Clustering Categorical Data using Bayesian Concept

Aranganayagi.S and Thangavel. K

***Abstract*— Clustering is the process of grouping similar objects. Naïve Bayes Classifier is the classification technique which is widely used to predict the unknown class labels. Here in this paper we extend this concept to unsupervised classification, clustering. As in K-modes the proposed method starts the clustering process with the modes. Based on the prior information bayes theorem is used to place the object in the respective clusters. The feature of the proposed algorithm is scalability and it need only one data scan. The proposed Bayesian clustering to cluster categorical data is experimented with the real data sets obtained from the UCI machine learning data repository and compared with the well known K-modes algorithm to cluster the categorical data. Experimental results prove that the proposed method is efficient than K-modes.**

***Keywords*— clustering, categorical data, Bayesian theorem, mode.**

I. INTRODUCTION

Clustering is understood as a decomposition or partition of data set into groups in such a way that the objects in one group are similar to each other but as different as possible from the objects in other groups [1, 6]. Thus, the main goal of clustering is to detect whether or not the general population is heterogeneous, that is, whether the data fall in to distinct groups. Clustering in a historical perspective rooted in mathematics, statistics and numerical analysis. From machine learning perspective, cluster corresponds to hidden patterns [8].

Grouping the object is carried out using some matching criteria. These criteria may be a simple Euclidean measure for numeric data. Geometric properties of objects are used in numeric clustering. These geometric properties can not be applied to categorical or nominal data. Categorical data is usually with small domains and which can not be ordered [4, 6, and 17]. Huang proposed the simple mismatching measure as, if the attribute values of two objects are unequal then the distance is assumed to be one else it is zero[16]. The concept of similarity alone is not sufficient for categorical data. Due to the special properties of categorical data it seems more complicated than that of numerical data [15]. Compared to continuous values, nominal values are with small domains.

Manuscript received January 21, 2009. Aranganayagi.S is with the J.K.K.Nataraja College of Arts & Science, Komarapalayam, Tamilnadu, India and doing research in the Department of Computer Science and Applications, Gandhigram Rural University, Gandhigram Tamilnadu India. Member of IAENG: Corresponding author, phone: 0424-2230855, 9842723085.

Dr.K.Thangavel is with the Periyar University, Salem, Tamilnadu, India as Professor in Computer Science.

Clustering is widely divided into partitional or hierarchical. K-means and K-modes come under the category of partitional clustering. K-means is an efficient and widely used technique to cluster the numeric objects. Huang proposed an algorithm called K-Modes which is an extension of K-means. Here instead of means, modes are used as centroids. The result depends on the initial selection of modes and 'K' the number of clusters. Most of the clustering algorithm requires the number of clusters or the threshold level as input. In the unsupervised learning without having knowledge about classes it is not possible to decide about 'K'. By repetitive execution minimum intra cluster similarity is achieved [16]. To find the effective 'K', the clustering technique is applied for different values of 'K' and the best partition is taken based on validity measure.

Modes are the most frequent attribute values in the partition. For Example consider a partition Ci with an attribute aj. If Ci contains say 30 objects, and the attribute aj contains values like [('x', 15), ('r', 14), ('t', 1)], then 'x' is selected as a mode of the attribute. But there is not much difference in the frequency of attribute values 'x' and 'r'. Even if the next tuple considered is with the value 'r', it is not taken in to account for similarity measure used in K-modes. Thus instead of just comparing the attribute values, all the values has to be considered to achieve a better clustering. So information in the partition has to be considered in the similarity measure to place the object in the appropriate cluster. In this paper we propose a method based on Bayes Theorem to cluster the objects.

The proposed method is based on the Bayesian concept, places the object in the cluster for which the posteriori probability is maximized [6]. It is an extension of the Bayesian classifier. Like K-modes, either most frequent attribute values or the distinct records are selected as modes. If the number of cluster is 'K', K modes are selected. All K clusters are initialized with one object. Treating this as prior information for Bayes theorem, posteriori probability is computed. Tuples/objects are read one by one and placed in the cluster with maximum posteriori probability. The purity of the clusters depends on the initial selection of modes and the number of clusters. Objects placed in the cluster contribute much in further clustering thus no updating of modes and no repetitive execution is needed. Experimental result shows that the proposed method performs well.

Section 2 describes some related methods to cluster the categorical data. Section 3 briefs the Bayesian concept. Section 4 discusses the proposed method. Experimentation details and the results are discussed in Section 5. Section 6 concludes the paper.

II. Related Work

A few existing categorical clustering algorithms are discussed in this section. The K-modes algorithm, an extension of K-means for categorical data use modes instead of means. Sieving through Iterated Relational 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. [3]

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 tuples. Informally the number of links between two tuples is the number of common neighbors that they have in the dataset [1, 10, 11, 14].

Clustering Categorical Data Using Summaries (CACTUS) attempts to split the database vertically and tries to cluster the set of projections of these tuples to only a pair of attributes [11]. The COOLCAT algorithm uses the entropy measure to group the records. The clustering process is carried out in two steps: initialization and incremental step. Algorithm groups objects in such a way that the expected entropy is minimized. In the first step 'k' most dissimilar records are selected and form the sample set by maximizing the minimum pairwise entropy of the chosen points. In the incremental step, the remaining records of the data set are placed in the appropriate clusters by computing the expected entropy [3]. The LIMBO algorithm clusters the categorical data using information bottle neck as a measure. This algorithm uses distributional summaries to deal with larger data set [9].

Instead of using simple matching measure, weighted frequencies concept is used in the variation of K-modes such as K-representative, fuzzy K-modes and K-histogram. The K-representative algorithm is an extension of K-means algorithm using relative frequency to cluster the data [7]. In fuzzy K-modes, instead of hard centroids, soft centroids are used [4]. The Squeezer algorithm reads each tuple't' in sequence, either assign 't' to an existing cluster or create 't' as a new cluster which is determined by the similarity between 't' and clusters. Similarity threshold level is used to group the objects [14]. The K-histogram algorithm extends the K-means algorithm by replacing the means of the clusters with histograms, and dynamically updates the histogram in the clustering process [15]. In all the K-modes related algorithms initial mode are selected and the modes are updated during each run.

In probabilistic clustering approach, data is considered to be sample independently chosen from a mixture model of several probability distributions. Expectation Maximization (EM) is an iterative refinement algorithm that can be used to find the parameter estimates. EM assigns an object to the cluster according to a weight representing the probability of membership. And assumes that the entire data fits in to main memory thus it is not scalable. AutoClass is a Bayesian clustering method that uses a variant of EM algorithm. Auto class can also estimate the number of clusters. AutoClass is an iterative clustering algorithm, if the data could not be loaded into memory, then the time cost is expensive [2]. COBWEB, the conceptual clustering algorithm, yields a clustering dendrogram called classification tree that characterizes each cluster with probabilistic distribution. COBWEB and its derivatives use the category utility measure to partition the data set [6].

III. Bayesian Classification

Bayesian classifier is a probabilistic model used to estimate the class for a new data item. Bayesian Classification is based on the Bayes theorem. Naïve Bayes classifier assumes the attributes are independent. Bayesian classifier is used to classify both numeric and nominal data.

A. Bayes Theorem

Bayesian theory gives a mathematical calculus of degree of belief. Let X is a data object/tuple and H is the hypothesis. X is also considered as evidence. If we have 'K' classes then the hypothesis H may be defined as the evidence X belongs to the class Ci. P(H/X) is the probability that the tuple X belongs to the class C or the hypothesis H holds the evidence X.

P(H/X) is the posteriori probability of H conditioned on X. P(H) is the priori probability of the hypothesis H. P(X/H) is the posteriori probability of X conditioned on H. P(X) is the priori probability of X.

Bayes theorem is defined as

$$P(H/X) = \frac{P(X/H)P(H)}{P(X)} \quad (1)$$

The posteriori probability P(H/X) can be computed using P(H), P(X/H) and P(X).

B. Naïve Bayes Classification

Bayesian classifier is a probabilistic model of what's happening in the data, which estimates the class for new data item. Here we assume that all attributes are independent thus the joint probabilities are obtained by multiplying the dimension wise probabilities. Challenge in the Bayesian classification is unknown distribution. Let T be the data set with *n* objects x_1, x_2, …, x_n, and each object contains *m* attributes. Assume that the number of classes be 'K'. i.e $C = \{C_1, C_2, C_3, ...C_k\}$. Naive bayes classifier predicts the tuple X to the class C_i if only if

$$P(C_i / X) > P(C_j / X) \quad \text{for } 1 \le j \le m, j \ne i.$$

Using Bayes Theorem (1), maximum posteriori hypothesis can be found using

$$P(C_i / X) = \frac{P(X/C_i)\,P(C_i)}{P(X)} \quad (2)$$

Estimate the conditional probability $P(C_i /X)$ using the priori probability or estimate. The conditional probability

$$P(X / C_i) = \prod_{k=1}^{d} P(x_k / C_i)$$

(3)

$$= P(x_1 / C_i) \text{ x } P(x_2 / C_i) \text{ x } P(x_3 / C_i) \text{ x} \ldots \text{x } P(x_d / C_i)$$

(4)

d is the cardinality of domain of attribute X.

$$P(x_k / C_i) = \frac{P(x_k \wedge C_i)}{P(C_i)}$$

(5)

where $P(x_k \wedge C_i)$ is the probability that the k^{th} dimension A_k has the value x_k and class C_i is given as,

$$P(x_k \wedge C_i) = \frac{\text{Number of cases with } C_i \text{ that have } A_k = x_k}{n}$$

(6)

Where A_k is the k^{th} attribute and n is the total number of objects in the entire database.

$$P(C_i) = n_i / n$$

(7)

Where 'n_i' is the number of objects in the class C_i. $n_i(x_k)$ is the number of objects in the class C_i with the value of A_k as x_k.

$$P(x_k / C_i) = \frac{n_i(x_k)}{n_i}$$

(8) and

$$P(X) = \sum_{i=1}^{k} P(X / C_i) P(C_i)$$

(9)

where 'k' is the number of classes or categories. If we consider the case when $n_i(x_k)$ is zero, then there is a chance of omitting such objects in subsequent process. But when we have large data set, there is a chance of having the attribute Ak with the value of x_k. Hence one is added to $n_i(x_k)$ as per Laplace adjustment.

In classification Naïve bayes classifier model is used to predict the unknown class labels. The data set is divided in to training set and test set. Training set samples are considered as prior information and the model is constructed [6].

IV. Proposed Method

In this paper we used the naïve bayes concept in clustering. With the assumption of K clusters, the objects are grouped based on the maximum posteriori probability. The process of clustering starts with K clusters each with one object as a member. Considering this as prior information, posteriori probability is computed for other objects, and the object is placed in the cluster with maximum posteriori probability. The objects are read one by one and placed in the respective clusters.

The proposed method is based on the concept of K-modes. Number of clusters and the initial set of modes are given as input. *K* distinct records are selected as initial values for *K* clusters.

Proposed Bayesian Clustering Method:

Input: Data set T, K-number of clusters.

i. Select K distinct records as initial objects for each cluster.
ii. Read the tuple X.
iii. Compute $P(C_i / X)$, $1 \le i \le K$.
iv. Place the object in the cluster which results in maximum posteriori probability.
v. Repeat (ii) to (iv) until all the objects in the dataset have been placed.

V. Results and Discussions

The proposed method is experimented with real data sets obtained from UCI data repository such as Mushroom, soybean, voting congress data sets etc.,[12]. The proposed method is compared with K-modes.

A. Data set

Mushroom: The mushroom data set contains 22 attributes with 8124 tuples. Each tuple describes the physical characteristic of a single mushroom. A classification label of poisonous or edible is provided with each tuple. The numbers of edible and poisonous mushrooms in the dataset are 4208 and 3916 respectively.

Congressional Votes: Each object represents one congressman's votes on 16 issues. All attributes are Boolean with yes and no values. A classification label of Republican or democrat is provided with each object. The data set contains 435 objects with 168 republicans and 267 democrats.

Soybean Small: The soybean data set has 47 instances, each being described by 35 attributes. Only 21 attributes are selected for experiment since others attributes are with single category. Each instance is labeled as one of the four diseases: Diaporthe Stem Canker, Charcoal Rot, Rhizoctonia root rot and Phytophthora rot.

Breast cancer: The breast cancer data set contains 699 instances with 11 attributes. A classification label of benign (2) or malign (4) is provided with each tuple. The first attribute id number is omitted.

Lymphography: The lymphography data set contains 148 instances with 19 attributes including class label. Each instance is labeled with one of the four classes: normal find, metastases, malign lymph and fibrosis.

Car : The car data set is with 1728 instances and each instance contains seven attributes. Attributes related to buying, maint, doors, persons, lug boot and safety. In this car evaluation data set each instance is classified as unacc, acc, good and vgood.

Hayes roth: Hayes roth and Hayes roth data set contains 132 instances with 6 attributes such as, name, hobby, age, education level, marital status and class label. The name field is coded with number from 1 to 132, and it is omitted for experimentation.

Balance Scale: The balance scale weight and distance data

set consists of 625 instances with 5 attributes. Each instance is labeled with 3 classes such as Balanced(B), Left(L) and Right(R).

Nursery: The nursery data set is with 12960 objects and 9 attributes. Each instance is classified into five categories such as not_recom, recommend, very_recom, priority, and spec_prior.

Attributes with unique value, with all distinct value and the class attributes are omitted for experimentation.

B.Accuracy measure

The clustering accuracy r is defined as,

$$r = \sum_{i=1}^{k} a_i / n \qquad (10)$$

where n is number of instances in the data set, ai is the number of instances occurring in both cluster 'i' and its corresponding class, which has the maximal value. Thus the clustering error is defined as e = 1 – r. If a partition has a clustering accuracy of 100%, it means that it has only pure clusters. Large clustering accuracy implies better clustering [14, 16].

C. Clustering Performance

In this section we compare the performance of the K-modes and the proposed Bayesian Method. The same set of records is considered as initial modes for both the algorithms. The number of cluster is given as the input and for the resulted clusters, purity measure is computed as per the definition given above and the results are tabulated.

Similar to K-modes results of the proposed method depends on the initial selection of modes. Based on the initial modes we get difference in the partitions. For example, confusion matrix of breast cancer and balance scale for two different modes is given in table–1, 2, 3 and 4. For breast cancer data set, purity rate of two different modes are approximately equal (Table-1 and 2) but in the balance scale data set when the initial modes are different, then there is a difference in the clusters found (table -3 and 4). To compute the efficiency of the cluster each data set is executed with five different modes and the average is taken (Table-5). Objects are placed in the cluster based on the attribute values in the partitions, so even if we repeat the execution with the new updated mode we get the same result. Thus the proposed method does not need the repetitive iteration as in K-modes.

TABLE – 1 CONFUSION MATRIX FOR BREAST CANCER DATA SET

(MODE-1)

Class/ category	2	4	Total	Max.	Purity
1	16	235	251	235	
2	442	6	448	442	
total	458	241	699	677	0.97

Table -1 is the result of breast cancer data set for K= 2. First cluster contains 251 objects in which 235 objects pertain to the class '4' and the second cluster contains 448 objects in which 442 objects belong the category of '2'. Out of 699 objects 677 objects have been placed correctly in the respective clusters. Thus the purity rate is 97%. Only 3% is misplaced. Whereas in table-2 679 objects were placed in the correct partition, and the purity rate is 97%.

TABLE – 2 CONFUSION MATRIX FOR BREAST CANCER DATA SET

(MODE 2)

Class/ category	2	4	Total	Max.	purity
1	16	237	253	237	
2	442	4	446	442	
total	458	241	699	679	0.97

Table-3 shows the result of the proposed method for balance scale data set, the objects of class "B" is evenly distributed among three clusters whereas 169 of class "L" belong to the cluster 1 and 156 of class "R" belong to the third cluster.

TABLE –3 CONFUSION MATRIX FOR BALANCE SCALE DATA SET

(MODE 1)

Class/ category	B	L	R	Max.	total
1	22	169	60	169	251
2	8	20	72	72	100
3	19	99	156	156	274
total	49	288	288	397	625
Purity rate = 397/625 = 0.6352					

TABLE –4 CONFUSION MATRIX FOR BALANCE SCALE DATA SET

(MODE 2)

Class/ category	B	L	R	Max.	total
1	18	90	93	93	201
2	17	96	112	112	225
3	14	102	83	102	199
total	49	288	288	307	625
Purity rate = 307/625 = 0.4912					

Table – 4 shows the results of balance scale data set for mode 2. Objects are distributed evenly in three clusters. For mode1 the purity rate is 63% whereas when using mode2 the purity rate is only 49%. As in K-modes the proposed method also depends on the initial selection of modes.

Table–5 lists the purity rate of K-Modes and the proposed Bayesian method for all eight data sets. We have taken the value of 'K' as the actual classes or categories in the data set. From Table-5 we have found that the proposed Bayesian method is efficient than the K-modes algorithm. Both the methods produce clusters with the same purity for car evaluation data set. When the K-modes method is used to cluster the objects at least three and at most we need five

iterations to get the optimal result.

TABLE 5 COMPARATIVE RESULTS OF K-MODES AND BAYESIAN METHOD

Data set	Number of Clusters	K-Modes	Bayesian Method
Mushroom	2	0.59	**0.615**
Congressional Votes	2	0.61	**0.88**
Soybean small	4	0.404	**0.787**
Breast cancer	2	0.65	**0.97**
Lymbhography	4	0.64	**0.68**
Car Evaluation	4	0.699	**0.699**
Hayes roth	3	0.440	**0.449**
Balance Scale	3	0.497	**0.604**
Nursery	5	0.386	**0.439**

Table-6 lists the purity rate of mushroom data set for K= 2 to 10. Except for K= 5 the proposed method is efficient than K-modes.

TABLE 6. PURITY RATE OF MUSHROOM DATASET

Number of Clusters	K-Modes	Bayesian Method
2	0.59	**0.615**
3	0.518	**0.603**
4	0.751	**0.958**
5	0.887	0.748
6	0.865	**0.949**
7	0.889	**0.923**
8	0.889	**0.929**
9	0.891	**0.930**
10	0.887	**0.919**

Purity rate for congressional votes data set for K = 2 to 10 are computed and tabulated in table-7. The proposed method is efficient than K-modes when K= 2, 3, 4, 7, and 8.

TABLE7. PURITY RATE OF CONGRESSIONAL VOTES DATASET

Number of Clusters	K-Modes	Bayesian Method
2	0.61	**0.88**
3	0.863	**0.909**
4	0.875	**0.909**
5	0.925	0.875
6	0.937	0.898
7	0.921	**0.921**
8	0.89	**0.89**
9	0.930	0.919
10	0.901	0.892

Proposed Bayesian method is efficient when K=2, 4 and 10 for soybean data set. (Table-8)

TABLE8. PURITY RATE OF SOYBEAN DATASET

Number of Clusters	K-Modes	Bayesian Method
2	0.532	**0.57**
3	0.787	0.702
4	0.404	**0.787**
5	0.936	0.851
6	0.912	0.702
7	0.89	0.809
8	0.915	0.766
9	0.979	0.745
10	0.71	**0.766**

Table-9 lists the purity rate of breast cancer data set. Table-10 lists purity rate for Lymbhography data set. The proposed method is efficient than K-modes for breast cancer, lymbhography and car evaluation dataset when K= 2 to 10. (Table – 11).

TABLE 9. PURITY RATE OF BREAST CANCER DATASET

Number of Clusters	K-Modes	Bayesian Method
2	0.65	**0.97**
3	0.943	**0.970**
4	0.916	**0.964**
5	0.943	**0.964**
6	0.948	**0.962**
7	0.936	**0.959**
8	0.946	**0.946**
9	0.924	**0.951**
10	0.855	**0.955**

TABLE 10. PURITY RATE OF LYMBHOGRAPHY DATASET

Number of Clusters	K-Modes	Bayesian Method
2	0.547	**0.667**
3	0.576	**0.695**
4	0.59	**0.64**
5	0.680	**0.797**
6	0.656	**0.747**
7	0.710	**0.755**
8	0.673	**0.795**
9	0.771	**0.772**
10	0.684	**0.766**

TABLE 11. PURITY RATE OF CAR DATASET

Number of Clusters	K-Modes	Bayesian Method
2	0.701	**0.701**

3	0.700	**0.700**
4	0.699	**0.699**
5	0.700	**0.700**
6	0.700	**0.700**
7	0.700	**0.750**
8	0.698	**0.769**
9	0.699	**0.720**
10	0.718	**0.72**

Bayesian method is efficient than K-modes for all values of K except when K= 2 and 4 for Hayesroth data set.(Table- 12).

TABLE 12. PURITY RATE OF HAYESROTH DATASET

Number of Clusters	K-Modes	Bayesian Method
2	0.410	0.393
3	0.440	**0.449**
4	0.434	0.409
5	0.460	**0.536**
6	0.449	**0.489**
7	0.468	**0.507**
8	0.479	**0.511**
9	0.475	**0.500**
10	0.514	**0.566**

For balance scale data set, Bayesian method is efficient than K-modes except when K= 5. (Table – 13).

TABLE 13 PURITY RATE OF BALANCE SCALE DATASET

Number of Clusters	K-Modes	Bayesian Method
2	0.545	**0.566**
3	0.497	**0.604**
4	0.499	**0.528**
5	0.616	0.590
6	0.547	**0.574**
7	0.625	**0.627**
8	0.547	**0.695**
9	0.620	**0.691**
10	0.616	**0.633**

From the above tables it is clear that, except the soybean data set and the voting congress data set in all the methods above 50% of the cases the proposed Bayesian clustering is efficient than K-modes. For all cases purity rate is above 50% except one or two.

To verify the scalability of the algorithm it is experimented for the nursery data set and the actual time taken is shown in figure-1. The proposed method increase linearly, and take less time compared to K-modes. The execution time of K-modes depends on the iterations needed to minimize the objective function. So there is a variation in the execution time when the size of the data set increases. But in the proposed method this discrepancy is eliminated and it is linearly proportional to the size of the data set and the number of attributes. Dynamically objects can be placed in to the existing clusters based on the summary information of clusters so it is not necessary to have the whole data set in primary memory as in EM algorithm. It is enough to have the attribute values and the frequency of them in the memory. As the objects are clustered based on the information in the clusters it is suitable for very large data set.

D. Sensitivity to the order of input:

To check the sensitivity of the input order, data set is reordered in ten different ways and checked with same set of modes and the results were tabulated in table-14.

TABLE -14 RESULTS OF BREAST CANCER DATA SET FOR DIFFERENT ORDER OF INPUTS

Purity rate
0.96
0.97
0.97
0.97
0.97
0.97
0.97
0.97
0.97
0.97

From the above results there is no difference in the clusters obtained for different order of inputs. Thus the proposed method is insensitive of input order.

E. Computational Complexity:

Complexity of the proposed method is O(dnk) where d is the number of attributes, n is the number of objects and k is the number of clusters. Complexity of K-modes algorithm is O(tdnk) where t represents the number of iteration. Thus the proposed method is scalable.

From the results it is clear that

i) Proposed method is scalable

ii) No repetitive execution is needed

iii) Produces efficient clusters for large data set.

iv) Insensitive to the order of input.

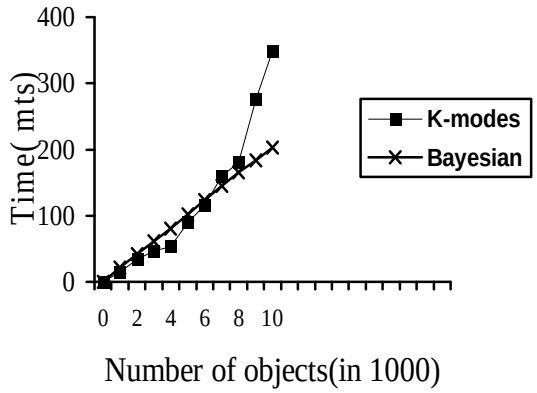

Figure-1. Comparison of actual time taken by K-modes and the proposed Bayesian method

VI. Conclusion

Compared to numerical clustering, categorical clustering seems to be more complicated. As the data mining deals with large data sets, the algorithms should be scalable. In algorithms based on similarity/dissimilarity matrix, the computational cost is high. K-modes is an efficient and scalable algorithm but a repetitive execution is needed to minimize the objective function. In the proposed method the repetition execution and updating of modes are not needed. Experimental results show that the proposed algorithm is efficient than K-modes and generates clusters with high purity and the results prove that the proposed method is effective in the case of large data sets when compared to small data sets. As no repetition is needed, the algorithm is scalable. As the results depend on the initial selection of modes, further we planned to extend this to mixed data and to improve the initialization methods.

References

[1] Arun.K.Pujari , Data Mining Techniques, Universities Press,2001,pp 114-147

[2] Cheeseman P, Stutz J (1996). "Bayesian Classification (AutoClass): Theory and Results." In PS U Fayyad G Piatetsky-Shapiro, R Uthurusamy (eds.), "Advances in Knowledge Discovery and Data Mining," pp. 153–180. AAAI Press/MIT Press.

[3] 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, 582 - 589 , 2002

[4] Dae-won kim, Kwang H.Lee, Doheon Lee, Fuzzy clustering of categorical data using centroids, Pattern recognition letters 25,1263-1271, Elsevier, (2004).

[5] George Karypis, Eui-Hong (Sam) Han, and Vipinkumar CHAMELEON: A hierarchical clustering algorithm using dynamic modeling, IEEE Computer, 1999.

[6] Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, Harcourt India Private Limited, 2001,2nd edition, pp 83,94,383-433.

[7] Ohm 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, 241–247, 2004

[8] Pavel Berkhin, Survey of Clustering Data Mining Techniques, Technical report, Accrue software, 2002

[9] Periklis Andristos, Clustering Categorical Data based On Information Loss Minimization, EDBT 2004: 123-146.

[10] 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

[11] Venkatesh Ganti, Johannes Gehrke, Raghu Ramakrishnan. CACTUS –Clustering Categorical Data using summaries, In Proc. of ACM SIGKDD, International Conference on Knowledge Discovery & Data Mining, 1999, San Diego, CA USA.

[12] www.ics.uci.edu/~mlearn/MLRepository.html

[13] Yang, Y., Guan, X., and You, J. 2002. CLOPE: a fast and effective clustering algorithm for transactional data, In Proceedings of the Eighth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Edmonton, Alberta, Canada, July 23 - 26, 2002). KDD '02, ACM Press, New York, NY, 682-687.

[14] Zengyou He, Xiaofei Xu, Shengchun Deng, Squeezer: An Efficient algorithm for clustering categorical data, Journal of Computer Science and Technology, Volume 17 Issue 5, 2002, Editorial Universitaria de Buenos Aires.

[15] .Zengyou He, Xiaofei Xu, Shengchun Deng, Bin Dong, K-Histograms: An Efficient Algorithm for Categorical Data set, www.citebase.org.

[16] 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.

[17] Zhexue Huang, Extensions to the K-means algorithm for clustering Large Data sets with categorical value, Data Mining and Knowledge Discovery 2, 283-304, Kluwer Academic publishers, 1998.

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, Rough sets and fuzzy logic.

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 Mathematics department, Gandhigram Rural University, in 1999. Currently he is working as a Professor in Computer Science Department, Periyar University, Salem and his experience in teaching started from 1988. His areas of interest include Medical Image processing, Artificial Intelligence, Neural Network, Data Mining, rough sets, Web mining, and fuzzy logic.