Special Issue of BIJIT on Fuzzy Logic
BIJIT - BVICAM's International Journal of Information Technology
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA)

Business Analysis and Decision Making Through Unsupervised Classification of Mixed Data Type of Attributes Through Genetic Algorithm

Rohit Rastogi[1], Saumya Agarwal[2], Palak Sharma[3], Uarvarshi Kaul[4] and **Shilpi Jain[5]**

Submitted in July, 2013; Accepted in February, 2014

Abstract - Grouping or unsupervised classification has variety of demands in which the major one is the capability of the chosen clustering approach to deal with scalability and to handle the mixed variety of data set. There are variety of data sets like categorical/nominal, ordinal, binary (symmetric or asymmetric), ratio and interval scaled variables. In the present scenario, latest approaches of unsupervised classification are Swarm Optimization based, Customer Segmentation based, Soft Computing methods like Fuzzy Based and GA based, Entropy Based methods and hierarchical approaches. These approaches have two serious bottlenecks…Either they are hybrid mathematical techniques or large computation demanding which increases their complexity and hence compromises with accuracy.

It is very easy to compare and analyze that unsupervised classification by Genetic Algorithm is feasible, suitable and efficient for high-dimensional data sets with mixed data values that are obtained from real life results, events and happenings.

Index Terms – Clustering, Clustering Algorithms, Categorical Dataset, Numerical Dataset, Data Mining, Pattern Discovery, Genetic Algorithm.

Core Idea of Our Paper

The proposed methodology deals with this problem with a new and efficiently generated Genetic Algorithm approach. It shapes into a better, lesser complex and computationally demanding pseudo codes which may in future lead into a revolutionary approach. We work upon multivariate data sets. In case of nominal variables, we calculate the Genetic Fitness Function criterion of each data item to decide their parent cluster and quantify the dataset by different methods to construct combined category quantifications and also plot the object scores. Here we propose an iterative procedure to calculate the cluster centers. To support our approach, a numerical experiment has been demonstrated. For all mixed variety of attributes Binary, Interval and Ratio scaled with ordinal type attributes, an efficient methodology by GA approach has been designed and has been suggested in this paper

[1, 2, 3, 4, 5]*ABES Engineering College, Ghaziabad (U.P.), India*
E-mail: [1]rohit.rastogi@abes.ac.in,
[2]mail4somya.ag@gmail.com,[3]sharma.palak595@ymail.com,
[4]saniakaul27@gmail.com and [5]shilpijain474@gmail.com

1.0 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 helps to partition objects-set into similar groups called as clusters in such a manner that objects in the same cluster are more similar to each other than objects in different clusters as per the same defined criteria.

Its major requirement is in the data mining concept to deal with high dimensionality for unsupervised classification distinct from other traditional applications of cluster analysis which may have thousands or millions of records with tens or hundreds of attributes. By this characteristic, many existing clustering algorithms stopped from being used in data mining applications. Data may be collected in the data mining sector which may contain 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 may produce useless results sometimes because most of the categorical attributes 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.

In order to handle large data sets with variety of mixed numeric and categorical and other data values, a new cost function may be defined for clustering by modification of the commonly 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 very much favorable for unsupervised classification of large data sets.

The rest of the paper is organized as follows. Forthcoming Section covers some literature survey and concluding remarks over some contemporaries unsupervised classification algorithms. Then 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. Last section summarizes the discussion.

2.0 LITERATURE SURVEY

LI Jie et al.[1] The authors writers have depicted in the paper and proposed a novel clustering algorithm for the mixed data sets. They have successfully shown the modification of the common cost function, trace of the within cluster dispersion matrix.

- The results have demonstrated the effectiveness of the algorithms in discovering structures in data. The scalability tests have shown that the algorithm is efficient when clustering large complex datasets in terms of both the number of records and the number of clusters. These properties are very important to data mining.
- The emphasis of this paper is put on the issue that employs the genetic algorithm to solve the clustering problem.
- However, in using this algorithm to solve practical data mining problems, they still faced the common problem:
- How many clusters are in the data? This will be a challenging topic for further research.
- A novel clustering method with network structure based on clonal algorithm.
- By analyzing the neurons of the obtained network with minimal spanning tree, one can easily get the cluster number and the related classification information. The test results with various data sets illustrate that the novel algorithm achieves more effective performance on cluster analyzing the data set with mixed numeric values and categorical values.
- The distance measure of their clustering algorithm indicated that the smaller the distance measure is, the better the clustering partition. For this case, the clonal algorithm asks for a bigger affinity value. Hence, they have defined the Ab-Ag affinity function by using the dissimilarity measure.

$$f(x_j, p_{ig}) = \frac{1}{1 + \sum_{i=1}^{t} x_{jl}^{r} - p_{ig,l}^{r}|^2 + \lambda \cdot \sum_{l=t+1}^{m} \delta(x_{il}^{c}, p_{ig,l}^{c})} \qquad i = 1,2,\cdots c, \quad g = 1,2,\cdots i_n$$

……..Eq. 2.1

The Ab-Ab affinity is defined as:

$$D_{ij} = \left\| p_{ig} - p_{jl} \right\| \qquad i, j = 1,2,\cdots c \quad g = 1,2\cdots i_n \quad l = 1,2,\cdots j_n$$

Where, $\|\cdot\|$ is any a norm; $D = (D_{ij})_{N \times N}$ is antibody-antibody affinity matrix, and $N = \sum_{i=1}^{c} i_n$ is number of neurons of networks. …….

.Eq. 2.2

- Since the new algorithm combines the clonal selection algorithm and the forbidden clone operator, the obtained network has not only the specificity but also the tolerance of immunity. The experimental results illustrate that the novel algorithm can effectively explore the cluster structures of the data set.
- Moreover, it does not depend on the prototype initialization and the priori information of cluster number, which makes it as a real unsupervised learning.
- Very popular and effective methodology in this context is a CSA-based clustering algorithm for large data sets with mixed numeric and categorical values.
- They have successfully shown the modification of the common cost function, trace of the within cluster dispersion matrix. The clonal selection algorithm (CSA) can be used to optimize the new cost function. By the Experiments, results illustrate that the CSA-based new clustering algorithm is suitable for the large data sets with mixed variety of the numeric and categorical values.
- For clustering analysis on the large data set with mixed numeric and categorical attributes, the CSA-based algorithm not only has a high convergence speed, but also is independent on the initialization of the prototypes and can converge to the global optimum with the probability of 1.These properties are very important to the applications of data mining.
- In this research article, authors have tried to present a data mining algorithm based on supervised clustering to learn data patterns and use these patterns for data classification. By this approach, they have proposed a scalable incremental learning of patterns from data with both numeric and nominal variables. There are two different methods of combining numeric and nominal variables in calculating the distance between clusters are investigated.
- One method suggests, numeric and nominal variables are combined into an overall distance measures and now separate distance measures are calculated for numeric and nominal variables, respectively.
- Second method says, nominal variables are converted into numeric variables, and then a distance measure is calculated using all variables.
- The prediction accuracy and reliability of the algorithm were analyzed, tested, and compared with those of several other data mining algorithms and then analyzed the computational complexity, and thus, the scalability, of the algorithm, and tested its performance on a number of data sets from various application domains.
- CCAS is a supervised clustering and classification algorithm which has been extended a scalable, incremental,

and—into ECCAS that has the capacity of handling data with both numeric and nominal variables.

- Two different methods of handling mixed data types are developed. The two methods of ECCAS are tested and compared on a data set with mixed variable types for intrusion detection. Both methods produced comparable performance to that of the winning algorithm in a data mining contest on the same data set.
- The ECCAS algorithm and the distance measure could be used in common data mining applications. The authors have developed methods to adaptively and dynamically adjust the parameters during training, including the grid-interval configuration and the threshold-controlling outlier removal.
- ECCAS (A) was also tested on two other data sets for medical diagnosis and salary prediction applications with comparable performance to those of other data mining algorithms applied to these data sets. The performance on different data sets showed the reliability of ECCAS. The testing results for one data set also showed that the five phases of ECCAS reduces the impact of the data presentation order on the prediction accuracy. The number of grid intervals showed the impact on the prediction accuracy of ECCAS. In this study, they tested different numbers of grid intervals empirically.[1]

JiangXi et al. [2] the authors represented their research ideas.
Strength- K-Prototype is one of the important and effective clustering analysis algorithms to deal with mixed data types. This article discussed fuzzy clustering algorithm based on K-Prototype in detail and made improvements to solve its initial value problems. The proposed method is simple, easy to understand and can be achieved easily.

Technology Gist

In order to improve the randomness of selecting initial cluster centers and enhance the stability of algorithm results, one selection method is proposed as follows: assume it will be divided into k clustering, consider in two steps:

(1) Value- The normalized treatment of source data set makes the numerical data in interval [0, 1]. The purpose is to prevent too much weight paralleling numerical attribute with categorical attribute, and reduce differences of the dissimilarity between the two attributes. Select equal division interval point for numerical attribute: 1 / (k + 1), 2/(k + 1), 3 / (k + 1), ... , k / (k + 1) corresponding to k cluster centers of numerical attributes.

(2) Categorical Attributes- The aim is to divide the dataset into k portions. The data number in each portion is [n / k] and put the rest into the last division directly. Then, find the value with the highest frequency in each division. It is also the mode corresponding to the categorical attributes of K initial cluster centers. Finally, combine the two parts into k initial cluster center.

Advantages and Disadvantages with Other Approaches

The problem of selecting initial cluster center is improved by analyzing fuzzy K-Prototype. This algorithm solves the problem of clustering of mixed data, reduces iteration times in the clustering process and improves the quality and efficiency.

Limitations-

Further research need to be done, such as the existing local convergence to optimize the clustering result, distributed process about massive data and so on.

3.0 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. To test its clustering performance against other algorithms, the algorithm can be run on real-life datasets. At the same time, its properties are also empirically studied. One property 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, to implicate that our algorithms will effective for large-scale data mining applications, too.

4.0 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={$x_1, x_2, \ldots\ldots, x_n$} define a set of n objects and $x_{i=}$ {$x_{i1}, x_{12}, \ldots\ldots, x_{im}$}^{T} be an object and there are m attribute values. Let k be a positive integer. The objective of clustering **X** is to find a partition that divides objects in **X** into k disjoint clusters. A simple way to solve it is to choose a clustering criterion to guide the search for a partition. A clustering criterion is called cost function**.**

Let the no. of possible partitions of the definite but highly large objects be n. Here we have to investigate every partition in order to find a better one for a classification problem.

5.0 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

$$C(W) = \sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij}^2 (d(x_j, x_i))^2 \quad , w_{ij} \epsilon \{0,1\} \quad \textbf{Eq. 5.1}$$

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

$$d(x_j, x_i) = \left(\sum_{l=1}^{m} |x_{jl} - x_{il}|^2\right)^{1/2}$$ **Eq. 5.2**

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.

6.0 COST FUNCTION FOR MIXED DATA CLUSTERING

6.1Max-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:-

$$n_i = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)} \times (Rh - Rl) + Rl$$ **Eq. 6.1**

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.

6.2 Normalizing ratio-scaled values:-

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

$$f(n) = \log(n)$$ **Eq. 6.2**

6.3Normalizing 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:

$$\varkappa(r) = \frac{1}{no.of different ordinal values - 1} \times (r-1)$$ **Eq. 6.3**

6.4 Normalizing categorical values:-

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

$$\delta(a,b) = \begin{cases} 0 & a = b \\ 1 & a \neq b \end{cases}$$ **Eq. 6.4**

6.5 Re-defining cost function:-

When N has attributes with numeric and mixed values, assuming that each object is denoted by=[

,

$$[n_{i1}^r, \ldots, n_{it}^r, n_{it+1}^s, \ldots, n_{im}^s, n_{im+1}^b, \ldots, n_{iy}^b, n_{iy+1}^o, \ldots, n_{iu}^o, n_{iu+1}^{rs}, \ldots, n_{in}^{rs}]$$

],

the dissimilarity between two mixed-type objects n_i and n_j can be measured by the following **Eq. 6.5**

$$d(n_i, n_j) = [\left(\sum_{l=1}^{t} |n_{il}^r - n_{jl}^r|^2 + \varkappa_1 \cdot \left(\sum_{l=t+1}^{m} |n_{il}^s, n_{jl}^s|^2 + \varkappa_2 \cdot \left(\sum_{l=m+1}^{y} |n_{il}^b - n_{jl}^b|^2 + \varkappa_3 \cdot \left(\sum_{l=y+1}^{u} |n_{il}^o - n_{jl}^o|^2\right)\right)\right)\right) + \varkappa_4 \cdot \sum_{l=y+1}^{u} |n_{il}^{rs} - n_{jl}^{rs}|^2)]^{1/2}$$

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

Using Eq. 6.5, for mixed-type objects, we can modify the cost function of Eq. 5.1, 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:

$$C(W) = \sum_{i=1}^{u}\left(\sum_{j=1}^{n} w_{ij}^2 \sum_{l=1}^{t} |x_{jl}^r - p_{il}^r|^2 + \varkappa_1 \sum_{j=1}^{n} w_{ij}^2 \sum_{l=t+1}^{m} |x_{jl}^s - p_{il}^s|^2 + \varkappa_2 \cdot \sum_{j=1}^{n} w_{ij}^2 \sum_{l=m+1}^{y} |n_{il}^b - n_{jl}^b|^2 + \varkappa_3 \cdot \sum_{j=1}^{n} w_{ij}^2 \sum_{l=y+1}^{u} |n_{il}^o - n_{il}^o|^2) + \varkappa_4 \cdot \sum_{l=y+1}^{u} |n_{il}^{rs} - n_{il}^{rs}|^2)$$

, $w_{ij} \varepsilon [0,1]$ **Eq. 6.6**

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

$$C_i^s = \varkappa_1 \sum_{j=1}^{n} w_{ij}^2 \sum_{l=t+1}^{m} |x_{jl}^s - p_{il}^s|^2 C_i^b = \varkappa_2 \cdot \sum_{j=1}^{n} w_{ij}^2 \sum_{l=m+1}^{y} |n_{il}^b - n_{jl}^b|^2 \; C_i^o = \varkappa_3 \cdot \sum_{j=1}^{n} w_{ij}^2 \sum_{l=y+1}^{u} |n_{il}^o - n_{il}^o|^2$$

Eq. 6.7

We rewrite Eq.6.6 as:

$$C(W) = \sum_{i=1}^{k}(C_i^r + C_i^s + C_i^b + C_i^o)$$ **Eq. 6.8**

7.0 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.

8.0 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 as 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.

9.0 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. 5.1 and Eq. 6.6, 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.

$$f(g) = \frac{1}{1+e^{c(W)}}$$ **Eq. 9.1**

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:

$$P_s(g_{(i)}) = \frac{f(g_i)}{\sum_{i=1}^{N} f(g_i)}$$ **Eq. 9.2**

The operation probabilities for crossover and mutation are adaptively assigned as Eq. **Eq. 9.3**

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

$$P_m(g_i) = \begin{cases} \frac{\alpha_3(f_{max} - f(g_i))}{f_{max} - \bar{f}} f(g_i) \geq \bar{f} \\ \alpha_4 otherwise \end{cases}$$ **Eq. 9.4**

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

$\bar{f} = \sum_{i=1}^{N} f(g_i)$, $f' = \{f(g_j)\}$,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:

$$w_{ij} = \sum_{i=1}^{k} \frac{(d(x_j, x_i))2}{(d(x_j, x_i))2}, i, j$$ **Eq. 9.5**

10. A PRACTICAL SAMPLE DATA TABLE OF MIXED VARIETY OF DATA TYPES:

We are representing the real life concept of our approach by taking 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 10.1: The Weighted Matrix W_{ij}

Test-1 contains 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 belongs (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 10.2: The Real Life Practical Dataset

Table 10.2 is converted into the normalized matrix using the above equations Eq. (6.1), Eq. (6.2), Eq. (6.3), and Eq. (6.4)

	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 10.3: The Normalized Matrix of Table 10.2

We calculate the value of the expression (stated below) to be further used in Eq.6.6

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

	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 10.4: The Matrix calculated by Eq. 6.6

Now for Eq.6.6 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 10.5: The Matrix for expression of Eq. 6.6

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 10.6: The Calculated Expression Value

Now using the Eq. 9.1, we find the fitness value of the above calculated values (above 5 tulles):

1	0.2066
2	0.3689
3	0.4498
4	0.2497
5	0.1259

Table 10.7: The Fitness Values of Calculated Values

First we arrange the above values in ascending order and label each one of them and then using Eq.9.2 calculate the selection probability $P_s(g_{(i)})$

1	0.1474
2	0.2633
3	0.3204
4	0.1782
5	0.0898

Table 10.8: The Selection Probability

11.0 ANALYSIS ON OUR EXPERIMENTAL RESULTS

From 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

12.0 CONCLUSION AND FUTURE SCOPE

Here we have presented the genetic algorithm to cluster large data sets. 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.

ACKNOWLEDGEMENT

The authors would like to thank the reviewers for their valuable suggestions. They would also like to thank HOD-CSE, Dr. R. RadhaKrishnan and HOD-IT, 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.

REFERENCES

[1]. LIJie, GAO Xinbo, JIAO Li-cheng, "A GA-Based Clustering Algorithm for Large Data Sets with Mixed Numeric and Categorical Values", National Key Lab. of Radar Signal Processing, Xidian Univ., Xi'an 710071, China, 1998.

[2]. Zhou Caiying of Science & Technology Division and Jiang Xi , "The Improvement of Initial Point Selection Method for Fuzzy K-Prototype Clustering Algorithm", 2010 2nd International Conference on Education Technology and Computer (ICETC), University of Science and Technology ,GanZhou, China Zhoucaiying, Longjun Software of Software, Jiang Xi Normal University ,Nanchang, China,2010.

[3]. Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Dept. of Computer Science, the University of British 3. M. R. Anderberg, "Cluster Analysis for Applications", Academic Press, New York,1973.

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

[5]. 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.

[6]. Zhexue Huang, "A fast clustering algorithm to cluster very large categorical data sets in data mining Columbia, Canada, pp.1-8.

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

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

Author's Profile

Mr. Rohit Rastogi received his B.E. degree in Computer Science and Engineering from C.C.S.Univ. Meerut in 2003, the M.E. degree in Computer Science from NITTTR-Chandigarh (National Institute of Technical Teachers Training and Research-affiliated to MHRD, Govt. of India), Punjab Univ. Chandigarh in 2010. He was Asst. Professor at IMS College, Ghaziabad in computer Sc. Dept. He has authored/co-authored, participated and presented research papers in various Science and Management areas in around 40 International Journals and International conferences including prestigious IEEE and Springer and 10 national conferences including SRM Univ., Amity Univ., JP Univ. and Bharti Vidyapeetha etc. He has guided 5 ME students in their thesis work and students of UG and PG in around 100 research papers. He has developed many commercial applications and projects and supervised around 30 B.E. students at graduation level projects.

At present, he is a Sr. Asst. Professor of CSE Dept. in ABES Engineering. College, Ghaziabad (U.P.-India), affiliated to Uttar Pradesh Tech. University, Lucknow.

His research interests include Data Ware Housing and Data Mining, Design Analysis of Algorithm, Theory of Computation & Formal Languages and Data Bases. At present, He is engaged in Clustering of Mixed Variety of Data and Attributes with real life application applied by Genetic Algorithm, Pattern Recognition and Artificial Intelligence. Also, he is preparing some interesting algorithms on Swarm Intelligence approaches like PSO, ACO, BCO etc.