DGLC: a Density-Based Global Logical Combinatorial Clustering Algorithm for Large Mixed Incomplete Data

[1]José Ruiz-Shulcloper
[2]Eduardo Alba-Cabrera
[3]Guillermo Sánchez-Díaz
[1] IRIS Lab., Electrical and Computer Engineering Department, The University of Tennessee, Knoxville, USA.
[1,2]Laboratory of Pattern Recognition, Institute of Cybernetics, Mathematics and Physics, Havana, Cuba.
{recpat, ealba}@cidet.icmf.inf.cu
[3]Laboratory of Digital Image Processing, Center of Computer Research, IPN, Mexico.
gsanchez@sagitario.cic.ipn.mx, gsanchezd@hotmail.com

ABSTRACT

Clustering has been widely used in areas as Pattern Recognition, Data Analysis and Image Processing. Recently, clustering algorithms have been recognized as one of a powerful tool for Data Mining. However, the well-known clustering algorithms offer no solution to the case of Large Mixed Incomplete Data Sets. In this paper we comment the possibilities of application of the methods, techniques and philosophy of the Logical Combinatorial approach for clustering in these kinds of data sets. We present the new clustering algorithm DGLC for discovering β_0-density connected components from large mixed incomplete data sets. This algorithm combines the ideas of Logical Combinatorial Pattern Recognition with the *Density Based Notion of Cluster*. Finally, an example is showed in order to illustrate the work of the algorithm.

1. INTRODUCTION

The introduced in this paper algorithm is based on the ideas of Logical Combinatorial Pattern Recognition [1-3] combined with the concept of density-based notion of cluster introduced by M. Ester, H.P. Kriegel and others [4].

By *mixed incomplete description* of object, we understand an n-uple of nominal, ordinal and/or numerical values, i.e., $I(O) = (x_1(O),...,x_n(O))$, where $x_i(O) \in M_i$; $i = 1,...,n$; and M_i is the admissible values set of the feature x_i.. Observe that these values could be even sets of values. These sets of admissible values could be a subset of real numbers; terms of some dictionary; propositions or predicates of some artificial or natural language; functions; matrixes; and so on. In each M_i will be present a special symbol: "*", that represent the absence of a value of the feature x_i in the description of an object O (missing data). All of these types of feature could be present simultaneously in the (complete or incomplete) descriptions of objects.

For mining clustering process we will establish differences between *Data Set* (DS), *Large Data Set* (LDS) and *Very Large Data Set* (VLDS). These differences are not exclusive for clustering process, but also we can use for any other pattern recognition problem. We define as DS such collection of object's descriptions that the size of the set of descriptions plus the size of the result of the pairwise comparisons of all object's descriptions do not exceeds the available memory size. By LDS we assume the case when only the size of the set of object's descriptions do not exceeds the available memory size and by VLDS, the case that both exceed the available memory size. It is important to underline that these concepts are relative to the size of the available memory. That is, the existence of these types of data sets is a relative problem and it is an always-present problem.

2. THE β_0 DENSE CONNECTIVITY

Suppose we have a large mixed incomplete data. That is, the size of object's descriptions set plus the size of the comparison set do not fit in available memory.

In [5] we found all the β_0-connected components in this type of data set. The algorithm GLC uses the clustering criteria in order to detect connected components, for which an object O_i belongs to a cluster G, if and only if there exists an object O_j such that the similarity measure between the two objects is greater or equal than a given threshold. Defining the similarity between two objects as $\beta(O_i,O_j) \rightarrow [0,1]$, and the similarity threshold as β_0, then $O_i \in G$ iff $\exists$ $O_j \in$ G such that $\beta(O_i,O_j) \geq \beta_0$.

There are several traditional methods for calculate connected sets in a data set, in [6] a methodology is described that perform this calculation as a part of the solution for another problem. In general, these techniques have two fundamental phases: the calculation of the similarity matrix of the objects and the generation of the clustering.

The main drawback that these algorithms shows, is the necessity of calculate and store the similarity matrix, and when the number of objects in the data set grows

This paper has been partially supported by the Division of Research and Postgraduate Studies at the IPN, as well as by the Redi-project CONACYT 1999, and by CONACYT with their system of scholarships for postgraduate students.

0-7803-6359-0/00/$10.00 © 2000 IEEE

considerably, then becomes practically inefficient their application (in size and time).

The GLC algorithm generates the connected set as it is reading the objects of the data set. It does not calculate nor store the similarity matrix as the conventional methodologies that perform the same procedure. Hence, the immediate advantage is that became possible to process a considerable quantity of data without having a large amount of memory destined to the storage of comparisons between objects, therefore using only the necessary memory to store the clusters that are gone be created by the algorithm. It is important to point out that the algorithm never compares two objects more than once.

One of the problems of *Mixed Incomplete Data Mining* (MID Mining) is to find a structuralization of the object's description set.

Let given a similarity function β, a similarity threshold β_0, a natural number *MinPts*, a large MID $MI = \{O_1,\ldots,O_m,\ldots\} \subseteq M$. Let M be a dynamical universe of objects described in terms of several kinds of features x_i, with M_i as admissible values set for $i = 1,\ldots,n$.

Definition 1. An object $O \in \prod_{i=1}^{n} M_i$ has a *β_0-dense neighborhood* with respect to (wrt) β, β_0, *MinPts*, iff $|V_{\beta_0}(O)| \geq MinPts$, where $V_{\beta_0}(O) = \{O_j \in M \,|\, \beta(O,O_j) \geq \beta_0\}$. We said that O is a *dense point.*

This definition is an analog of ε-neighborhood in [4]. The principal difference is the function β which is not necessary a distance.

Definition 2. A non–empty set $C = \{O_1,\ldots,O_s\}$ is named a *β_0- chain* wrt β, β_0, iff for all $O_j \in C$, $\beta(O_j,O_{j+1}) \geq \beta_0$. In other words, $C = \{O_1,\ldots,O_s|$ for $j = 1,\ldots,s\text{-}1$ $O_{j+1} \in V_{\beta_0}(O_j)\}$.

Definition 3. A β_0- chain wrt β, β_0, minPts $C = \{O_1,\ldots,O_s\}$ is named a *β_0-dense chain* wrt β, β_0, *MinPts*, iff for all $O_j \in C$, O_j is a dense point. In other words $C = \{O_1,\ldots,O_s|$ [for $j=1,\ldots,s\text{-}1, O_{j+1} \in V_{\beta_0}(O_j)] \wedge |V_{\beta_0}(O_j)| \geq MinPts, j=1,\ldots,s\}$.

Definition 4. A non–empty set $NK = \{O_1,\ldots,O_m\} \subseteq K \subseteq M$ is named *nucleus of the β_0-dense connected component K* wrt β, β_0, *MinPts*, iff for all $O_j \in NK$ and for all $O \in M$ holds: $O \in NK$ iff there is $C = \{O_{i_1},\ldots,O_{i_s}\} \subseteq M$, a β_0-dense chain such that $O_j = O_{i_1}$, $O = O_{i_s}$, $O_{i_t} \in NK$, for $t=1,\ldots,s\text{-}1$.

Definition 5. A non–empty set $BK = \{O_1,\ldots,O_m\} \subseteq M$ is named *border of the β_0-dense connected component K* wrt β, β_0, *MinPts*, iff for all $O_j \in BK$ there is $V_{\beta_0}(O_j)$, $0 < |V_{\beta_0}(O_j)| <$ MinPts and there is $O \in NK$ such that $O \in V_{\beta_0}(O_j)$.

Definition 6. A non-empty set $K = \{O_1,\ldots,O_m,\ldots\} \subseteq M$ is named a *β_0-dense connected component* wrt β, β_0, *MinPts* iff $K = NK \cup BK$.

3. DGLC ALGORITHM

Let M be a dynamical large mixed incomplete data set, β, β_0 and *MinPts* as was before defined.

Step 1.-

a) Apply GLC algorithm to the set M:

b) When appears an object β_0-similar with O, increasing a counter associated to each object O of M while the density of O is not equal to *MinPts*. If *MinPts* is reached by the counter, then labeling this object.

Step 2.- The no labeled objects will be compared with the non compared objects into its β_0-connected components following step 1b).

At this moment the algorithm can get up all the β_0-connected components and its respective dense objects.

Step 3.- If in a given β_0-connected component all of its objects are dense object, then it is a β_0-dense connected component. Else, apply again the GLC algorithm but in this case to the set of labeled objects.

Step 4.- Compare all no labeled objects O_j with all labeled objects O_i while $\beta(O_j,O_i) < \beta_0$. If $\beta(O_j,O_i) \geq \beta_0$ then labeling O_j and put it in the same set of O_i.

The output will be the set of all β_0-dense connected components of each β_0-connected component obtained in the step 1.

All points, which do not belong to some β_0-dense connected components, is called *noise*.

4. AN EXAMPLE

The figure 1 shows the β_0-connectivity of 19 objects when $\beta_0 = 0.8$. The line that join a pair of objects mains that similarity between these objects is 0.8 or more.

Applying the DGLC algorithm, we obtain de following β_0-dense connected components wrt $\beta_0 = 0.8$ and $MinPts = 5$:

$K_1 = \{O_9, O_{10}, O_{11}, O_{12}, O_{13}, O_{14}, O_{15}, O_{16}\}$, where $NK_1 = \{O_{10}, O_{11}\}$ and

$K_2 = \{O_1, O_2, O_3, O_4, O_5, O_{17}, O_{18}, O_{19}\}$, where $NK_2 = \{O_2, O_3, O_4\}$.

In the table 1, some experimental results of applying DGLC to data sets are shown. Also it is show the time requiered to just calculate the similarity matrix, without applying any algorithm for clustering, due to for data with more than 5000 objects the required memory is larger than the RAM of the computer used in the experiments.

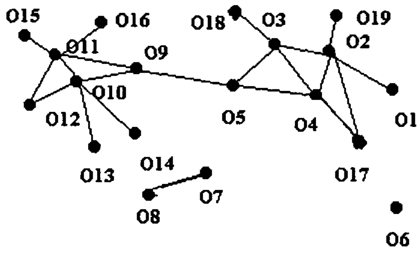

Fig. 1. The β_0-connected components of objects.

Table 1. Experimental tests with several data sets.

Number of Objects	Number of attributes	Run time of DGLC	Run time of the calculation of the similarity matrix	Generated β_0-dense connected components	Value of Min-Pts	Value of β_0
500	15	0	1	1	5	0.59
1000	15	1	2	3	5	0.65
3000	15	2	18	1	5	0.58
4000	15	6	34	1	5	0.60
5000	15	4	56	1	5	0.60
10000	15	14	233	1	5	0.60
20000	15	105	944	1	5	0.63
32561	15	188	2506	1	5	0.60
40000	55	2953	7129	20	5	0.82
50000	55	3624	11110	19	5	0.82

The experiments were implemented in C language in a personal computer based on the pentium processor at 350 Mhz, with 64 Megabytes of RAM.

Figure 2 shows the behavior of the run time required by the DGLC algorithm in order to create the β_0-denses connected components, as well as the run time required to calculate the similarity matrix of the objects by a traditional algorithm.

5. CONCLUSIONS

DGLC is the clustering algorithm which works with large mixed dataset because of its incremental behavior and the no necessity to calculate nor stored the comparison matrix. This last fact implies more computational efficient than any other algorithms calculating all connected components of a set.

REFERENCES

[1] V. Valev and Y.I. Zhuravlev, "Integer-valued problems of transforming the training tables in k-valued code in pattern recognition problems", Pattern Recognition 24, 1991, pp. 283-288.

[2] M. Lazo-Cortés and J. Ruiz-Shulcloper, "Determining the feature relevance for non-classically described objects and a new algorithm to compute typical fuzzy testors", Pattern Recognition Letters 16, 1995, pp. 1259-1265.

[3] J. Ruiz-Shulcloper and M. Lazo-Cortés, "Mathematical Algorithms for the Supervised Classification Based on Fuzzy Partial Precedence", Mathematical and Computer Modelling 29, 1999, pp. 111-119.

[4] M. Ester, H.P. Kriegel, J. Sander and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", Proceedings of the Second International Conference on Knowledge Discovery & Data Mining; Edited by E. Simoudis, J. Han, U. Fayad. August 2-4, Portland, Oregon, 1996, pp. 226-231.

[5] G. Sánchez-Díaz, J. Ruiz-Shulcloper and J.L. Díaz de León-Santiago, "GLC: Un Nuevo Algoritmo de Agrupamiento para Grandes Conjuntos de Datos Mezclados", Technical Report, Serie Roja No. 56, CIC-IPN, Mexico, 1999.

[6] J.F. Martínez Trinidad, José Ruiz Shulcloper and M. Lazo Cortés, "Structuralization of Universes", Fuzzy Sets & Systems, Vol. 112/3, 2000, pp. 485-500.

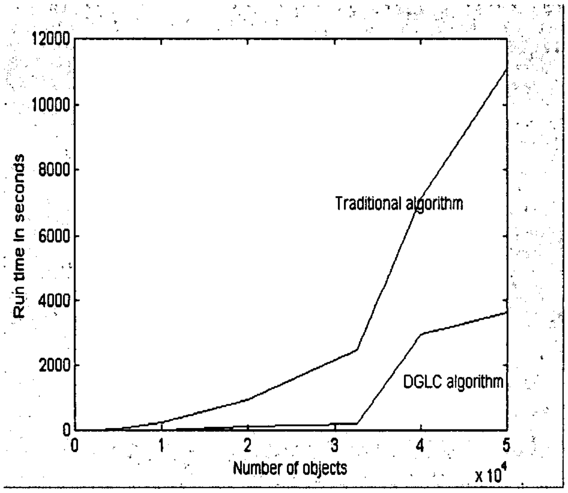

Fig. 2. Graphic objects-time requiered by DGLC and a traditional algorithm.