Genetic Algorithms applied to Clustering Problem and Data Mining

J.F. JIMENEZ[a], F.J. CUEVAS[b], J.M. CARPIO[a]
[a]Instituto Tecnológico de León., Av. Tecnológico s/n, Fracc. Julián de Obregón, CP.37290 León, Guanajuato, México.
[b]Centro de Investigaciones en Óptica A.C., Loma del Bosque 115, CP. 37150, León, Guanajuato, México

http://www.cio.mx and http://simba.itleon.edu.mx/principal2.html

Abstract: - Clustering techniques have obtained adequate results when are applied to data mining problems. However, different runs of the same clustering technique on a specific dataset may result in different solutions. The cause of this difference is the choice of the initial cluster setting and the values of the parameters associated with the technique. A definition of good initial settings and optimal parameters values is not an easy task, particularly because both vary largely from one dataset to another. In this paper the authors investigate the use of Genetic Algorithms to determine the best initialization of clusters, as well as the optimization of the initial parameters. The experimental results show the great potential of the Genetic Algorithms for the improvement of the clusters, since they do not only optimize the clusters, but resolve the problem of the number K cluster, which had been giving it form a priori. The techniques of clustering are most used in the analysis of information or Data Mining, this method was applied to Data Set at mining.

Key-Words: - Clustering Techniques, Data Mining, *k*-means, Genetic Algorithms

1 Introduction

Clustering has always been a key task in the process of acquiring knowledge. The complexity and specially the diversity of phenomena have forced society to organize things based on their similarities. The objective of cluster analysis is to sort the observations into clusters such as the degree of natural association which is high among members of the same cluster and low between members of different clusters (Berry, 2003; Tou and Gonzalez, 1974; Webb, 2002), the complexity of such task is easily recognized due to the number of possible arrangements. Sensitivity to initial points and convergence to local optimum are usually among the problems affecting the interactive techniques such as *k*-means (Bradley and Fayyad, 1998). Largely used, cluster analysis has called the attention of a very large number of academic disciplines. Most of the work done on internal spatial and social structure of cities has in some way used classification as a basis for data sets analysis using Data mining.

There are several established methods for generating a clustering algorithmically (Everitt, 1992; Kaufman and Rousseeuw, 1990; Gersho and Gray, 1992). The most cited and widely used method is the *k*-means algorithm (McQueen, 1967). It begins with an initial solution, which is iteratively improved using two different optimality criteria in turn until a local minimum has been reached. The algorithm is easy to implement and it gives positive results in most cases.

The problem of the techniques clustering includes two search and selection sub-problems: (1) number of clusters to forming (*k*), and (2) the initial elements of these clusters. Clustering of adequate quality has been obtained by genetic algorithms (GA) (Kivijarvi and Frati, 2003; Naldi and Carvalho, 2003; Wang, et al. 2006), Solving the problem of initialization of the clusters. However, it does not solve the selection problem of the number of clusters.

In this paper we propose an adaptive genetic algorithm for the clustering problem, our aim is to give an effective algorithm which obtains good solutions for the optimization problem without explicit parameter tuning, and each individual of the GA population contains a set of parameter values. These parameters are used for the generation of clusters.

2 Data mining and Clustering

Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science. Moreover, depending on the data mining approach used, techniques from other disciplines may be applied, such as neural networks, fuzzy and/or rough set theory, knowledge

representation, inductive logic programming, or high performance computing. Depending on the kinds of data to be mined or on the given data mining application, the data mining system may also integrate techniques from spatial data analysis, information retrieval, pattern recognition, image analysis, signal processing, computer graphics, Web technology, economics, or psychology.

A data mining system has the potential to generate thousands or even millions of patterns, or rules. Are all of the patterns interesting? The answer in not, only a small fraction of the patterns potentially generated would actually be of interest to any given user. Clustering also has been studied in the fields of machine learning and statistical pattern recognition as a type of unsupervised learning because it does not rely on predefined class-labeled training examples (Duda, Hart, & Stork, 2001).

The kinds of knowledge to be mined: This specifies the data mining functions to be performed, such as characterization, discrimination, association, classification, clustering, or evolution analysis. For instance, if studying the buying habits of customers in Mexico, you may choose to mine associations between customers.

Inside the data mining they working in practical pattern-classification and knowledge-discovery problems require the selection of a subset of attributes or features to represent the patterns to be classified, some works manage with genetic algorithms, memetic algorithms, in some cases cultural algorithms (Sikora and Piramuthu, 2007; Ochoa, et al, 2007), in our case we will use clustering techniques and will optimize them with GA

3 Clustering

Clustering methods partition a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria.

The clustering problem is defined as follows. Given a set of N data objects x_i, partition the data set into K clusters in such a way that similar objects are cluster together and objects with dissimilar features belong to different clusters. M patterns $x_1, x_2, ..., x_M$, a process clustering consists of searching K clusters $S_j, j = 1,2,..,K$. Every cluster is characterized to have centroid (mean), it is the optimal pattern of the cluster, and is formed for $Z_i, i = 1,2,...,K$. Scheme functionally clustering techniques as is indicated in the Figure 1.

The general clustering problem includes two sub-problems: (1) Initialization of centroids and patterns processing, (2) decision of the number of clusters.

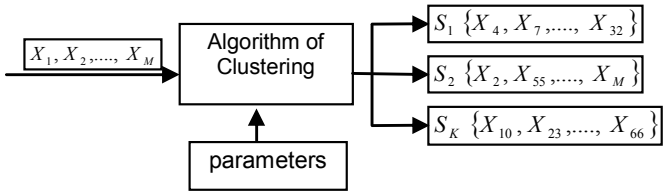

Fig. 1. Scheme functionally clustering techniques.

The most cited and widely used method is the *k*-means algorithm (McQueen, 1967). It begins with an initial solution, which is iteratively improved using two different optimality criteria in turn until a local minimum has been reached. The algorithm is easy to implement and it gives reasonable results in most cases.

Typically the *k*-means algorithm starts with an initialization process in which seed positions are defined. This initial step can have a significant impact on the performance of the method (Bradley and Fayyad, 1998) and can be done in a number of ways (Bradley and Fayyad, 1998). After the initial seed had been defined each data element is assigned to the nearest seed. The next step consists on repositioning the seeds, this can be done after all elements are assigned to the nearest seed or as each one of the elements is assigned. After this, a new assignment step is necessary and the process will go on until no further improvement can be made, in other words a local optimum has been found. Considering that the assignments will be done on the basis of the distance to the nearest seed, implicitly this process will produce a minimization of the sum of the distance squared between each data point and its nearest centroid of the cluster (Bradley and Fayyad, 1998).

The measurement of similarity simpler is the distance, if d it is a measurement of dissimilarity defined between two patterns there turns out to be evident:

$$d(X_i, X_i) = 0$$
$$d(X_i, X_j) \neq 0 \quad \forall j \neq i \tag{1}$$

In the scientific literature (Bow, 2002; Tou and Gonzalez, 1974; Webb, 2002) they can find different expressions, Euclidean distance is a widely used

distance function in the clustering context, and it is calculated as:

$$d(x_1, x_2) = \sqrt{\sum_{M=1}^{M} \left(x_1^M - x_2^M\right)} \quad (2)$$

The most important choice in the clustering method is the objective function for evaluating the quality of a cluster, a commonly used objective criterion is to minimize the sum of squared distances of the data objects to their cluster representatives, and it is calculated as:

If Z_i is the centroid of the clustering S_j, calculated like

$$Z_i = \frac{1}{N_i} \sum_{X \in S_i} X \quad (3)$$

The sum of squared errors is:

$$J_e = \sum_{i=1}^{K} \sum_{X \in S_i} \|X - Z_i\|^2 \quad (4)$$

The specifications of the algorithm of k-means are the following:

A. Algorithm using to join a data set M information in K clusters.
B. The algorithm converges to local optimum Z_i.
   1. Select at random the centroid $\{z_1, z_2, ..., z_K\}$.
   2. Repeat until the criterion stop is satisfied J_e.
      a) Every pattern of the data set assigns the most nearby cluster

$$x \in M, \quad d(x, Z_i) \le d(x, Z_j), \quad \forall i \ne j \quad (5)$$

      b) update from the new assigned patterns

$$Z_i = E(x), \quad x \in S_i, \quad 1 \le i \le K \quad (6)$$

4 Genetic Algorithms

Evolution has proven to be a very powerful mechanism in finding good solutions to difficult problems. One can look at the natural selection as an optimization method, which tries to produce adequate solutions to particular environments.

In spite of the large number of applications of GA in different types of optimization problems, there is very little research on using this kind of approach to the clustering problem (Kivijarvi and Frati ,2003; Naldi and Carvalho, 2006; Wang, 2003). In fact, the quality of the solutions that this technique has showed in different types of fields and problems (Mitchell, 1996) it makes perfect sense to try to use it in clustering problems.

The flexibility associated with GA is one important aspect and advantage to consider. With the same genome representation and just by changing the fitness function one can have a different algorithm. In the case of spatial analysis this is particularly important since one can try different fitness functions in an exploratory phase.

5 Solving the Clustering Problem using *k*-means method

An individual ι in the genetic traditional algorithm codifies a solution ω_t. In this case the solution is given in the individual, which encodes it. The reason of this conceptual distinction between the individual and the solution is that an individual includes the information of the parameter entry of the algorithm of *k*-means.

Before starting the codification of the Genetic Algorithm, we must codify the chromosomes chain that contains all the genetic information of our system.

In this case, analyze a massive repository of information, the scheme of the information is as follows:

Table 1. Scheme of the set information

	Car_1	Car_2	.....	Car_N
X_1				
. . .	. . .	. . .	. . .	. . .
X_M				

where M is the number of samples, and N is the number of characteristics or dimensions of every sample. In the chromosome is encoded the following parameters of the *k*-means algorithm: (1) the number of clusters *k* and (2) the number of characteristics or attributes that will be used in the clustering process in the range of $[1, N]$ denominated C. Then the chromosome structure can be represented as follows,

Number of Clusters (K)	Car_1	Car_2	. . .	Car_C

With information previously described there is generated a chromosome chain ω_t, which this composed for: the number of k clusters, this number

is selected random in the range of $[1, K\_\max]$, this is to explore the space of search of the clusters, also the solution ω_t this composed by the numbers of the characteristics to using in the clustering techniques.

As example of chromosome take the maximum $K = 7$ and C = 3 of 7 possible ones, which can be represented each one by 3 bits (see Table 2).

Table 2. Example of the genotype of the chromosome chains

	Number of Cluster (K)	Car_1	Car_2	Car_3	Chromosome
1	011	110	001	101	011110001101
2	101	011	100	001	101011100001
3	111	100	101	011	111100101011

The chromosome is generated of form random, where the number of clusters $K \in [1, K\_\max]$ and $Car_i \in [1,$ maximum number of characteristics or column of the data set] for i=1,2, and 3. The following step generates his phenotype, which is the representation in decimal form.

Table 3. Example of the phenotype of the chromosome chains

	Number of Clusters (K)	Car_1	Car_2	Car_3	Chromosome
1	3	6	1	5	3,6,1,5
2	5	3	4	1	5,3,4,1
3	7	4	5	3	7,4,5,3

These characteristics are the parameters of entry to the algorithm of *k*-means.

Scheme of the algorithm proposed for the resolution of clustering problems:

1. To obtain the crossover probability (P_C) and mutation (P_M), population size (G), and maximum number of generations (T).
2. Generate G random individuals to form the initial generation.
3. Iterate the following T generations
   a) Apply *k*-means to G individuals.
   b) To obtain the statistics of the clusters of every individual G.
   c) Select G_B surviving individuals for the new generation.
   d) Select $G - G_B$ pairs of individuals as the set of parents.
   e) For each pair of parents (ι_a, ι_b) do the following:
      i) Create the solution ω_t of ι_n the offspring by crossing the solutions of the parents.
      ii) Mutate ω_t with probability pm.
      iii) To evaluate the quality of the solution ω_t in a function of fitness
      iv) Add ι_n to the new generation
   f) Replace the current generation by the new generation.
4. Output the best solution of the final generation.

In every run of the *k*-means, such statistics are obtained for ω_t like standard deviation and distances between clusters, to be able to evaluate them in function fitness.

The chromosome chain ω_t, this to be evaluated by function fitness, in this case the criteria of the clustering technique are: (1) to maximize the distance between clusters. This function can be written like:

$$distC = \frac{\sqrt{\sum_{i=1}^{K}\sum_{j=1}^{K}\left(Z_i - Z_j\right)^2}}{K} \tag{7}$$

Where $distC$ is the average of the distances of the centroids. And (2) to minimize the internal standard deviation of every cluster, this function can be written like

$$desvC = 1\Big/\frac{\sum_{i=1}^{K}\left(\sigma_i - \bar{\sigma}\right)^2}{\sum_{i=1}\sigma_i} \tag{8}$$

The $desvC$ minimizes the Sum of Squared Errors of the standard deviation of the clusters.

Finally combining functions $distC$ and $desvC$, we obtain the function of fitness

$$f(M) = \frac{\sqrt{\sum_{i=1}^{K}\sum_{j=1}^{K}\left(Z_i - Z_j\right)^2}}{K} + 1\Big/\frac{\sum_{i=1}^{K}\left(\sigma_i - \bar{\sigma}\right)^2}{\sum_{i=1}\sigma_i} \tag{9}$$

Where M is the set of patterns, before applying this function fitness, M is evaluated by one clustering techniques in this case that of k-means.

6 Test Results

The data set used in the simulated test had M=2000 samples, N=12 characteristics, only two characteristics where used to generate five clusters. Clusters were generated taking 5 random centroids. The samples were spread using a Gaussian distribution. The other characteristics were generated using a uniform distribution. In Figure 3 the original computer generated clusters are shown.

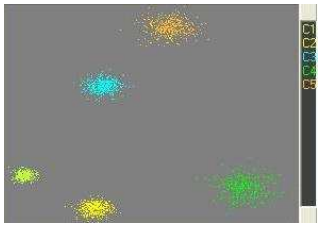

Fig3. Clusters generated by Computer simulation

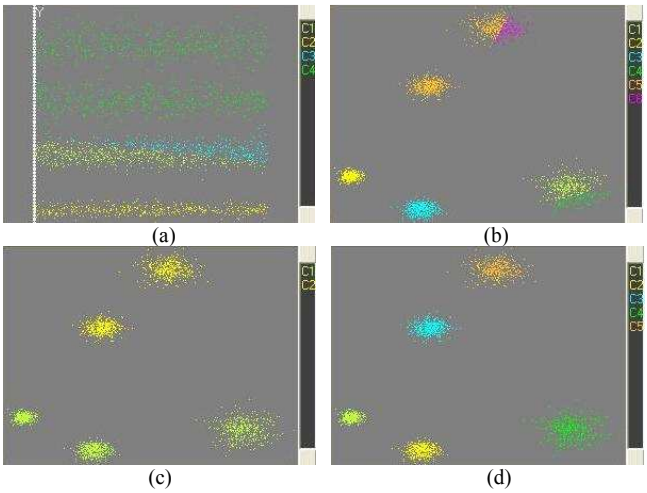

Fig. 4. Graphs of the results of k-means and the proposed algorithm, with (a) Gen 2 (b) Gen 6 (c) Gen 10 (d) Gen 35

The input parameters used in the GA, were the following: P_C=0.8, P_M=0.001, G=20, T=30, and a Boltzmann selection method where used. Figure 4 shows the results of the proposed GA technique in Generations 2, 6, 10 and 35. Finally, on Generation 35 the original clusters are recovered.

7 Conclusion

The good result of a clustering method depends to a great extent on initial parameters, in this paper we proposed a genetic algorithm that adapts the initial parameters. The GA technique is applied in k-means clustering method to determine the number of clusters, and the characteristics to take in consideration in the clustering process.

The future work consists in using different representation schemes for the GA and compares the qualities and shortcomings of the different representations.

References:

[1] Berry Michael W.: Surver of Text Mining: Clustering, Classification, and Retrieval. John Wiley & Sons (2003).

[2] Bow Sing-Tze.: Pattern Recognition and Image Preprocessing. Marcel Dekker Inc. (2002).

[3] Bradley P, Fayyad U.: Refining Initial Points for K-Means Clustering, In J. Shavlik, editor, Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann (1998).

[4] Duda Richard O, Hart Peter E.: Pattern Classification. John Wiley & Sons (2001).

[5] Goldberg David E.: Genetic Algorithms in Search Optimization and Machine Learning. Addison-Wesley Publishing (1989).

[6] Gonzalez Rafael C., Woods Richard E.: Digital Image Processing. Addison Wesley (2002).

[7] Hartigan, J.: Clustering Algorithms. Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons (1975).

[8] Huapt Randy L, Huapt Sue Ellen.: Practical Genentic Algorithm. John Wiley & Sons (2005).

[9] Jain A.k, Dubes R.C.: Algorithms for Clustering Data. Prentice-Hall (1998).

[10] Kivijarvi Juha, Frati Pasi.: Self-Adaptative Genetic Algorithm for Clustering. Journal for Heuristics, Kluwer Academic Publishers 9: 113-129 (2003).

[11] Marques de Sá J.P.: Pattern Recognition: Concept, Methods and Aplications. Springer (2001).

[12] Mitchel, Melanie.: An Introduction to Genetic Algorithms. MIT Press, London (1999).

[13] Naldi Murillo C, Carvalho André.: Partitional clustering improvement with Genetic Algorithms. (2006).

[14] Ochoa Alberto, Ponce Julio, Baltazar Rosario.: An approach to Cultural Algorithms from Data Mining. (COMCEV07) Mexican congress of Evolutionary Computation (2007).

[15] Pedrycz Witold.: Knowledge Based Clustering. John Wiley & Sons (2005).

[16] Sato M, Sato Y, Jain L.: Fuzzy Clustering Models and Applications Springer-Verlag (1997).
[17] Sikora Riyaz, Piramuthu Selwyn.: Framework for efficient feature selection in genetic algorithm based data mining. European Journal of Operational Research 180(2): 723-737 (2007).
[18] Tou Julius T, Gonzalez Rafael C.: Pattern Recognition Principles. Addison-Wesley (1974).
[19] Una-May O´Reilly, Tina Yu.: Genetic Programming Theory and Practice II. Springer (2005).
[20] Wang Chang, Zengqiang Chen, Qinlin Sun, Zhuzhi Yuan.: Clustering of Amino Acid Sequences based on K-Medoids Method. Journal of Computer Engineering, Vol.29 No.8 (2003).
[21] Wang Chang, Zengqiang Chen, Zhuzhi Yuan.: K-Means Clustering Based on Genetic Algorithm. Journal of Computer Science, Vol.30 No.2 (2003).
[22] Webb Andrew R.: Statistical Pattern Recognition Principles. John Wiley & Sons (2002).