Proceedings of the 6th WSEAS Int. Conf. on EVOLUTIONARY COMPUTING, Lisbon, Portugal, June 16-18, 2005 (pp269-274)

A Data Mining technique for Data Clustering based on Genetic Algorithm

J. Aguilar
CEMISID. Departamento de Computación
Facultad de Ingeniería. Universidad de los Andes
Mérida 5101, Venezuela

ABSTRACT

Data mining is the process of deriving knowledge from data. The data clustering is a classical activity in data mining. In this paper we propose a method to carry out data clustering using genetic algorithms. We use evolutionary characteristics to define the data clustering procedure. In addition, we present an example of application of our approach, the definition of healthcare centers for a given venezuelan region.

Introduction

Most 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]. By other hand, Genetic Algorithm (GA) 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. In this paper, we present a GA based clustering method to discover data groups. We show that the proposed method has good potential for useful applications.

2. Theoretical Aspects

2.1 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 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]. 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].

2.2. Genetic Algorithms.

GA 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
```

3. Our method for Data Clustering based on GAs

We present a new clustering approach based on the use of GAs in order to find the optimal grouping of data records. 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 DM.
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. *GA Execution:* this step corresponds to the execution of the GA and the result analyzes.

3.1 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
A. Chromosome definition.
B. GA parameters definition.
  Evolutionary phase
  B.1 Random Initialization of the individuals.
  B.2 Evaluation of each individual using the Objective Function.
  B.3 Generation of new individuals using the genetic operators and the best individuals.
  B.4 Evaluation of the new individuals.
  B.5 Replacement of the worst old individuals by the best new individuals.
  If we have not arrived to a convergence criterion we return to step B.3.
C. Extraction of the best individual (this is a new cluster) and update the information on the work DB.
D. Determine if we can determine new clusters. In that case, we return to step B.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.

4. Experiments

In this section we apply our approach for a given problem. 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.

4.1 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 1: Type of Accidents (T11)

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

Table 2: 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 3: Cause of the accident (T13)

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

Table 4: Specialty by type of accident (T14)

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

Table 5: 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 6: Specialty Definition (T21)

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

Table 7: Characteristics of the healthcare centers (T22)

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

Table 8: 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:

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

where:

$$\mathbf{S_i} = \begin{cases} 1 , & \text{If } \mathbf{dist\ (T12(UA_{\ i})),\ Ind_j\ (CTR))} \\ & \mathbf{< Ind_j\ (LONG)} \text{ and} \\ & \mathbf{T14(T12(CTA_i\ )) =} \\ & \mathbf{Ind_j\ (\ CE_{\ k}\ )} \\ 0 , & \textbf{otherwise.} \end{cases}$$

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.

4.1.1 Description of our GA based Clustering Algorithm.

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

Individual	CTR	LONG	CE1	CE2
1	**05-19**	**12**	**01**	**13**
2	**06-24**	**85**	**08**	**05**
3	**84-84**	**175**	**17**	**05**

Table 9: Initial individuals Example.

The places of an individual **i** (CTR) or an accident (UA) are described by the North and Est 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**). We call that **Zona_Ind.** In addition, we must verify if the individual has the specialty required by the accident type, we call that **Cubierto_Ind.**

- **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 CTR_i and the place of the accident is smaller than $LONG_i$:

      If $dist(CTR_i, T12(UA_j)) < LONG_i$
    - The type of accident *j* is covered for one of the specialties of the individual *i*.

      If $T14(T12(CTA_j)) = CE_i(k) \quad \forall k.$
    - If these conditions are true for the individual i, we can store the register UA_j , CTA_j , CE_j in the temporal file i.
  - **2.2.2. Compare each register i of the existent healthcare centers of the table T22 with each register store on the temporal file, to verify if the accidents are covered by existent healthcare centers. That means, verify if:**
    - The distance between the existent healthcare center i and the place of the accident store on the temporal file j is smaller than $LONG_i$,

      If $dist(T22(CTR_i), TEMP(UA_j)) < LONG_i$
    - If the accident type on the temporal file j is covered for one of the specialties of the existent healthcare center i,

      If $T23(T22(CH_i)) = TEMP(CE_j)$
    - If we find registers that verify these conditions, we delete these register from the temporal file j.
  - **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 ($COUNT_i$).**
  - **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**
      - **2.2.4.2.1. 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**".
      - **2.2.4.2.2. 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 $\mathbf{COUNT_{(best\ individual)} \geq CONSTANT}$
- 2.3.1.1. Update tables T22 and T23 with the information of this individual.

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

4.2 Simulations.

We have developed a system, called "Data Mining System based on GA" (DMSGA). We have developed this system using Visual Foxpro and ArcView Gis.

4.2.1 Results.

We have used the DB of the institute *Corporación de Salud del Edo. Mérida [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=80.000 Mts., Number of specialties=10.** Table 10 shows some of the results that we have obtained, (* is the result of the standard case).

Parameter Modified		**Individuals Generated**			
Mutation Prob.	***Cross. Prob.***	***LONG***	***Est***	***North***	***Specialties***
0.1	**0.9**	**75.517,26**	**299.620,53**	**909.953,50**	**04-05-07-09-10-13-14-16**
		78.903,07	**161.461,05**	**876.982,74**	**01-04-05-06-07-09-10-12-14**
*** 0.2**	**0.8**	**46.548,94**	**159.394,73**	**852.280,25**	**01-02-04-05-06-08-09-10-16**
		47.316,75	**334.121,54**	**981.323,14**	**01-02-09-10-12**
		40.359,28	**221.611,94**	**901.741,36**	**02-03-07-08-09-10-12-14-16**
0.3	**0.7**	**68.667,91**	**272.224,81**	**963.245,74**	**05-06-08-09-10-11-13-15-16**
		74.551,50	**134.820,74**	**837.549,29**	**01-02-04-08-09-10-16**
Constant of new groups		***LONG***	***EST***	***North***	***Specialties***
05		**109.285,67**	**216.871,36**	**903.569,04**	**01-02-03-05-06-08-09-10-12**
15		**75.237,15**	**303.237,83**	**966.884,94**	**03-07-09-10-12-16**
		71.822,18	**121.952,24**	**856.333,66**	**01-02-03-05-06-08-09-10-12**
Number of generations		***LONG***	***EST***	***North***	***Specialties***
10		**106.815,65**	**219.650,11**	**909.245,64**	**03-04-05-07-09-10-11-13-16**
50		**75.495,23**	**131.500,51**	**843.908,63**	**02-07-09-10-11-12-13-16**
		74.200,37	**329.590,67**	**1.001.512,6**	**01-05-06-09-10-11-12-13-16**
100		**47.265,28**	**195.488,67**	**904.202,55**	**05-09-10-12-13**
		45.959,56	**310.534,14**	**991.622,36**	**01-06-09-10-12-13**
		47.504,86	**112.526,13**	**811.348,11**	**02-03-08-09-10-11-12-14-16**
Maximal length		***LONG***	***EST***	***North***	***Specialties***
40.000		**74.962,88**	**141.510,86**	**842.385,35**	**02-03-06-09-10-12-13-14-15**
		75.349,39	**300.791,78**	**980.523,80**	**04-06-08-09-10-12-13-14**
60.000		**101.206,16**	**254.170,40**	**900.344,97**	**01-06-07-08-09-10-12-13-14**
100.000		**96.968,02**	**268.233,31**	**910.595,24**	**03-05-09-10-12-16**

Table 10: Results

4.2.2 Analyze.

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).

5. Conclusions

In this work we have developed a clustering algorithm based on GA. According to the results that we have obtained, we conclude: i) We need an expert to define the chromosome structure and the fitness function for the specific problem to solve. ii) According to our results, DMSGA propose new healthcare centers close to the places where there are more accidents and there aren't healthcare centers. We must extent our procedure with other intelligent techniques to reduce the necessity of experts.

References

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

[2] Aguilar J, Hidrobo F. "Algoritmos Genéticos Paralelos en problemas de Optimización Combinatoria", *Revista Técnica de Ingeniería*, Universidad del Zulia, Vol. 21, No. 1, pp. 47-58, 1998.

[3] Aguilar, J, Cerrada M. “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] Aguilar, J., Rivas, F. (eds.) “Computación Inteligente”, MERITEC, June 2001, Venezuela.

[5] Corporación de Salud Mundial. “Clasificación de Tipos y Causa de Accidentes”, Gobernación del Edo. Mérida, Venezuela, 1999.

[6] Fayyad U., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. “Advances in Knowledge Discovery and Data Mining", The MIT Press, 1996.

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

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

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

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