and Future Work:

A new clustering method capable of adapting itself to the set of input patterns has been presented. When compared with other methods, it ha been shown that it is able to provide satisfactory results.

This method was conceived for the classification of pixels at 256 colors, and it is part of a recognition process of the elements of a liver tissue sample currently being developed.

A sample of liver tissue is composed of approximately 120 images of 640x480 pixels. Thsi volume of data has to be reduced in order for to be processed within a suitable response time. The method proposed here is independent from the size of the image and uses the cardinality criteria of each pattern.

However, when the size of the set of data to classify is large, the evaluation of the environment of each pattern may have a high computational cost. This leads to obtain problem-dependent solutions.

All documentation is available at the L.I.D.I. (Laboratorio de Investigación y Desarrollo en Informática, Laboratory of Research and Development in Computer Sciences), 50 y 115 1er. Piso, La Plata, Argentina.

References

1. Baxes Gregory (1994). *Digital Image Processing.* Wiley
2. Gonzalez and Woods (1992), *Digital Image Processing.* Addison-Wesley.
3. Jain Anil (1989). *Fundamentals of Digital Image Processing.* Prentice Hall.
4. Maravall Gomez-Allende (1994). *Reconocimiento de Formas y Visión Artificial.* Addison-Wesley Iberoamericana
5. Meneganti Massimo, Saviello F., Tagliaferri R. (1998). "Fuzzy Neural Networks for classification and detection of anomalies". *IEEE transactions on Neural Networks.* , Vol. 9, nr. 5, pp 843-860.
6. Newton S., Pemmaraju S. and Mitra S (1992). "Adaptive Fuzzy Leader Clustering of Complex Data Sets in Pattern Recognition". *IEEE transactions on Neural Networks.* , Vol. 3, nr. 5, pp 794-800.
7. Niemann, H. (1989). *Pattern Analysis and Understanding.* Springer-Verlag.
8. Simpson Patrick (1991). "Fuzzy Min-Max Neural Networks – Part 1: Clustering". *IEEE Transactions on Neural Networks*, Vol. 3, nr. 5 , pp 776-786.
9. Simpson Patrick (1993). "Fuzzy Min-Max Neural Networks – Part 2: Clustering". *IEEE Transactions on Fuzzy Systems*, Vol. 1, nr 1 , pp 32-45.
10. Torbjorn Eltoft (1998). "A new neural network for cluster-detection-and-labeling". *IEEE Transactions on Neural Networks*, Vol. 9, nr. 5, pp 1021-1035.