International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:3, 2009

Improved K-Modes for Categorical Clustering Using Weighted Dissimilarity Measure

S.Aranganayagi and K.Thangavel

Abstract—K-Modes is an extension of K-Means clustering algorithm, developed to cluster the categorical data, where the mean is replaced by the mode. The similarity measure proposed by Huang is the simple matching or mismatching measure. Weight of attribute values contribute much in clustering; thus in this paper we propose a new weighted dissimilarity measure for K-Modes, based on the ratio of frequency of attribute values in the cluster and in the data set. The new weighted measure is experimented with the data sets obtained from the UCI data repository. The results are compared with K-Modes and K-representative, which show that the new measure generates clusters with high purity.

Keywords—Clustering, categorical data, K-Modes, weighted dissimilarity measure

I. Introduction

Data mining is a rapidly growing interdisciplinary field which merges together database, statistics, machine learning and related areas in order to extract hidden information from the data. Clustering categorical data is the recent research problem in data mining. Clustering is the process of organizing objects into classes/groups such that the objects within the same cluster have a high degree of similarity, while objects belonging to other clusters has a high degree of dissimilarity. To group the objects, similarity or distance measure is used. The data used in the clustering is of various types like numerical, nominal etc. For numerical clustering, the distance measure based on geometric concepts such as Euclidean distance or Manhattan distance is used [1, 7]. Since the categorical data contains nominal values like $[male, female]$, $[low, medium, high]$,the geometric distance measures are not applicable for categorical or nominal data. It is not possible to use the geometric measure for these types of data [3, 5, 15]. Huang proposed a simple matching or mismatching measure in K-Modes to find the similarity between objects. For e.g. If the object t1 = [x, y, a] and t2 = [x, s, a], the distance between t1 and t2 is computed as 1 according to the hamming distance. This type of measure does not consider the implicit similarity relationship embedded in categorical values, which result in weaker intra cluster similarity [14].

Modes are the attribute values with high frequency. In K-Modes, 'K' distinct records or most frequent attribute values are selected as modes. Each object is compared with the modes using simple matching measure and the object is placed in the nearest cluster and the modes are also updated [14]. From the experimental results we found the occurrence of $modetie$ after some iteration. In order to avoid drawbacks we proposed a weighted measure to obtain the fine difference and place the object in the appropriate cluster.

In the weighted measure, the frequency of attribute values is considered. If the values of an attribute are equal, then the frequency of the attribute value in the dataset or cluster are taken into account for clustering. As in K-histogram, if the ratio of frequency of attribute values in the data set is considered, it can be found in one scan of the data set and it will be constant throughout the process. Thus this measure has little difference than that of the K-Modes. When the ratio of frequency of attribute value within the cluster is considered for clustering, no weight is given for the most frequent values in the data set. In this paper, we propose a new weighted measure which is the product of the ratio of frequency of attribute values in the cluster and the ratio of frequency of attribute values in the data set.

Few existing algorithms such as variation of K-Modes using frequency based dissimilarity measures, Sen et al. proposed relative frequency based measure for K-Modes in [6]. In relative frequency based methods, if the attribute values are equal, then the proportion of attribute value in the cluster is considered.

The proposed method is experimented with eight data sets viz. voting, soybean small, lymbhography, zoo, balance scale, car, hayesroth, and mushroom obtained from the UCI machine learning data repository and compared.

Section 2 describes related work. Section 3 summarizes the notations and definitions used in this paper. Section 4 deals with the function of K-Modes algorithm. Section 5 discusses the proposed method. Experimentation details and the results are discussed in Section 6. Section 7 concludes the paper.

II. Related Work

Few algorithms have been proposed in recent years for clustering categorical data,like Expectation Maximization(EM), STIRR[2], ROCK[1,9,10,12], CACTUS[10], COOLCAT[2], LIMBO[8], K-representative[6], K-histograms[13] and Squeezer[12] etc.

The EM algorithm is a general method of finding the maximum-likelihood estimate of the parameters of an underlying distribution from a given data set when the data is incomplete or has missing values. EM assigns probability to each class or category. Sieving Through Iterated Relational

Aranganayagi.S is with the J.K.K.Nataraja College of Arts & Science, Komarapalayam, Tamilnadu, India and with the Department of Computer Science and Applications, Gandhigram Rural University, Gandhigram, Tamilnadu, India. (Corresponding author: Phone: 0424 -2230855; E-mail:arangbas@yahoo.com,arangbas@gmail.com)

Dr.K.Thangavel is with the Department of Computer Science, Periyar University, Salem, India. (Phone: 0427-2330911,0427-2345766, drktvelu@yahoo.com)

Reinforcement (STIRR), is an iterative algorithm based on nonlinear dynamical systems. It represents each attribute value as a weighted vertex in a graph. Starting with the set of weights on all vertices, the system is iterated until a fixed point is reached[2]. Robust hierarchical Clustering with linKs (ROCK) is an adaptation of an agglomerative hierarchical clustering algorithm, which heuristically optimizes a criterion function, defined in terms of the number of links between objects. Informally the number of links between two objects is the number of common neighbors that they have in the dataset [1, 9, 10, 12]. Clustering Categorical Data Using Summaries (CACTUS) attempts to split the database vertically and tries to cluster the set of projections of these objects to only a pair of attributes [10]. The COOLCAT algorithm uses the entropy measure in clustering [2]. Set of K initial objects are selected such that the minimum pair wise distance among them is maximized. Remaining objects are placed in the cluster with minimum entropy. The ScaLable Information Bottleneck(LIMBO) algorithm clusters the categorical data using information bottle neck as a measure. LIMBO uses distributional summaries to handle large data sets. Instead of keeping objects or whole clusters in main memory, only the statistics are maintained. In the first phase the DCF tree is constructed. In the second phase the leaves are combined to form clusters. In the third phase the objects are assigned to clusters [8].

K-Modes is an extension of K-Means clustering algorithm, but the working principle of both is same. Instead of means, the concept of modes are used. By varying the dissimilarity measure, fuzzy K-Modes, K-representative and K-histogram are developed. In fuzzy K-Modes, instead of hard centroid, soft centroid is used[3]. In K-representative algorithm, the measure relative frequency is used. Frequency of attribute value in the cluster divided by cluster length is used as a measure in K-representative [6]. In K-histogram, proportion of attribute value in the data set is considered. K-Modes is extended with fuzzy, genetic and fuzzy-genetic concepts. As the proposed measure is similar to K-Modes and K-representative, we compared the proposed measure with K-Modes and K-representative.

III. Definitions and Notation

Let $D = x_1, x_2, ..., x_n$ be the data set with 'm' categorical attributes. Let the attributes be $A_1, A_2, ..., A_m$ with domains $D_1, D_2, .., D_m$ respectively. The domain of $A_i, D_i = V_{i_1}, V_{i_2}, ..., V_{i_s}$, that is the i^{th} attribute contains 's' distinct values. Number of objects in the data set D with attribute value V_{i_j} is denoted as $f(V_{i_j})/D$. Number of objects in the cluster C_i with attribute value V_{i_j} is denoted as $f(V_{i_j}/C_i)$. Number of objects in the cluster C_i is represented as $f(C_i)$.

Huang proposed a similarity measure to find the distance between two categorical objects[14]. Let X, Y be two categorical objects described by 'm' categorical attributes. The dissimilarity between the two objects is the total mismatches of the corresponding attribute values. Objects are grouped based on the minimum dissimilarity.

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \tag{1}$$

$$\delta(x_j, y_j) = \begin{cases} 0 & \text{if} x_j = y_j \\ 1 & \text{if} x_j \neq y_j \end{cases} \tag{2}$$

In this paper, instead of using simple mismatching measure, we propose a new dissimilarity measure which is the product of proportion of attribute value in the cluster and the proportion of attribute value in the data set, which is defined as,

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \tag{3}$$

$$\delta(x_j, y_j) = \begin{cases} 1 - w_{l_j} & \text{if} x_j = y_j \\ 1 & \text{if} x_j \neq y_j \end{cases} \tag{4}$$

where w_{l_j} is defined as,

$$w_{l_j} = (f(v_{l_j}/C_l)/|C_l|) * (f(v_{l_j}/D)/|D|) \tag{5}$$

IV. K-Modes Algorithm

The data set and 'K', the number of cluster are inputs to the K-Modes algorithm. The 'K' initial modes are selected as either 'K' distinct objects or most frequent occurring attribute values.

Algorithm: K-Modes

1) Select 'K' initial modes.
2) Allocate an object to the cluster whose mode is the nearest to the cluster, using definition (1). Update the mode of the cluster after each allocation.
3) After all objects have been allocated to the respective cluster, retest the objects with new modes and update the clusters.
4) Repeat steps(2) and (3) until there is no change in clusters.[14]

The K-Modes is an extension of K-Means, which partitions the dataset into 'K' clusters that minimize the objective function P, with unknown variables U and Z is defined as follows:

$$P(U,Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} u_{i,l} d(x_{i,j}, z_{l,j}) \tag{6}$$

subject to

$$\sum_{l=1}^{k} u_{i,l} = 1, \quad 1 \leq i \leq n \tag{7}$$

$$u_{i,l} \in 0, 1, \quad 1 \leq i \leq n \quad , \quad 1 \leq l \leq k \tag{8}$$

where U is an n x k partition matrix, $u_{i,l}$ is a binary variable and $u_{i,l} = 1$ indicates that object X_i is allocated to cluster C_i and Z is a cluster center. The optimization problem can be solved by iteratively solving the objective function to get the minimum value and each time, the center is also updated.

TABLE I
DESCRIPTION OF THE DATA SET

Data set	Number of objects	Number of attributes	Classes
Soybean Small	47	21	4
Mushroom	8124	22	2
Congressional Votes	435	16	2
Zoo	101	16	7
Lymbhography	148	19	4
Hayesroth	132	4	3
Balance scale	625	4	3
Car evaluation	1728	6	4
Audiology	200	62	24

To maximize the intracluster similarity, the value $(f(v_{l_j}/C_l)/|C_l|) * (f(v_{l_j}/D)/|D|)$, should be maximum. When the frequency of attribute value within the cluster or within the data set is considered, the object may be placed in cluster with less similarity. As the product of proportion is considered, the object will be placed in the cluster with high similarity. Thus the proposed measure improves the accuracy of the algorithm.

V. K-MODES WITH PROPOSED WEIGHTED MEASURE

As in K-Modes, modes are initialized either by frequency of attribute values or by distinct records. 'K', the number of cluster is given as input. During clustering process the object is placed in the nearest cluster according to the definition(3). The mode is updated during the object allocation. The process is repeated until there is no change in the clusters and the objective function is minimized.

The proposed method with new similarity measure is given below.

Algorithm : K-Modes with proposed measure

1) Select 'K' initial modes.
2) Allocate an object to the cluster whose mode is nearest to it according to the proportion of the attribute value in the cluster and the dataset. Update the mode of the cluster, after each allocation.
3) After all objects have been allocated to the cluster, retest the objects with new modes and update the clusters.
4) Repeat steps (2) and (3) until there is no change in the clusters.

The proposed method is tested with data sets obtained from the UCI machine learning data repository. As the proposed method is similar to K-representative, we have compared the performance of the proposed algorithm with K-Modes and K-representative.

VI. EXPERIMENTS AND RESULTS

A. Data Set

Table-1 lists the details of the data set obtained from the UCI data repository, used in the experiments for analysis [11]. In all the data sets, attribute which relates to class label is omitted for the purpose of clustering.

TABLE II
PURITY MEASURE FOR K-MODES, K-REPRESENTATIVE AND PROPOSED METHOD

Data Set	Number of Clusters	K-Modes	K-represen-tative	Proposed Algorithm
Soybean small	4	66	89	89
Mushroom	2	59	61	61
Congressional Votes	2	62	87	88
Zoo	7	88	89	90
Lymbhography	4	64	58	67
Hayesroth	3	41	42	42
Balance scale	3	50	52	52
Car evaluation	4	70	70	71
Audiology	24	62	61	62

B. Purity Measure

A cluster is called a pure cluster if all the objects belong to the same class. To measure the efficiency of the proposed method, we have used the clustering accuracy measure, suggested by Huang (1998). The clustering accuracy r is defined as,

$$r = 1/n \sum_{i=1}^{k} a_i \quad (9)$$

where a_i refers to the data objects that occur in both cluster C_i and its corresponding labeled class, and 'n' is the number of objects in the data set. The clustering error 'e' is defined as $e = 1 - r$. If a partition has a clustering accuracy of 100 percent, it means that it has only pure clusters. Large clustering accuracy implies better clustering [12].

'K' distinct records are selected at random and considered as initial modes for all the three algorithms. Purity measure is computed and is tabulated in Table-II. Relative performance of the algorithms has been analyzed by ranking the algorithms and shown in Fig-1.

It is observed that the proposed method produces the clusters with high purity when compared with K-Modes and K-representative.

For the data sets considered for experimentation, the number of classes or categories is known. Purity measure [12] is related to the class or categories. If the value of 'K' chosen is less than the number of actual categories, then purity measure will be less. Consider the soybean small data set with four classes and the classes contain records such as 10, 10, 10 and 17 respectively. Let the four classes be D1, D2, D3 and D4. For K= 2 the confusion matrix obtained is shown in Table-III. Even though all objects belonging to a particular class is placed in the same cluster, the purity measure is (10 + 17) / 47 i.e. 0.57. Thus we get only 50 percent of purity and the error rate is also 50 percent. So we choose the value of K as the

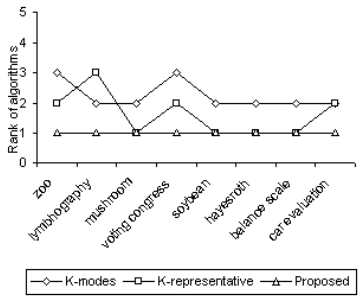

Fig. 1. Relative Performance of the Algorithms

TABLE III
CONFUSION MATRIX - EXAMPLE

Class/ Cluster	D1	D2	D3	D4
1	10	10		
2			10	17

TABLE IV
PURITY MEASURE FOR ZOO DATA SET

Number of Clusters	K-Modes	K-representative	Proposed Algorithm
8	0.835	0.835	0.835
9	0.827	0.845	0.845
10	0.955	0.955	0.955

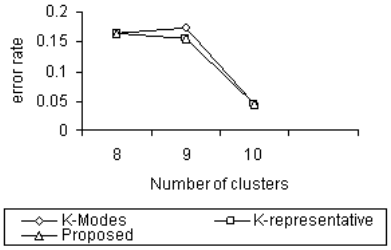

Fig. 2. Error rate of zoo data set

number of classes and vary it up to 10. The purity measure for different values of 'K', for the three methods K-Mode, K-Representative and the proposed method are tabulated. (Table-IV to Table- XI).

From the Table-IV, it is clear that the performance of the proposed method is the same as K-representative for all three cases of K. The proposed method produces the best clusters rather than the K-Modes, when K = 9.

For lymbhography data set, the proposed method is efficient for all the values of K except K = 9, the proposed method produces the same result as K-representative in two cases when K= 7 and K = 10. For the remaining cases the proposed method produces the clusters with high purity as shown in Table-V.

TABLE V
PURITY MEASURE FOR LYMBHOGRAPHY DATA SET

Number of Clusters	K-Modes	K-representative	Proposed Algorithm
5	0.68	0.77	0.71
6	0.66	0.71	0.77
7	0.74	0.71	0.71
8	0.67	0.72	0.79
9	0.77	0.75	0.75
10	0.68	0.77	0.77

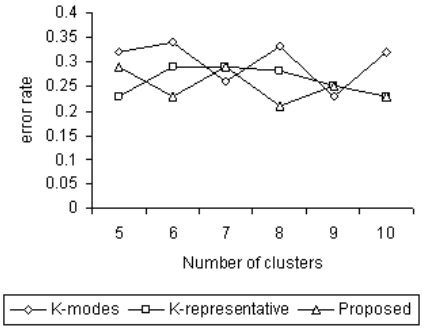

Fig. 3. Error rate of Lymbhography data set

TABLE VI
PURITY MEASURE FOR MUSHROOM DATA SET

Number of Clusters	K-Modes	K-representative	Proposed Algorithm
3	0.52	0.62	0.62
4	0.75	0.77	0.74
5	0.88	0.86	0.91
6	0.81	0.91	0.89
7	0.88	0.75	0.88
8	0.88	0.93	0.91
9	0.89	0.93	0.93
10	0.89	0.93	0.90

For Mushroom data set, the proposed method excels in all the cases when compared to K-Modes, and with K-representative the proposed method generates clusters with equal purity for K=3 and K=9, and generates better clusters when K = 5 and K = 7. Results are tabulated in Table - VI.

The proposed method is efficient in all the cases for the congress data set. All the three algorithms generate clusters with the same purity when K = 9, and the proposed and K-Modes are equal when K= 3, 5, 6, and 9 and the results are shown in Table - VII.

All three methods produce clusters with the same purity when K= 6, 7, 9 and 10 for the soybean small dataset. When K= 5, K-representative and the proposed method are with the same purity. And when K = 8, the proposed method produces clusters with high purity. Results are tabulated in Table - VIII.

For Hayesroth data set, the proposed method excels in all cases. For balance scale data set, the proposed and K-

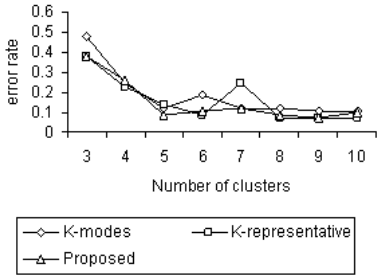

Fig. 4. Error rate of Mushroom data set

TABLE VII
PURITY MEASURE FOR VOTING CONGRESS DATA SET

Number of Clusters	K-Modes	K-representative	Proposed Algorithm
3	0.93	0.92	0.93
4	0.91	0.92	0.93
5	0.93	0.90	0.93
6	0.94	0.93	0.94
7	0.92	0.92	0.93
8	0.90	0.93	0.94
9	0.93	0.93	0.93
10	0.90	0.91	0.92

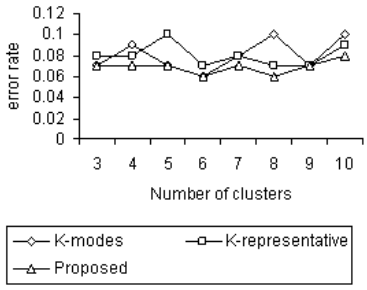

Fig. 5. Error rate of Voting Congress data set

TABLE VIII
PURITY MEASURE FOR SOYBEAN DATA SET

Number of Clusters	K-Modes	K-representative	Proposed Algorithm
5	0.94	1.00	1.00
6	1.00	1.00	1.00
7	1.00	1.00	1.00
8	0.91	0.91	1.00
9	0.98	0.98	0.98
10	1.00	1.00	1.00

representative generates clusters with the same purity, except when K = 10. For Car Evaluation data set, proposed method excels in all cases. All the three are equal when K = 5 and K = 6. Results are shown in tables IX ,X and XI.

From Figure-2 to Figure-9, it is observed that the proposed

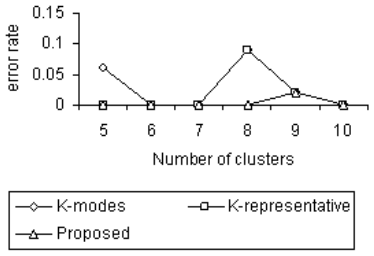

Fig. 6. Error rate of soybean data set

TABLE IX
PURITY MEASURE FOR HAYESROTH DATA SET

Number of Clusters	K-Modes	K-representative	Proposed Algorithm
4	0.42	0.42	0.43
5	0.44	0.44	0.46
6	0.45	0.47	0.47
7	0.45	0.45	0.46
8	0.48	0.51	0.51
9	0.47	0.52	0.52
10	0.46	0.46	0.46

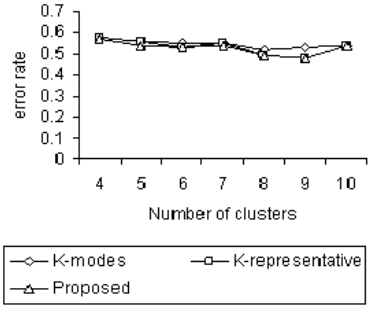

Fig. 7. Error rate of Hayesroth data set

TABLE X
PURITY MEASURE FOR BALANCE SCALE DATA SET

Number of Clusters	K-Modes	K-representative	Proposed Algorithm
4	0.50	0.56	0.56
5	0.56	0.58	0.58
6	0.55	0.57	0.57
7	0.63	0.67	0.67
8	0.55	0.60	0.60
9	0.62	0.62	0.62
10	0.62	0.66	0.67

method generates clusters with a lower error rate. The experiments are carried out for five different modes and the average purity measure is calculated.

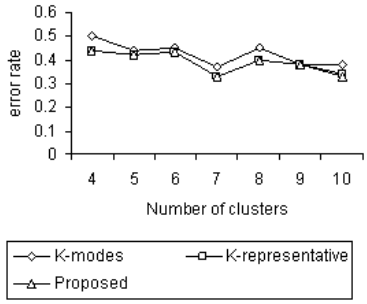

Fig. 8. Error rate of Balance Scale data set

TABLE XI
PURITY MEASURE FOR CAR EVALUATION DATA SET

Number of Clusters	K-Modes	K-representative	Proposed Algorithm
5	0.70	0.70	0.70
6	0.70	0.70	0.70
7	0.70	0.74	0.74
8	0.70	0.71	0.71
9	0.70	0.71	0.72
10	0.72	0.70	0.76

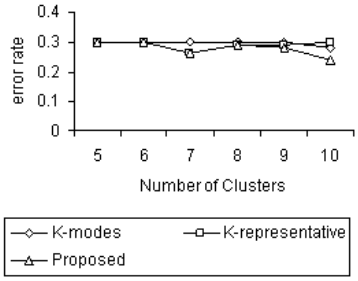

Fig. 9. Error rate of Car Evaluation data set

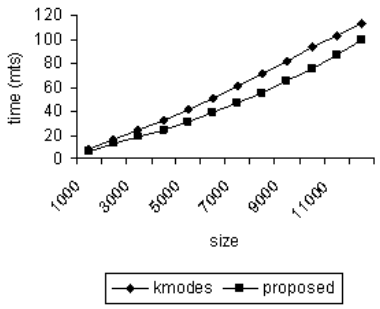

Fig. 10. nursery data set- execution time for K-Modes and the proposed

C. Scalability

To test the scalability of the proposed method it is experimented with the nursery data set and the audiology data set. Audiology data set contains more attributes and nursery data set contains more objects. The nursery data set contains 12960 objects with 8 attributes and 5 classes. The actual time taken is shown in the fig-10.

D. Complexity

One preprocessing step is required to compute the frequency of attribute value in the data set, which needs a single pass and is finished in O (n) time. The time complexity of all K-Modes and its variations is O(tmkn), where t is the total number of iterations required, k- the number of clusters, n -total number of objects, m - number of attributes in the data set.

VII. Conclusion

This paper thus proposes a new weighted dissimilarity measure which relates to the attribute values in the cluster and in the data set. Experimental Analysis shows that the proposed algorithm produces quality clusters, when compared to K-Modes and K-representative. Results of K-Modes and its variations depend on the initial selection of modes. Better initialization leads to better and efficient clusters. Thus our future intention is to find out efficient methods to initialize modes.

References

[1] Arun.K.Pujari, ”Data Mining Techniques”, Universities Press, 2001.
[2] Daniel Barbara, Julia Couto, Yi Li, ”COOLCAT An entropy based algorithm for categorical clustering”, Proceedings of the eleventh international conference on Information and knowledge management, 2002, 582 - 589.
[3] Dae-won kim, Kwang H.Lee, Doheon Lee, ”Fuzzy clustering of categorical data using centroids”, Pattern recognition letters 25, Elseveir, (2004), 1263-1271.
[4] George Karypis, Eui-Hong (Sam) Han, Vipinkumar, ”CHAMELEON: A hierarchical clustering algorithm using dynamic modeling”, IEEE Computer, 1999.
[5] Jiawei Han, Micheline Kamber, ”Data Mining Concepts and Techniques”, Harcourt India Private Limited, 2001.
[6] [6] Ohn Mar San, Van-Nam Huynh, Yoshiteru Nakamori, ”An Alternative Extension of The K-Means algorithm For Clustering Categorical Data”, J. Appl. Math. Comput. Sci, Vol. 14, No. 2, 2004, 241-247.
[7] Pavel Berkhin, ”Survey of Clustering Data Mining Techniques”, Technical report, Accrue software,2002
[8] Periklis Andristos, Clustering Categorical Data based On Information Loss Minimization, EDBT 2004: 123-146.
[9] Sudipto Guga, Rajeev Rastogi, Kyuseok Shim, ”ROCK, A Robust Clustering Algorithm For Categorical Attributes”, ICDE ’99: Proceedings of the 15th International Conference on Data Engineering, 512, IEEE Computer Society, Washington, DC, USA,1999
[10] Venkatesh Ganti, Johannes Gehrke, Raghu Ramakrishnan, ”CACTUS -Clustering Categorical Data using summaries”, In Proc. of ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining, 1999, San Diego, CA USA.
[11] www.ics.uci.edu/ mlearn/MLRepository.html
[12] Zengyou He, Xiaofei Xu, Shengchun Deng, ”Squeezer: An Efficient algorithm for clustering categorical data”, Journal of Computer Science and Technology, Volume 17 Issue 5, Editorial Universitaria de Buenos Aires, 2002.
[13] Zengyou He, Xiaofei Xu, Shengchun Deng, Bin Dong,” K-Histograms: An Efficient Algorithm for Catgorical Data set”, www.citebase.org.
[14] Zhexue Huang , ”A Fast Clustering Algorithm to cluster Very Large Categorical Datasets in Data Mining”, In Proc. SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997.
[15] Zhexue Huang, ”Extensions to the K-means algorithm for clustering Large Data sets with categorical value”, Data Mining and Knowledge Discovery 2, Kluwer Academic publishers, 1998. 283-304.

Aranganayagi.S She received the degree Master of Computer Applications from Pondicherry Engineering College, Pondicherry, India in 1989. Currently she is working as a Selection Grade Lecturer at J.K.K.Nataraja College of Arts & Science, Komarapalayam, Tamilnadu, India and her experience in teaching started from the year 1990. She is doing research in the Department of Computer Science and Applications, Gandhigram Rural University, Gandhigram, India. Her areas of interests include Data Mining, Clustering, soft computing.

Thangavel.K He received the degree of Master of Science from Department of Mathematics, Bharathidasan University, Tiruchi, in 1986, and Master of Computer Applications from Madurai Kamaraj University, India in 2001. He obtained his Ph.D from Department of Mathematics, Gandhigram Rural University, in 1999. Currently he is working as a Professor, Computer Science, Periyar University, Salem and his experience in teaching started from 1988. His areas of interest include Medical Image processing, Artificial Intelligence, Data Mining, bioinformatics and soft computing.