Resolution of the Clustering Problem using Genetic Algorithms

Jose Aguilar

Abstract—Clustering is the process of grouping objects together in such a way that the objects belonging to the same group are similar and those belonging to different groups are dissimilar. In this paper we propose a method to carry out data clustering using Genetic Algorithms. We use evolutionary characteristics to define a general data clustering procedure. In addition, we present examples of application of our approach: the definition of healthcare centers or new public universities for a given Country.

Keywords—Data Clustering, Genetic Algorithms, Data Mining.

I. INTRODUCTION

M OST organizations possess large volumes of data about their business processes and resources. While this data can provide plenty of statistical information, very little useful knowledge can be procured from it. In order to gain such useful knowledge, we need to discover patterns in the data, associated with the past behavior of business processes. These patterns are used to dictate future strategy so as to maximize performance and profit. Such a knowledge discovery process is called Data Mining (DM) [6, 9]. Among the possible interesting patterns that can be discovered, those related to the discovering of clusters in data can be particularly useful [1, 6].

The data clustering is a classical activity in DM. In general, clustering is the search for those partitions that reflect the structure of an object set. Traditional clustering algorithms search only a small sub-set of all possible clustering (the solution space) and consequently, there is no guarantee that the solution found will be optimal.

In this paper we report the application of Genetic Algorithms (GAs) to the problem of clustering. GAs is one of the techniques that belong to the domain of evolutionary computation [2, 4, 7, 8]. This domain is inspired on the evolutive process of the species, in order to propose a general algorithm to solve complex problems. GAs are search algorithms based on natural genetic and selection, combining the concept of survival of the fittest with a structured interchange, but random, of the information. These concepts involve the preservation of the characteristics of the best exponents of a generation in the next generation; moreover introducing random changes in the new generation composition by means of crossing over and mutation

operations. This random component prevents getting stuck into a local maximum from which you can not escape to reach a global maximum. Our general GAs based clustering method to discover data groups has good potential for useful applications. We present in this paper two of them.

II. THEORETICAL ASPECTS

A. Data Mining and Data Clustering

DM is concerned with the discovery of interesting patterns and knowledge in large data repositories. The technology of DM (mining of data) has been gained the attention of the market [4, 6, 9, 10]. DM offers a powerful alternative to companies to discover new chances of business and to trace new strategies for the future. The tools of DM analyze the data, discover problems or chances hidden in the relationships of the data, and then diagnosis the behavior of the businesses, requiring the minimum intervention of the user. Clustering is a hard combinatorial problem and is defined as the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). Clustering is the process of grouping data into clusters so that data within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters [1, 6]. That is, the formation of clusters is based on the principle of maximizing the similarity between objects of the same cluster while simultaneously minimizing the similarity between objects belonging to distinct clusters. Similarity can be expressed in terms of a distance function, which is typically, though not necessary, a metric [6]. For example, for each pair of data objects p1, p2, the distance D(p1,p2) can be defined. In addition to a distance function, there is a separate "quality" function that measures the "goodness" of a cluster. Even though similarity between objects and goodness of clusters can be defined, it is much harder to define "similar enough" and "good enough". The answer to this question is typically highly subjective and remains an open issue in cluster analysis [8].

The clustering problem has been addressed in many contexts and by researchers in many disciplines. [14] presents an overview of pattern clustering methods from a statistical pattern recognition perspective. They present a taxonomy of clustering techniques, and identify cross-cutting themes and recent advances. They also describe some important applications of clustering algorithms such as image segmentation, object recognition, and information retrieval.

[11] deals with the different aspects of Web data mining and provides an overview about the various techniques used in this field. [12] presents a tool for database clustering using a rule-based genetic algorithm (RBCGA). RBCGA evolves individuals consisting of a fixed set of clustering rules, where each rule includes non-binary intervals, one for each feature. [15] introduces an algorithm for personalized clustering based on a range tree structure, used for identifying all web documents satisfying a set of predefined personal user preferences. [16] proposes two new approaches to using PSO to cluster data. It is shown how PSO can be used to find the centroids of a user specified number of clusters. The algorithm is then extended to use K-means clustering to seed the initial swarm. This second algorithm basically uses PSO to refine the clusters formed by K-means. A popular heuristic for k-means clustering is Lloyd's algorithm. [17] explores basic aspects of the immune system and proposes a novel immune network model with the main goals of clustering and filtering unlabeled numerical data sets. As important results of the model, the network evolved will be capable of reducing redundancy, describing data structure, including the shape of the clusters. Fuzzy C-Means (FCM) clustering algorithm cannot be used for the subsequent data (adaptive data). [13] presents an alternative adaptive FCM which is able to cope with this limitation. In [18] is introduce an approach, called the Constructive Genetic Algorithm (CGA), where the problems are modeled as bi-objective optimization problems that consider the evaluation of two fitness functions. This double fitness process, called fg-fitness, evaluates schemata and structures in a common basis. The CGA has been applied to two clustering problems in graphs. The clustering problems studied are the classical p-median and the capacitated p-median. Good results are shown for problem instances taken from the literature.

B. Genetic Algorithms

GAs might be considered as a model of learning machine whose behavior is derived from certain natural evolution mechanism [7, 8]. GA is an optimization algorithm based on the principles of evolution in biology. A GA follows an "intelligent evolution" process for individuals based on the utilization of evolution operators such as mutation, inversion, selection and crossover. The idea is to find a good local optimum, starting from a set of initial solutions, by applying the evolution operators to successive solutions. The procedure evolves until it remains trapped in a local minimum. Normally, in this method several parameters we studied: the maximum number of generations, the number of individuals on the population and the probabilities to use the evolution operators (mutation, crossover, etc.). During the execution of a GA over an individual population, which represents the solution candidates to a given problem, the population will be subjected to a set of transformations (Genetic Operators) in order to update the search. Then, it will be subjected to a selection process, which will choose the best. Each transformation+selection cycle gives rise to a generation. The

representing procedure for such a process is:
Generation of individuals which represent potential solutions
Repeat until system convergence
  Evaluation of every individual
  Selection of the best individual for Reproduction
  Reproduction of the individual using evolutionary
    operators
  Replacement of the worst old individuals by the new
    individuals

III. OUR METHOD FOR DATA CLUSTERING

We present a new clustering approach based on the use of GA in order to find the optimal grouping of data records. The proposed algorithm has several useful features: (i) it is able to distinguish relevant attribute values for the characterization of different groups of records from those that are not; (ii) it is able to efficiently handle data of high dimensionality; and (iii) it does not require any domain knowledge to work effectively. In order to use our GA based clustering algorithm, we propose the next procedure:

1. Problem Definition. In this phase we describe the problem.
2. Goals Definition: in this phase we describe the goals that we hope to solve using the Clustering Algorithm.
3. Variables Determination: in this phase we analyze the DBs of the organization to determine the files, fields, etc. that we will used.
4. Data Extraction: sometimes we have not the data on the DBs of the organization. For these cases, we need to extract these data from other sources (figures, etc.).
5. Data Integration: In this phase we define the DB that our GA will use (called work DB). In this DB we integrate the data from the different sources. For example, the different parts of the organization DBs necessary for the data clustering procedure.
6. Chromosome Definition: In general, each individual represents a solution to our problem (a possible "cluster"). The individuals must code the goals of the problem and consider: i) The variables choose from the organization DBs. ii) The keys of the organization DBs. The genes represent the attributes that describe the clusters and an individual represents a possible cluster.
7. GA Parameters Definition: In this phase we define the next set of parameters: i) Objective Function: It is defined according to the goals of the problem. Specifically, the clustering problem becomes to determine whether a given cluster (individual) groups a large number of data. That must be measured by the fitness function. ii) Convergence Criteria.
8. GAs Execution: this step corresponds to the execution of the GA and the result analyzes.

A. The GA-Based clustering algorithm

Our algorithm is composed by two phases: a Data Mining phase where we determine if we must continue to search new clusters, and an evolutionary phase where the GA proposes new clusters.

Data Mining phase
1. Define Chromosome.
2. Define GA parameters.
   Evolutionary phase
   2.1 Initialize randomly the individuals.
   2.2 Evaluate each individual using the Objective Function.
   2.3 Generate new individuals using the genetic operators and the best individuals.
   2.4 Evaluate the new individuals.
   2.5 Replace the worst old individuals by the best new individuals.
   2.6 If we have not arrived to a convergence criterion return to step 2.3.
3. Extract the best individual (this is a new cluster) and update the information on the work DB.
4. Determine if we can determine new clusters. In that case, we return to step 2.1.

With the genes of the chromosomes we search the information from the organization DBs. In this way, we can compare the information recovery for each chromosome. The chromosome that recoveries the largest number of registers from the organization DB is the best one.

IV. EXPERIMENTS

In this section we apply our approach for two clustering problems.

A. *Definition of Healthcare Center places*

We study the mortality due to violent accidents on the Merida State, Venezuela. Some of these problems are generated because there are not healthcare centers close to the place where the accidents occur [5]. For this reason, we propose a system to determine where we need to install healthcare centers and theirs medical specialties, according to the type of accidents around them.

Our approach applied to this problem

Problem: reduce the number of death due to violent accidents in Merida state.

Objectives: define new healthcare centers and theirs medical specialties.

Variables (DBs):
   DB1: (Accidents)

TABLE I
TYPE OF ACCIDENTS (T11)

Name	Description
CTA	Code of the accident.
NOM	Name of the type of accident.

TABLE II
ACCIDENT CHARACTERISTICS (T12)

Name	Description
CI	ID of the person with the accident.
UA	Place of the accident.
CTA	Code of the accident.
FA	Date of the accident.

TABLE III
CAUSE OF THE ACCIDENT (T13)

Name	Description
CTA	Code of the accident.
CC	Code of the cause.
NOM	Name of the cause.

TABLE IV
SPECIALTY BY TYPE OF ACCIDENT (T14)

Name	Description
CTA	Code of the accident.
CE	Code of the specialty.

TABLE V
CHARACTERISTICS OF THE PEOPLE (T15)

Name	Description
CI	ID.
NOM	Name of the person.
FN	Birth date.
PROF	Profession.
DIR	Address.

DB2: (Healthcare centers)

TABLE VI
SPECIALTY DEFINITION (T21)

Name	Description
CE	Code of the specialty.
NOM	Name of the specialty

TABLE VII
CHARACTERISTICS OF THE HEALTHCARE CENTERS (T22)

Name	Description
CH	Code of the healthcare center
CAP	Capacity.
LONG	Region that is cover by the healthcare centers.
CTR	Address.
NOM	Name.

TABLE VIII
SPECIALTIES BY HEALTHCARE CENTERS (T23)

Name	Description
CH	Code of the healthcare centers
CE	Code of the specialties
JE	Name of the head of the specialty.

Chromosome Structure: it must contain all the elements to describe a cluster (a new healthcare center), that is:
- Its specialties (CE1, CE2, CE3, ...).
- Where the healthcare center is placed (CTR).
- The region that is covered by the healthcare center (LONG).

In this way, the structure of an individual (ind) is:

CTR LONG CE1 CE2 CE3
Specialties

Objective Function: In our problem, an individual is better than other if the number of accidents cover by it is bigger than this other:

$$FA_{(j)} = \sum S_i$$

where:

$$S_i = \begin{cases} 1 & \text{If } [\text{dist } (T12(UA_i)), Ind_j(CTR) \\ & < Ind_j(LONG)] \text{ and } [T14(T12(CTA_i)) \\ & = Ind_j(CE_k)] \\ 0 & \text{otherwise.} \end{cases}$$

where: k = 1,. # of specialties of the individual j
i = 1 ... # of accidents.
j = 1 ... # of individuals.

Criteria of convergence:
Evolutionary phase:
- Number of iterations.
- GA gives the same clusters during different generations.

Data Mining phase:
- Until all accidents are covered by the existent healthcare centers.

The GA based Clustering Algorithm
1. Chromosome definition.
2. Repeat Until (not more new cluster)
   2.1. Population: we generate a set of individuals where CTR and LONG are determined randomly (uniform distribution). In addition, we choose randomly (uniform distribution) a set of specialty codes to be assigned to each individual.

TABLE IX
INITIAL INDIVIDUALS EXAMPLE

Individual	CTR	LONG	CE1	CE2
1	05-19	12	01	13
2	06-24	55	08	05
3	84-84	75	17	05

The places of an individual i (CTR) or an accident (UA) are described by the North and East coordinates. In this way we can compare the registers of the table 2 with the individuals and know if a given accident can be covered by a given individual (LONG). In addition, we must verify if the individual has the specialty required by the accident type.

2.2 Repeat Until (not convergence of the evolutionary cycle)

2.2.1. Compare each register j of the table 2 (all accidents store in the DB) with each individual i. That means, we verify if:
- The distance between CTRi and the place of the accident is smaller than LONGi
  If dist(CTRi ,T12(UAj)) < LONGi
- The type of accident j is covered for one of the specialties of the individual i.
  If T14(T12 (CTAj)) = CEi(k) $\forall$ k.
- If these conditions are true for the individual i, we can store the register UAj, CTAj, CEj in the temporal file i.

2.2.2 Compare each register j of the existent healthcare centers of the table T22 with each register store on the temporal file i, to verify if the accidents are covered by existent healthcare centers. That means, verify if:
- The distance between the existent healthcare center j and the place of the accident store on the temporal file i is smaller than LONGi,
  If dist(T22(CTRi) ,TEMP(UAj)) < LONGi
- If the accident type on the temporal file i is covered for one of the specialties of the existent healthcare center j,
  If T23(T22(CHi)) = TEMP(CEj)
If we find registers that verify these conditions, we delete these register from the temporal file i.

2.2.3 Count the number of registers on the temporal file i (this is the value of the fitness function for this individual) and store this value (COUNTi).

2.2.4 Reproduction:
2.2.4.1 Selection: we choose the best " T" individuals.
2.2.4.2 Repeat until create a given number of new individuals
- Crossover: we choose randomly two individuals from the "T" individuals, then we choose randomly a cross point (a field) and we exchange these parts among the individuals. We can use as cross points the fields: "CTR", "LONG" and "SPECIALITIES".
- Mutation: for the field's "CTR" and "LONG", we generate new values randomly. For the "SPECIALITIES" field we can replace some of them for existent codes of specialties.

2.2.5 Replacement: we replace the worst old individual for the best new individuals.

2.3. Extraction of the best individual (We define a constant as the minimal number of accident covered by an individual):
If COUNT (best individual) $\geq$ CONSTANT
  Update tables T22 and T23 with the information of this individual.
Else
  Stop, we have arrived to a condition where we can't create new clusters.

Simulations

We have used the DB of the institute Corporación de Salud del Edo. Mérida (this is a Venezuelan state) [5]. Each result is an average of 30 experiments. The set of values of the standard case is: Number of individuals=20, Mutation Probability=0.2, Crossover Probability=0.8, Number of generations=20, Maximal Length=80 Km., Number of specialties=20. Table 10 shows some of the results that we have obtained, (* is the result of the standard case).

TABLE X
RESULTS

Parameter Modified		Individuals Generated			
Mutation Prob.	Cross. Prob.	LONG	East	North	Specialties
0.1	0.9	75	299.620,53	909.953,50	04-05-07-09-10-13-14-16
		78	161.461,05	876.982,74	01-04-05-06-07-09-10-12-14
*0.2	0.8	46	159.394,73	852.280,25	01-02-04-05-06-08-09-10-16
		47	334.121,54	981.323,14	01-02-09-10-12
		40	221.611,94	901.741,36	02-03-07-08-09-10-12-14-16
0.3	0.7	68	272.224,81	963.245,74	05-06-08-09-10-11-13-15-16
		74	134.820,74	837.549,29	01-02-04-08-09-10-16
Constant of new groups		LONG	East	North	Specialties
05		19	216.871,36	903.569,04	01-02-03-05-06-08-09-10-12
15		75	303.237,83	966.884,94	03-07-09-10-12-16
		71	121.952,24	856.333,66	01-02-03-05-06-08-09-10-12
Number of generations		LONG	East	North	Specialties
10		56	219.650,11	909.245,64	03-04-05-07-09-10-11-13-16
50		75	131.500,51	843.908,63	02-07-09-10-11-12-13-16
		74	329.590,67	1.001.512,6	01-05-06-09-10-11-12-13-16
100		47	195.488,67	904.202,55	05-09-10-12-13
		45	310.534,14	991.622,36	01-06-09-10-12-13
		47	112.526,13	811.348,11	02-03-08-09-10-11-12-14-16
Maximal length		LONG	East	North	Specialties
40.000		74	141.510,86	842.385,35	02-03-06-09-10-12-13-14-15
		75	300.791,78	980.523,80	04-06-08-09-10-12-13-14
60.000		19	254.170,40	900.344,97	01-06-07-08-09-10-12-13-14
100.000		76	268.233,31	910.595,24	03-05-09-10-12-16

We obtain different number of clusters according to the values of the parameters, but they cover more or less the same space. According to the results, our approach defines new healthcare centers (individuals) far of the large cities or towns. If some individuals are near of some existent healthcare centers is because some of the specialties that requires the most frequent accidents aren't offered by these healthcare centers. With respect to the specialist, some of them can be found in all the individuals generated (for example, 10 (Radiology)). In this example, the execution time of our algorithm to converge on a PC is less than 15 minutes, which includes the access time to the databases (PostgreSQL), where T12 has 904.340 registers and T22 has 103 registers (they are the biggest tables and T12 corresponding registers of one month).

B. Definition of new Public Universities

In this section we use our approach to define new public universities in a given country. Our approach determine where the country need to create new universities (with the specific departments, number of professors, student capacity, etc.), according to the young people population in a region, type of economical (agricultural, tourist, industrial, etc), political (Court, Governor's Office, Mayor's Office, etc), cultural (musical, etc.) and social activities around them, among other.

Our approach applied to this problem

Objectives: define new universities
Variables (DBs):
  DB1: (Regions)

TABLE XI
TYPE OF ACTIVITY (T31)

Name	Description
CAC	Code of the Activity.
NAM	Name of the Activity.

TABLE XII
REGION ACTIVITIES (T32)

Name	Description
CI	ID of the region.
CAC	Code of the Activity.

TABLE XIII
SPECIALTY BY TYPE OF ACTIVITY (T33)

Name	Description
CAC	Code of the Activity.
CD	Code of the specialty (or department).

TABLE XIV
CHARACTERISTICS OF THE REGION (T34)

Name	Description
CI	ID of the region.
NAM	Name of the region.
ADD	Address.
NUM	Number of young in the region

DB2: (University)

TABLE XV
DEPARTMENT DEFINITION (T41)

Name	Description
CD	Code of the department.
NAM	Name of the department

TABLE XVI
CHARACTERISTICS OF THE UNIVERSITIES (T42)

Name	Description
CU	Code of the university
LONG	Zone that is cover by the university.
CTR	Address.
NAM	Name.

TABLE XVII
DEPARTMENT BY UNIVERSITIES (T43)

Name	Description
CU	Code of the university.
CD	Code of the department
JE	Name of the head of the department.
CAP	Capacity of students.
NUM	Number of professors

Chromosome Structure: it must contain all the elements to describe a cluster (a new university):

- Its departments (CD1, CD2, CD3, ...).
- Where the university is placed (CTR).
- The region that is covered by the university (LONG).

In this way, the structure of an individual (ind) is:

CTR	LONG	CD1	CD2	CD3

Departments

Objective Function: In our problem, an individual is better than other if the number of students cover by it is bigger than this other:

$$FA_{(j)} = \sum S_i$$

where:

$$S_i = \begin{cases} T34(NUM_i) & \text{If } [dist\,(T34(ADD_i)), \\ & Ind_j(CTR) < Ind_j(LONG)] \\ & \text{and } [T33(T32(CAC)_i) \\ & = Ind_j(CD_k)] \\ 0 & \text{otherwise.} \end{cases}$$

where: k = 1,. # of departments of the individual j
i = 1 ... # of regions.
j = 1 ... # of individuals.

Criteria of convergence: The same of the previous experiments.

The GA based Clustering Algorithm
1. Chromosome definition.
2. Repeat Until (not more *new cluster*)

2.1. Population: we generate a set of individuals where CTR and LONG are determined randomly. In addition, we choose randomly (uniform distribution) a set of department codes to be assigned to each individual.

TABLE XVIII
INITIAL INDIVIDUALS EXAMPLE

Individual	CTR	LONG	CE1	CE2
1	25-91	92	18	23
2	54-45	113	08	18
3	10-45	45	07	12

The places of an individual i (CTR) or a region (ADD) are described by the North and East coordinates. In this way we can compare the registers of the table 14 with the individuals and know if a given region can be covered by a given individual (LONG). In addition, we must verify if the individual (university) has the specialties (departments) required by the region.

2.2 Repeat Until (*not convergence of the evolutionary cycle*)

  2.2.1. Compare each register j of the table 14 (all regions store in the DB) with each individual i.

    - That means, we need to verify if the distance between CTR_i and the region is smaller than $LONG_i$
        If $dist(CTR_i, T34(ADD_j)) < LONG_i$

    - Then, we need to verify if the activities of the region *j* are covered for the specialties (department) of the individual *i*.
        If $T33(T32(CAC_j)) = CD_i(k) \;\; \forall\, j,k.$

    - If these conditions are true for the individual i, we can store the register ADD_j, CAC_j, CD_j in the temporal file i.

  2.2.2 Compare each register j of the existent universities of the table T42 with each register store on the temporal file i, to verify if the regions are covered by existent universities. That means, verify if:

    - The distance between the existent university j and the place of the region store on the temporal file i is smaller than $LONG_i$,
        If $dist(T42(CTR_i), TEMP(ADD_j)) < LONG_i$

    - The region activities on the temporal file i are covered for the departments of the existent universities j,
        If $T43(T42(CU_i)) = TEMP(CD_j)$

    If we find registers that verify these conditions, we delete these register from the temporal file i.

  2.2.3 With the registers on the temporal file i calculate the fitness function for this individual and store this value ($COUNT_i$).

  2.2.4 Reproduction:

    2.2.4.1 Selection: we choose the best "T" individuals.

    2.2.4.2 Repeat until create a given number of new individuals

      - Crossover: we choose randomly two individuals from the "T" individuals, then we

choose randomly a cross point (a field) and we exchange these parts among the individuals. We can use as cross points the fields: "CTR", "LONG" and "DEPARTMENTS".

- Mutation: for the field's "CTR" and "LONG", we generate new values randomly. For the "DEPARTMENTS" field we can replace some of them for existent codes of departments.

2.2.5 Replacement: we replace the worst old individual for the best new individuals.

2.3. Extraction of the best individual (We define a constant as the minimal number of regions covered by an individual):

If COUNT $_{(best\ individual)}$ ≥ CONSTANT

Update tables T42 and T43 with the information of this individual, where the capacity of student by department (T43(CAP$_i$)) is COUNT$_i$/number of departments of individual I, and the number of professor (T43(NUM$_i$)) is T43(CAP$_i$)/80.

Else

Stop, we have arrived to a condition where we can't create new clusters.

Simulations

We have used a DB about the Venezuela region [19]. Each result is an average of 30 experiments. The set of values of the standard case is: Number of individuals=50, Mutation Probability=0.1, Crossover Probability=0.8, Number of generations=15, Maximal Length=150 Km., Number of department=25. Table 19 shows some of the new universities proposed by our approach.

TABLE XIX
RESULTS

	Individuals Generated		
LONG	East	North	Specialties
75	291.870,53	900.903,50	04-07-09-10-13-14-16-20
101	243.270,40	909.000,97	01-05-07-08-09-11-15-18-24
146	191.674,73	820.080,25	01-08-14-18-20-21-22-25
127	349.111,54	980.003,14	03-05-10-21-22
86	219.761,94	911.761,36	02-05-06-09-12-15-24-25
108	262.234,81	903.265,74	05-07-09-12-15-18-21-24-25
93	144.845,74	827.249,29	01-05-08-10-16-20-21
67	51.340,86	892.325,35	02-03-06-09-10-13-16-19-21

We obtain different number of clusters (universities), but each one covers a set of different regions. The department of the Universities are defined according to the activities of the regions cover with them (for example, 10 (Anthropology) in regions with Indians population, or 16 and 20 (Industrial Engineering) in regions with industrial activities and political activities (Governor's Office)). In this example, the execution time of our algorithm to converge on a PC is less than 3 minutes, which includes the access time to the databases (PostgreSQL), where T32 has 54 registers and T42 has 18 registers (they are the main tables to access).

V. CONCLUSIONS

In this work we have developed a clustering algorithm based on GA. The investigations attempt to alleviate certain drawbacks related to the classical clustering algorithm by suggesting a flexible fitness function which takes into consideration cluster asymmetry, coverage and specific information of the study case. According to the results that we have obtained, we conclude that our approach proposes new healthcare centers (or universities) close to the places where there are more accidents and there aren't healthcare centers (more regions without universities). We must extent our procedure with other intelligent techniques to reduce the necessity of experts.

With respect to [14], in this work they need a lot of information for the statistical interpretation of the results; in our approach we present a heuristic algorithm that obtains good results without a lot of information about the problem. With respect to [12] the rule-based genetic algorithm (RBCGA) needs to define the set of rules to be used, which can be hard in some problems. With respect to [18], they have a restriction, the Constructive Genetic Algorithm (CGA) requires use bi-objective optimization functions. That can not be obtained in some problems.

REFERENCES

[1] J. Aguilar., C. Becerra. "Los Algoritmos Genéticos en Minería de Datos". *Technical Report. # 8-99,* CEMISID, Universidad de los Andes, Mérida, Venezuela, 1999.

[2] J. Aguilar, F. Hidrobo. "Algoritmos Genéticos Paralelos en problemas de Optimización Combinatoria", *Revista Técnica de Ingeniería, Universidad del Zulia*, vol. 21, pp. 47-58, 1998.

[3] J. Aguilar, M. Cerrada. "Genetic Programming-Based Approach for System Identification", *Advances in Fuzzy Systems and Evolutionary Computation, Artificial Intelligence* (Ed. N. Mastorakis), World Scientific and Engineering Society Press, pp. 329-334, 2001.

[4] J. Aguilar, F. Rivas. (eds.), *Computación Inteligente*, MERITEC, 2001.

[5] Corporación de Salud Mundial. *Clasificación de Tipos y Causa de Accidentes*, Gobernación del Edo. Mérida, Venezuela, 2003.

[6] U. Fayyad, P Shapiro, P. Smyth, R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining'*, the MIT Press, 1996.

[7] D. Goldberg. *Genetic Algorithms in Optimization and Machine Learning*, Addison-Wesley Publishing Company, 1989.

[8] T. Mitchell. *Machine Learning*, McGraw-Hill, 1997.

[9] A. Pieter, D. Zantinge. *Data Mining*, Edinburgh Gate, England, 1996.

[10] P. Shapiro, G. Frawley. *Knowledge Discovery in Databases'*, MIT Press, 1991.

[11] R. Ivancsy, F. Kovacs. "Clustering Techniques Utilized in Web Usage Mining", in *Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases*, 2006, pp. 237-242.

[12] I. Sarafis, A. Zalzala, P. Trinder. "A genetic rule-based data clustering toolkit", *in Proceeding Evolutionary Computation Congress*, 2002, pp. 1238-1243.

[13] C. Champathong, S. Wongthanavasu, "Alternative Adaptive Fuzzy C-Means Clustering", *in Proceedings of the 7th WSEAS International Conference on Evolutionary Computing*, 2006, pp. 7-11.

[14] A. Jain, M. Murty, P. Flynn. "Data clustering: a review Full text", *ACM Computing Surveys*, vol. 31, pp. 264 – 323, 1999.

[15] M. Rigou, S. Sirmakessis, G. Tzimas. "A method for personalized clustering in data intensive web applications", *in Proceedings of the Joint International Workshop on Adaptivity, Personalization & the Semantic Web*, 2006, pp. 35 – 40.

[16] D. Van der Merwe, D. Engelbrecht. "Data clustering using particle swarm optimization*", in Proceeding of the Evolutionary Computation Congress,* 2003, pp. 215- 220.

[17] L. Nunes, F. Von Zuben. "An Evolutionary Immune Network for Data Clustering", *in Proceeding of the VI Brazilian Symposium on Neural Networks (SBRN'00)*, 2000, p. 84-92.

[18] L. Nogueira, J. Furtado. "Constructive Genetic Algorithm for Clustering Problems", *Evolutionary Computation* vol. 9, pp. 154-163, 2001.

[19] Regiones de Venezuela, *Ministerio de Planificación y Desarrollo*, República Bolivariana de Venezuela, 2006.

Professor Aguilar Jose received the B. S. degree in System Engineering in 1987 (Universidad de Los Andes-Venezuela), the M. Sc. degree in Computer Sciences in 1991 (Universite Paul Sabatier-France), and the Ph. D degree in Computer Sciences in 1995 (Universite Rene Descartes-France). He was a Postdoctoral Research Fellow in the Department of Computer Sciences at the University of Houston (1999-2000). He is a Titular Professor in the Department of Computer Science at the Universidad de Los Andes. He has published more than 200 papers and 5 books, in the field of parallel and distributed systems, computational intelligence, science and technology management, etc. Dr. Aguilar has been a visiting research/professor in different universities and laboratories, has been the coordinator or inviting research in more than 20 research or industrial projects, and has supervised more than 20 M. S. and Doctoral students in their thesis.