6th INTERNATIONAL CONFERENCE ON ADVANCED COMPUTING AND COMMUNICATION TECHNOLOGIES
IEEE COMPUTER SOCIETY DELHI SECTION AND INSTITUTION OF ELECTRONIC AND TELECOMMUNICATION ENGINEERS
Nov.-03-2012

Clustering of Mixed Data Type of Attributes Using Genetic Algorithm (Numeric, Categorical, Ordinal, Binary, Ratio-Scaled)

Mr. Aman Agarwal
(B.Tech. CSE-III Yr.)
CSE-Dept.-ABES Engg.College,
Ghaziabad (U.P.), India
(amanagarwal9891@gmail.com)

Mr. Rohit Rastogi
(Sr. Asst. Professor)
CSE-Dept.-ABES Engg.College,
Ghaziabad (U.P.), India
(rohit.rastogi@abes.ac.in)

Mr. Vipul Aggarwal
(B.Tech. CSE-III Yr.),
CSE-Dept.-ABES Engg.College,
Ghaziabad (U.P.), India
(vipul.aggarwal18@gmail.com)

Abstract

Data mining discloses hidden, previously unknown, and potentially useful information from large amounts of data. As comparison to the traditional statistical and machine learning data analysis techniques, data mining emphasizes to provide a convenient and complete environment for the data analysis. Data mining has become a popular technology in analyzing complex data. Clustering is one of the data mining core techniques.

In the field of data mining and data clustering, it is a highly desirable task to perform cluster analysis on large data sets with mixed numeric, categorical, ordinal, ratio-scaled, with binary and nominal values. However, most already available data merging and grouping through clustering algorithms are effective for the numeric data rather than the mixed data set. For this purpose, this paper makes efforts to present a new amalgamation algorithm for these mixed data sets by modifying the common cost function, trace of the within cluster dispersion matrix.

The genetic algorithm (GA) is used to optimize the new cost function to obtain valid clustering result. We can compare and analyze that the GA-based clustering algorithm is feasible for the high-dimensional data sets with mixed data values that are obtained in real life results.

Core Idea of Our Paper

By this paper, we try to describe a technique for estimating the cost function metrics from mixed numeric, categorical and other type databases by using a uncertain grade-of-membership clustering model with the efficiency of Genetic Algorithm. This technique can be applied to the problem of opportunity analysis for business decision-making.

This general approach could be adapted to many other applications where a decision agent needs to assess the value of items from a set of opportunities with respect to a reference set representing its business. For processing numeric attributes, instead of generalizing them, a prototype may be developed for experiments with synthetic and real data sets, and comparison with those of the traditional approaches. The results confirmed the feasibility of the framework and the superiority of the extended techniques.

Keywords

Clustering algorithms, categorical dataset, numerical dataset, clustering, data mining, pattern discovery, genetic algorithm.

1. Introduction

The basic operation in Data Mining is partitioning a set of objects in database into homogeneous groups or clusters. It is useful in a number of tasks, such as unsupervised classification, image processing, sequence analysis, market research, pattern recognition, spatial analysis, economics etc. Clustering is a popular and widely used approach to implement the partitioning operation. It partitions a set of objects into clusters such that objects in the same cluster are more similar to each other than objects indifferent clusters according to some defined criteria.

However, data mining, distinct from other traditional applications of cluster analysis, deals with large high dimensional data (thousands or millions of records with tens or hundreds of attributes). This characteristic stops many existing clustering algorithms from being used in data mining applications. Another important characteristic is that data in data mining often contains all types of mixed attributes in real life practical applications. The traditional way to treat categorical, nominal, ratio-scaled or ordinal attributes as numeric with the help of dissimilarity matrices (after calculating Euclidean/Manhattan/Minkowaski distances and applying normalization (standard deviation/ Z-score or min-max normalization on the results) and applying the related algorithms for numeric values, but the drawback of this process is that it does not always produce meaningful results, because many categorical domains are not ordered.

Most already available unsupervised classification algorithms either can handle both data types but are not efficient when clustering large data sets or can handle only the numeric attributes efficiently. Few algorithms can perform both well, such as k-prototypes and etc.

To process such large data sets with mixed numeric and categorical and other values, we define a new cost function for clustering by modifying the common used trace of the within cluster dispersion matrix. For minimizing the cost function (to get optimal solution) we introduce genetic algorithm (GA) in clustering process. Since GA uses search strategy globally and fits for implementing in parallel, the benefit of high search efficiency is achieved in GA based clustering process, which is suitable for clustering large data sets.

The rest of the paper is organized as follows. The next section gives some mathematical preliminaries of the

algorithm. Then we discuss the Genetic-Algorithm briefly with modified and efficient cost function for all the data sets. In last section there are summaries the discussions.

2. Betterment by the Use of Genetic Algorithm

With the basic features of GA like encoding, crossover, mutation, appropriate fitness function and reproduction with survivor selection, the GA can be able to design better clustering and unsupervised classification operations.

The proposed approach can be described with experiments and their results. The algorithm can be run on real-life datasets to test its clustering performance against other algorithms. At the same time, its properties are also empirically studied. One observation from the above analysis is that our algorithm's computation complexity is determined by the component clustering algorithms. So far, many efficient clustering algorithms for large databases are available, it implicate that our algorithms will effective for large-scale data mining applications, too.

3. The definition of cost function

Cost function is a function that determines the amount of residual error in a comparison and needs to be minimized in optimization experiment. Let X=$\{x_1, x_2, \ldots\ldots, x_n\}$ denote a set of n objects and $x_{i=}$ $\{x_{i1}, x_{12}, \ldots\ldots, x_{im}\}^T$ be an object represented by m attribute values. Let k be a positive integer. The objective of clustering **X** is to finds a partition that divides objects in **X** into k disjoint clusters.

For a given number of objects n, the number of possible partitions of the object-set is definite but highly large. It is not practical to investigate every partition in order to find a better one for a classification problem. A common solution is to choose a clustering criterion to guide the search for a partition. A clustering criterion is called cost function**.**

4. Cost Function for Numeric Data Clustering

The widely used cost function is the trace of the within cluster dispersion matrix. One way to define the cost function is

$$C(W) = \sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij}^2 (\mathrm{d}\,(x_j,\ x_i))^2 \ , w_{ij} \epsilon \{0,1\} \quad (1)$$

Here,$w_{ij}$is the membership degree of $x_j$belonging to cluster i . **W** is a $k \times n$ order partition matrix. The function d(.) is a dissimilarity measure often defined as the Euclidean distance. For the data set with real attributes, i.e., X $\subset$ R^m, we have

$$\mathrm{d}\,(x_j, x_i) = \left(\sum_{l=1}^{m} | x_{jl} - x_{il} \,|\, 2\right)½ \quad (2)$$

Since, w_{ij}indicates $x_j$belonging to cluster i, and$w_{ij} \epsilon$ [0,1], we call **W** to be a hard k-partition.

5. Cost Function for Mixed Data Clustering:

5.1 Max-Min Normalization for numeric data

For clustering the numeric data, first we will normalize numeric data so as to prevent the dominance of particular attribute. For which the normalization formula is as follows:-

$$\mathrm{n_i} = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)} \times (Rh - Rl) + Rl \quad (3)$$

Where,x_i is the i-th object. Rh and Rl are the higher and lower ranges respectively. N is the new normalized matrix containing all types of data.

5.2 Normalizing ratio-scaled values:-

First, we will take log of the ratio-scaled values, given as

$$f(n) = \log(n) \quad (4)$$

5.3Normalizing ordinal values:-

First we assign ranks to the values as, better the value higher the rank and vice versa. Now, based on their ranks we will normalize them. Give 1 to the highest rank and 0 to the lowest one and other ranks get the value as:

$$\prec(\mathrm{r}) = \frac{1}{no.of\ different\ ordinal\ values - 1} \times (\mathrm{r\text{-}1}) \quad (5)$$

5.4 Normalizing categorical values:-

If the two values match put value 1 and otherwise 0.

$$\delta(a, b) = \begin{cases} 0 & a = b \\ 1 & a \neq b \end{cases} \quad (6)$$

6. Re-defining cost function:-

When N has attributes with numeric and mixed values, assuming that each object is denoted byn_i=[$n_{i1}^r, \ldots\ldots, n_{it}^r, n_{i,t+1}^c, \ldots\ldots, n_{im}^c, n_{i,m+1}^b, \ldots\ldots, n_{i,y}^b, n_{i,y+1}^o, \ldots\ldots, n_{iu}^o, n_{i,q+1}^{rs}, \ldots\ldots, n_{is}^{rs}$,],the dissimilarity between two mixed-type objects $n_i$and $n_j$can be measured by the following Eq.(7)

$$d(n_i, n_j) = [\left(\sum_{l=1}^{t} |n_{il}^r - n_{jl}^r|\right)2 + \prec_1 . \left(\sum_{l=t+1}^{m} |n_{il}^c,\ n_{jl}^c|\right)2 + \prec_2 . \left(\sum_{l=m+1}^{y} |n_{il}^b - n_{jl}^b|\right)2 + \prec_3 . \left(\sum_{l=y+1}^{u} |n_{il}^o - n_{il}^o|\right)2\Bigg) + \prec_4 . \sum_{l=y+1}^{u} |n_{il}^{rs} - n_{il}^{rs}\,|2)]½$$

where all the terms are squared Euclidean distance measure on the mixed attributes.

Using Eq. (7) for mixed-type objects, we can modify the cost function of Eq. (1) for mixed data clustering. In addition, to extend the hard k-partition to fuzzy situation, we further modify the cost function for fuzzy clustering as:

$$C(W) = \sum_{I=1}^{K} \Big(\sum_{j=1}^{n} w_{ij}^2 \sum_{l=1}^{t} | x_{jl}^r - p_{il}^r |2 + \prec_1 \sum_{j=1}^{n} w_{ij}^2 \sum_{l=t+1}^{m} |x_{jl}^c - p_{il}^c|2 + \prec_2 . \sum_{j=1}^{n} w_{ij}^2 \sum_{l=m+1}^{y} |n_{il}^b - n_{jl}^b\,|2 + \prec_3 . \sum_{j=1}^{n} w_{ij}^2 \sum_{l=y+1}^{u} |n_{il}^o - n_{il}^o\,|2\Big) + \prec_4 . \sum_{l=y+1}^{u} |n_{il}^{rs} - n_{il}^{rs}\,|2), \qquad w_{ij}\ \varepsilon\ [0,1] \quad (8)$$

Let $C_i^r = \sum_{j=1}^{n} w_{ij}^2 \sum_{l=1}^{t} | x_{jl}^r - p_{il}^r | 2$

$$C_i^c = \prec_1 \sum_{j=1}^{n} w_{ij}^2 \sum_{l=t+1}^{m} |x_{jl}^c - p_{il}^c| 2 \quad C_i^b$$
$$= \prec_2 . \sum_{j=1}^{n} w_{ij}^2 \sum_{l=m+1}^{y} |n_{il}^b - n_{jl}^b| 2 \quad C_i^o$$
$$= \prec_3 . \sum_{j=1}^{n} w_{ij}^2 \sum_{l=y+1}^{u} |n_{il}^o - n_{il}^o| 2 \quad (9)$$

We rewrite Eq.(8) as:

$$C(W) = \sum_{i=1}^{k} (C_i^r + C_i^c + C_i^b + C_i^o) \quad (10)$$

7. GA-based clustering algorithm for mixed data

To obtain the optimal fuzzy clustering of the large data set with mixed values, genetic algorithms are employed to minimize the cost function. Since GA is a global search strategy in random fashion, it has high probability to achieve the global optima. Moreover, GA is fit for implementation in parallel, so GA-based clustering algorithm will be suitable for large data set.

8. Genetic algorithm

Genetic algorithm is a search strategy based on the mechanism of natural selection and group inheritance in the process of biology evolution. It simulates the cases of reproduction, mating and mutation in reproduction. GA looks each potential solution as an individual in a group (all possible solutions), and encodes each individual into a character string. By a pre-specified objective function, GA can evaluate each individual with a fitness value. In the beginning, GA generates a set of individuals randomly, then some genetic operations, such crossover, mutation and etc., are used to perform on these individuals to produce a group of offspring. Since these new individuals inherit the merit of their parents, they must be better solution over their predecessors. In this way, the group of solution will evolve toward more optimal direction.

9. GA-based Clustering Algorithm

9.1 Algorithm:

Step1. Begin
Step2. Define pop-size as desired population size
Step3.Randomly initializes pop-size population
Step4.While (Ideal best found or certain number of generations met)
O Evaluate fitness
O While (number of children=population size)
O Select parents
O Apply evolutionary operators to create children
O End while
Step5. End While
Step6. Return Best solution
Step7. End

To employ GA to solve the clustering, the following three problems should be settled first.

(1) How to encode the clustering solution into the gene string?

(2) How to design a reasonable fitness function for our clustering problem?

(3) How to select or design genetic operators including their parameters to guarantee fast convergence.

9.2 Encoding: From the cost function in Eq. (1) and (8), it is clear that the objective of clustering is to obtain a (fuzzy) partition matrix **W.** Then using the fitness function (stated below) we can improve the chances of a particular data point to be chosen. Then after selecting that particular cluster we can further subdivide the data points in the cluster, based on their fitness values.

Note that since we process data with mixed attributes, besides the numeric parameters, there are other mixed parameters in gene string. Due to this, it is not ordered for the binary attributes; they can be directly encoded rather than should be normalized first.

9.3 Fitness function: We are taking the fitness function such that fitness value is inversely proportional to the cost function value, i.e., the smaller the cost function is, the better the fuzzy clustering partition. For this case, the GA asks for a bigger fitness value. Hence, we define the fitness function by using the clustering cost function. Exponentially increased cost function will sharply reduce the fitness function.

$$f(g) = \frac{1}{1+e^{C(W)}} \quad (11)$$

9.4 Genetic operators: Our GA-based clustering algorithm involves all the basic genetic operators, such as selection, reproduction, crossover and mutation. What we need to do is to specify the probability parameters for each operator. For the N individuals in a generation of population, we sort their fitness value in ascending order and label each individual with its order. The selection probability is specified as:

$$P_s(g_{(i)}) = \frac{f(g_i)}{\sum_{i=1}^{n} f(g_i)} \quad (12)$$

The operation probabilities for crossover and mutation are adaptively assigned as Eq. (13)

$$P_c(g_i, g_j) = \begin{cases} \dfrac{\alpha_1 (f_{max} - f')}{f_{max} - \bar{f}} & f' \geq \bar{f} \\ \alpha_2 & otherwise \end{cases}$$

$$P_m(g_i) = \begin{cases} \dfrac{\alpha_3 (f_{max} - f(g_i))}{f_{max} - \bar{f}} & f(g_i) \geq \bar{f} \\ \alpha_4 & otherwise \end{cases} \quad (14)$$

where, $f_{max} = max_{l=1}^{N} \{f(g_l)\}$

$\overline{f} = \sum_{l=1}^{N} f(g_l)$, $f' =, f(g_j)\}$, and $\alpha_i \in [0, 1]$

Apart from above operators, we define a new operator for the clustering algorithm,

Gradient operator. The changes in the existing weights are done as per the formula:

The gradient operator includes two steps iteration as:

$$w_{ij} = \sum_{l=1}^{k} \frac{\left(d(x_j, x_i)\right)2}{\left(d(x_j, x_i)\right)2}, \forall i, \text{j} \qquad (15)$$

10. A Real-Life Practical sample data table of mixed data types:

We are representing the real life concept of our approach by taking the data of 5 employess working in a company. Here we will use every kind of data (related to all data types) to show that our method works for every kind of data.In this example :

We are taking the weighted matrix (W_{ij}) as:

	1	2	3	4	5
1	0				
2	0.4	0			
3	0.2	0.2	0		
4	0.1	0.3	0.2	0	
5	0.5	0.2	0.1	0.4	0

(Table 1)

Test-1 cantains salary of an employee (numeric data)

Test-2 shows whether the employee is male or female (binary data- Male=1/ Female=0)

Test-3 shows the department to which employee belong (categorical data)

Test-4 depicts the ability of an employee (ordinal values) Exc.-Excellent, Fair or Good

Test-5 shows avg. credit points alloted according to their performance (ratio-scaled values)

Last Column shows the log value of ratio-scaled data type.

Object- id	Test-1	Test-2	Test-3	Test-4	Test-5	Log
1	25K	M	Code-A	Exc.	445	2.65
2	40K	F	Code-B	Fair	22	1.34
3	55K	M	Code-C	Good	164	2.21
4	27K	M	Code-A	Exc.	1210	3.08
5	53K	F	Code-B	Fair	38	1.58

(Table 2)

The Table 2 is converted into the normalized matrix using the above equations Eq.(3), Eq.(4), Eq.(5), Eq.(6)

	1	2	3	4	5
1	0	0	0	0	0
2	0.5	1	1	1	1.31
3	1	0	1	0.5	0.44
4	0.0666	0	0	0	0.43
5	0.9333	1	1	1	1.07

(Table 3)

We calculate the value of the expression (stated below) to be further used in Eq. (8)

$$\sum_{j=1}^{n} d(x_j, x_i)$$

	1	2	3	4	5
1	0				
2	4.9961	0			
3	2.4436	2.2569	0		
4	0.1893	3.962	2.1213	0	
5	5.0159	0.2453	1.6153	4.1607	0

(Table 4)

Now for Eq. (8) we are calculating the value of the expression:

$$\sum_{j=1}^{n} w_{ij}^2 \left(d(x_j, x_i)\right) 2$$

	1	2	3	4	5
1	0				
2	0.0799	0			
3	0.0977	0.0903	0		
4	0.0018	0.3566	0.0848	0	
5	1.2539	0.0098	0.0165	0.6571	0

(Table 5)

Now we will calculate the expression:

$$\sum_{i=1}^{k}\sum_{j=1}^{n} w_{ij}^{2}\left(d(x_j, x_i)\right)2$$

1	**1.3455**
2	**0.5366**
3	**0.2013**
4	**1.1063**
5	**1.9373**

(Table 6)

Now using the Eq. (11) we find the fitness value of the above calculated values (above 5 tuples):

1	**0.2066**
2	**0.3689**
3	**0.4498**
4	**0.2497**
5	**0.1259**

(Table 7)

First we arrange the above values in ascending order and label each one of them and then using Eq.(12) calculate the selection probability $P_s(g_{(i)})$

1	**0.1474**
2	**0.2633**
3	**0.3204**
4	**0.1782**
5	**0.0898**

(Table 8)

11. Analysis on our Experimental Results

By the above calculated tables, we can easily verify the dissimilarity matrices of our real life experimental data shown in tabular structure,

We can comfortably decide the set of clusters based on the fitness values. We are taking the threshold value for our method to be 0.22. Data item 1 and 5 whose fitness value lie below the threshold value can be grouped together in the cluster and the other three tuples can be grouped in another. Now these clusters can be improved using GA and using the selection probability.

Result:

So there can be two clusters:

C1:- data items 1 and 5.

C2:- data items 2, 3 and 4

12. Conclusions

Here we have presented the genetic algorithm to cluster large data sets. The clustering performance of the algorithm can be evaluated using a large data set. The proposed results can be used to demonstrate the effectiveness of the algorithms in discovering structures in data.

The emphasis of this paper is put on the issue that employs the genetic algorithm to solve the clustering problem. Though the application is specific for the business, our approach is general purpose and could be used with a variety of mixed-type databases or spreadsheets with categorical, numeric and other data values, and temporal information. With improved metrics, artificial intelligence algorithms and decision analysis tools can yield more meaningful results and agents can make better decisions.

This approach, then, can ultimately lead to vastly improved decision-making and coordinating among business units and agents alike. If a class attribute is involved in the data, relevance analysis between the class attribute and the others (or feature selection) should be performed before training to ensure the quality of cluster analysis. Moreover, most variants of the GA use Euclidean-based distance metrics. It is interesting to investigate other possible metrics like the Manhattan distance or Cosine-correlation in the future. To faithfully preserve the topological structure of the mixed data on the trained map, we integrate distance hierarchy with GA for expressing the distance of categorical values reasonably.

13. Acknowledgement

The authors would like to thank the reviewers for their valuable suggestions. They would also like to thank Prof. A.K. Sinha (Dean CRAP-ABES-Engineering College, Ghaziabad) for his involvement and valuable suggestions on soft-computing in the early stage of this paper.

14. References

[1] LI Jie, GAO Xinbo, JIAO Li-cheng,"A GA-Based Clustering Algorithm for Large Data Sets with MixedNumeric and Categorical Values",National Key Lab. of Radar Signal Processing, Xidian Univ., Xi'an 710071, China

[2] M. R. Anderberg. Cluster Analysis for Applications. Academic Press, New York,1973.

[3] B. Everitt. Cluster Analysis. Heinemann Educational Books Ltd., 1974.

[4] Zhexue Huang, Michael K.Ng. A fuzzy *k*-modes algorithm for clustering categorical data. IEEE Trans. on Fuzzy Systems, 7(4): 446-452, August, 1999.

[5] Zhexue Huang. A fast clustering algorithm to cluster very large categorical data sets in data Mining. Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Dept. of Computer Science, the University of British Columbia, Canada, pp.1-8.

[6] R. Krovi. Genetic Algorithm for Clustering: A Preliminary Investigation. IEEE press, Pp.504-544.

[7] J. H. Holland. Adoption in Natural and Artificial System. Ann Arbor, MI: Univ. Mich. Press, 1975.