International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:4, No:1, 2010

Incremental Algorithm to Cluster the Categorical Data with Frequency Based Similarity Measure

S.Aranganayagi and K.Thangavel

***Abstract*—**Clustering categorical data is more complicated than the numerical clustering because of its special properties. Scalability and memory constraint is the challenging problem in clustering large data set. This paper presents an incremental algorithm to cluster the categorical data. Frequencies of attribute values contribute much in clustering similar categorical objects. In this paper we propose new similarity measures based on the frequencies of attribute values and its cardinalities. The proposed measures and the algorithm are experimented with the data sets from UCI data repository. Results prove that the proposed method generates better clusters than the existing one.

***Keywords*—**Clustering, Categorical, Incremental, Frequency, Domain

I. Introduction

CLUSTERING is the unsupervised classification of patterns into groups. Data clustering is a data analyzing technique and has been considered as a primary data mining method for knowledge discovery. Clustering is defined as the process of grouping most similar/homogeneous objects [3, 9]. In clustering, the object has to be grouped without the prior knowledge about the number of classes or groups. Clustering is broadly divided into partitional and hierarchical. In partitional algorithms, data set is divided in to 'K' partitions, where 'K' is the number of clusters. Finding appropriate 'K' is a complicated task without prior information. In hierarchical clustering algorithms, the objects are grouped /divided based on the merging criteria. The merging process is continued until we get the desired number of clusters. When the size of the data set increases then the computational time increases exponentially. Always it is not possible to get the actual number of clusters needed, instead it may be less than or greater than 'K'.

Results of clustering methods depend on the similarity measure used to group the data. Based on the geometric property of the data numerous similarity measures exist. Categorical data is the one which cannot be ordered and with limited domains. Due to the special properties of the categorical data, the geometrical measurements are not applicable. Thus the categorical similarity measures are based on the matching attribute values.

The partitional method does not remove the outliers; moreover it places the outlier in the nearest cluster. For example consider the data set with 12 attributes and with the number of clusters 'K' as 2, with modes Q1 and Q_2. Let the distance between the object *t* and Q_1, Q_2 be, d (t, Q1) = 10 and d (t, Q2) = 11, using the similarity measure proposed by Huang. Minimum of these two is 10. Thus the object *t* is placed in the first cluster, but out of 12 attributes only 2 are equal. Thus the less similar objects are also placed in some cluster. Based on the threshold value when the objects are grouped, less similar objects will be placed in a separate cluster. If the similarity matrix is used to group the objects, memory constraint is a problem. Most of the hierarchical algorithm uses this approach. To alleviate these problems the objects has to be grouped incrementally.

Non-incremental methods process all data patterns at a time. These algorithms require the entire datasets being loaded into memory and therefore have high requirement in memory space. A major advantage of the incremental clustering algorithm is limited space requirement since the entire data set is not in memory. Therefore these algorithms are well suited for a dynamic environment and for very large datasets[8]. Efficiency of the clusters obtained depends on the similarity measures and the threshold value. Hence in the paper, incremental algorithm is proposed, with new similarity measures. The similarity measures proposed are based on the frequency of attribute values and the cardinality of the domain of attribute values. The proposed method reads the object one by one and either places the object in the existing cluster or forms a new cluster based on the threshold value. The first object is considered as a seed or mode of the first cluster. Remaining objects are read and the similarity between the object and the representative of the existing clusters are computed as per the measures proposed. Maximum similarity is chosen and it is compared with the threshold value, if it is greater than the threshold value then the object is placed in the existing cluster else form a new cluster with the object as an initial seed. Resultant clusters are analyzed, after changing the threshold value; again the algorithm is executed for the same

Aranganayagi 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@gmail.com).

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

data set. From these results, the best number of clusters 'K' can be selected.

The similarity measures proposed in this paper are

i) Based on the frequency of matching attribute value in the cluster with respect to the number of attributes.
ii) Based on the frequency of matching attribute value with respect to domain and the cluster size.
iii) Based on the cardinality of the domain of matching attribute values.
iv) Products of the frequency of matching attribute values in the data set and in the cluster.[2]

These measures are applied to a single pass incremental algorithm. The fourth measure is the variation of dissimilarity measure proposed by author in [2]. The proposed measures were experimented with the data sets obtained from UCI data repository and it is compared with the similar methods Squeezer and M-Squeezer.

The rest of the paper is organized as follows section 2 briefs the related work, section 3 discusses a definitions and notations used, section 4 describes the proposed method, Section 5 discusses the experimental results and section 6 concludes the paper.

II. RELATED WORK

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 agglomerative hierarchical clustering algorithm, group the objects based on the links, where the number of links between two objects is the number of common neighbors that they have in the dataset [3, 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 [13]. The COOLCAT algorithm uses the entropy measure in clustering. Objects are placed in the cluster with minimum entropy [5]. 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 [11].

K-Modes cluster the categorical data using modes instead of means[17]. By varying the dissimilarity measure, K-Representative, K-Histogram and Improved K-Modes using weighted measures[1] were proposed. In K-Representative the measure relative frequency such as the frequency of attribute value in the cluster divided by cluster length is used as a measure[10]. In K-Histogram, histograms were used instead of modes, distance between the object and the histogram is computed by adding the frequency of attribute values if the value is present in the histogram, and the object is placed in the cluster with minimum distance[16]. QROCK is a hierarchical clustering algorithm, uses the new similarity measure to group the data[7]. Squeezer reads the object one by one and places it in the existing cluster or form a new cluster based on the average similarity. Sample of the data set is used to compute the average similarity[15]. M-Squeezer proposed by author[1], reads the object one by one and clusters the object based on the threshold value. The simple mismatching measure is used to find the distance between two objects.

III. DEFINITIONS AND NOTATIONS

Consider the dataset T with 'm' attributes and 'n' objects $T=\{X_1, X_2,...,X_n\}$. The attribute set A is defined as $A=\{A_1, A_2,...,A_m\}$. The set of domain of the *m* attribute is $D=\{D_1, D_2,...,D_m\}$, the domain of ith attribute A_i, D_i contains 's' distinct values, such as $|D_i|=s_i$, where $D_i = \{v_{i_1}, v_{i_2}, ..., v_{i_{s_i}}\}$. Cluster representatives Q is $\{Q_1, Q_2,...,Q_k\}$.

Where each X_i and Q_i is defined as $X_i=\{x_{i1}, x_{i2},...,x_{im}\}$ and $Q_i=\{q_{i1}, q_{i2},...,q_{im}\}$ respectively.

$f(v_{i_j}/T)$ = Number of objects in the data set T with the attribute value v_{i_j}

$f(v_{i_j}/C_l)$ = Number of objects in the cluster C_l with the attribute value v_{i_j}

$f(C_l)$ = Number of objects in the cluster C_l

Definition – **1:** Weighted measure based on the frequency of matching attribute value in the cluster with respect to the number of attributes.(WMFAC)

$$sim(X_i, Q_j)=\sum_{l=1}^{m}\partial(x_{il}, q_{jl}) \quad (1)$$

$$\partial(x_{il}, q_{jl})=\begin{cases}\dfrac{f(v_{l_t}/C_j)}{f(C_j)m} & \text{if } x_{il}=q_{jl}\\ 0 & \text{if } x_{il}\neq q_{jl}\end{cases} \quad (2)$$

Definition – **2:** Weighted measure based on the frequency of matching attribute value with respect to domain and the cluster size.(WMFACD)

$$sim(X_i, Q_j)=\sum_{l=1}^{m}\partial(x_{il}, q_{jl}) \quad (3)$$

$$\partial(x_{il}, q_{jl}) = \begin{cases} \dfrac{f(v_{l_t}/C_j)}{f(C_j)|D_l|m} & \text{if } x_{il} = q_{jl} \\ 0 & \text{if } x_{il} \neq q_{jl} \end{cases} \quad (4)$$

Definition – **3:** Weighted measure based on the cardinality of the domain of matching attribute values. (WMCD)

$$sim(X_i, Q_j) = \frac{\sum_{l=1}^{m} \partial(x_{il}, q_{jl})}{m + 1/\sum_{l=1}^{m} w_l} \quad (5)$$

$$\partial(x_{il}, q_{jl}) = \begin{cases} 1 & \text{if } x_{il} = q_{jl} \\ 0 & \text{if } x_{il} \neq q_{jl} \end{cases} \quad (6)$$

$$w_l = \begin{cases} |D_l| & \text{if } x_{il} = q_{jl} \\ 0 & \text{if } x_{il} \neq q_{jl} \end{cases} \quad (7)$$

The proposed measure is a variation of the measure proposed in QROCK [7], instead of $|X_i \cap X_j|$ the count of matching attributes is used .

The measure $|X_i \cap X_j|$ and $|X_i \cup X_j|$, yields different results with respect to the values of the attribute. Categorical or nominal data contains more or less same values for the attribute. For example consider two objects from Zoo data set,

$$Xi = \{0001011111010101\}$$
$$Xj = \{0001011111010101\}$$

$|X_i \cap X_j| = \{0,1\} = 2$

$sim(X_i, X_j) = 16$ using (6).

This data set contains only {0, 1} as a possible value for each attribute, thus intersection of two attributes is always 2. Domain of all the attribute in the zoo data set contains only 0 and 1 as the values. Consider the case where there is no match between the two objects X_i and X_j , then

$|X_i \cap X_j| = |\{0,1\}| = 2$ and $|X_i \cup X_j| = |\{0,1\}| = 2$

Thus similarity between the objects is equal to 1, but no attributes are equal, whereas when the proposed measure is used the similarity will be the proportion of the matching attribute value with the summation of the cardinalities of the domain. Hence the object can be placed in the nearest cluster.

Similarly if two objects from mushroom data set are compared, such as

X_i = [x s n t p f c n k e e s s w w p w o p k s u]
X_j = [x s n t p f c n k e e s s w w p w o p n s u]
$|X_i \cap X_j| = |\{x\ s\ n\ t\ p\ f\ c\ k\ e\ w\ o\ u\}| = 12$

$sim(X_i, X_j) = 21$ using (6) .

Definition – 4: Weighted Measure based on the product of proportion of the frequency of matching attribute values in the data set and in the cluster. (WMPFA)

When the categorical objects are grouped, the frequency of the attribute values plays an important role.

$$sim(X_i, Q_j) = \sum_{l=1}^{m} \partial(x_{il}, q_{jl}) \quad (8)$$

$$\partial(x_{il}, q_{jl}) = \begin{cases} w_l & if \quad x_{il} = q_{jl} \\ 0 & if \quad x_{il} \neq q_{jl} \end{cases} \quad (9)$$

If $x_{il} = q_{jl} = v_{l_t}$

$$w_l = \left(\frac{f\left(v_{l_t}/C_j\right)}{|C_j|} \right) \left(\frac{f\left(v_{l_t}/T\right)}{|T|} \right) \quad (10)$$

IV. PROPOSED INCREMENTAL ALGORITHM

Incremental clustering algorithms are dynamic, and it is enough to have the summary of the cluster in the memory, thus no memory constraint. The proposed algorithm is similar to M-Squeezer. By varying the threshold value, we can select the best number of clusters from the result obtained. Threshold value for the measure WMFAC, WMFACD and WMCD varies from 0 to 1, whereas for the measure WMPFA, the minimum threshold value is zero and the maximum threshold value depends on the summation of the maximum of v_{i_k} for each attribute i divided by the total number of objects 'n'. Hence we get different threshold values for each data set.

A. Proposed Incremental Algorithm

Input:

Data set *T* with *m* attributes and *n* objects, Threshold value, '*th*'

Output: K clusters

Step 1 : Initialize 1 to i and k.
Step 2: Compute the Domain(i) for all *m* attributes.
Step 3: If (i == 1) then call Newcluster(i, X_i, k)

Step 4: For i = 2 to n do

4.1Compare the object with modes of existing clusters.

4.2 Compute m1 = maximum (sim(X_i, Q_j)).

4.3 if m1 > *th* then add the i^{th} object in the j^{th} cluster and update the modes else call N*ewcluster(i,* X_i , k).

Step 5: Output K clusters.

NewCluster(i, X_i, k)

Step 1: Increment the number of clusters by *k* by one.

Step 2: Create a new cluster with i^{th} object as a member and assign X_i as a mode of the cluster.

V. Experiments

Validating the clustering results is a complicated task. Here we have used the external quality measure Purity to evaluate the clustering results obtained.

A. Purity Measure

A cluster is called a pure cluster if all the objects belong to a single class. The clustering accuracy *r* is defined as,

$$r = 1/n \sum_{i=1}^{k} a_i \qquad (11)$$

where a_i is the number of data objects that occur in both cluster C_l and its corresponding labeled class, which has the maximal value 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%, it means that it has only pure clusters. Large clustering accuracy implies better clustering. Low error rate indicates the best clustering [18].

The proposed incremental algorithm is similar to Squeezer and M-Squeezer, hence it is compared with these two methods. To evaluate the efficiency of the proposed method it is experimented with the data sets such as soybean small, Zoo, congressional votes and mushroom from UCI data repository. Objects are grouped based on the cluster summary and the average similarity in Squeezer. Similarity matrix is constructed to compute the average similarity between objects. One tenth of the data set is taken as a sample to find the average similarity. The simple mismatching measure is used as a dissimilarity measure in M-Squeezer. Hence lower threshold value groups more similar objects. Thus the experimentation is carried out for different threshold values from 0.2. The proposed measures are similarity measures, thus the clustering process is started with the threshold value of 0.6. By decreasing and increasing the threshold value the process is repeated. From the resultant clusters, based on the quality measure best clusters are selected.

B. Data set

1) Soybean data set

The soybean data set consists of 47 cases of soybean disease each characterized by 35 multivalued categorical values. These cases are drawn from four population each one of them representing one of the following four diseases. D1 – Diaporthe stem canker, D2- Charcoat rot, D3- Rhizoctonia root rot and D4 – Phytophthorat rot. Attributes with unique values are omitted for clustering. Except for Phytophthora Rot that has 17 instances, all other diseases have 10 instances each.

Average similarity of soybean data set is 10. We tested the original squeezer with s = 10. The results are tabulated in table-I. M-Squeezer algorithm is executed for threshold values from 0.2 to 0.5 and the results are shown in table –II.

The proposed algorithm with measure WMFAC, WMFACD, WMCD and WMPFA is executed for different threshold values. The results are tabulated in Table–III, IV, V and VI. M-Squeezer generates clusters with 100% purity but the number of clusters obtained is 9. Whereas the proposed measures generate clusters with 100% purity and the number of clusters obtained is equal to the actual class labels.

TABLE I
RESULTS OF SQUEEZER FOR SOYBEAN DATA SET

Average Similarity	Purity	Number of clusters Obtained	Number of clusters selected
10	0.79	3	3

TABLE II
RESULTS OF M-SQUEEZER FOR SOYBEAN DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
0.2	1	18	12
0.3	1	9	9
0.4	0.9362	5	5
0.5	0.9574	4	4

TABLE III
RESULTS OF WMFAC FOR SOYBEAN DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
0.3	0.5745	2	2
0.4	0.7872	3	3
0.5	1.0	4	4

TABLE IV
RESULTS OF WMFACD FOR SOYBEAN DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
0.2	1.0	4	4
0.3	1.0	10	7

TABLE V
RESULTS OF WMCD FOR SOYBEAN DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
0.5	0.9574	4	4
0.6	0.9362	5	4
0.7	1	9	9
0.8	1	18	6

TABLE VI
RESULTS OF WMPFA FOR SOYBEAN DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
5	0.766	3	3
6	1	4	4
7	1	7	4
8	1	13	5

Fig-1 shows the comparative results of clusters obtained and the clusters selected based on high purity. Fig-2 depicts the comparative chart of purity measure of all the six methods for soybean data set.

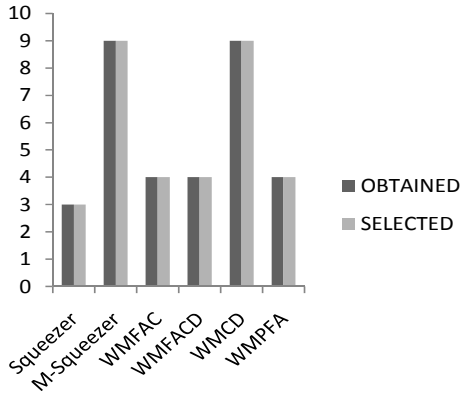

Fig. 1 Clusters obtained based on higher purity for soybean data set

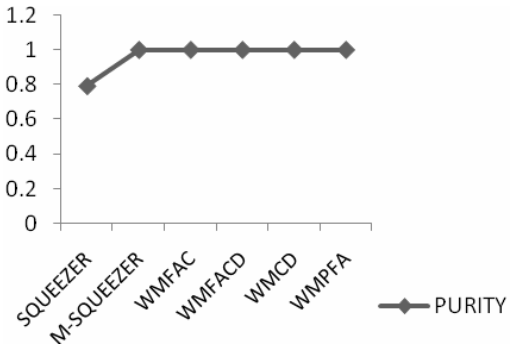

Fig. 2 Purity rate of soybean data set

C. Results of Zoo Data Set

This data set contains 101 instances of animals with 18 features. Each attribute describes the characteristics of animals like feathers, airborne, backbone, fins, leg and so on. The name of the animal constitutes the first attribute. This attribute is neglected. The character attribute corresponds to the number of legs lying in the set 0, 2, 4, 5, 6, and 8. The data set consists of 7 different categories of animals. Average similarity of this data set is 13. Squeezer results in 11 clusters(Table- VII), after removing the small clusters, eight clusters are selected. All the eight clusters are pure clusters. Table-VIII shows the result of M-Squeezer for the threshold values ranges from 0.2 to 0.5. If the higher purity is selected the number of clusters obtained is 11. But close to the actual class labels we get 6 clusters when threshold value is 0.3. Table-IX to Table-XII displays the results of the proposed measures. Depends on the nature of the dataset, clusters with high purity are obtained for different threshold values. The measure based on cardinality of the domains WMFACD and WMCD generate high purity clusters. But in WMCD, out of 32 clusters only 3 are selected, whereas WMFACD selects 12 clusters out of 17 clusters. The measures WMFACD and WMPFA generates clusters equal to the actual class labels, ie k =7. Fig-3 shows the number of clusters obtained and number of clusters selected based on the higher purity for zoo data set. Fig – 4 shows the purity rate of the clusters selected.

TABLE VII
RESULTS OF SQUEEZER FOR ZOO DATA SET

Average Similarity	Purity	Number of clusters Obtained	Number of clusters selected
13	1	11	8

TABLE VIII
RESULTS OF M-SQUEEZER FOR ZOO DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
0.2	0.9604	14	11
0.3	0.88	7	6
0.4	0.79	5	4
0.5	0.72	3	3

TABLE IX
RESULTS OF WMFAC FOR ZOO DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
0.5	0.7822	4	4
0.6	0.8317	5	5

TABLE X
RESULTS OF WMFACD FOR ZOO DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
0.2	0.6040	3	3
0.3	0.8218	6	6
0.4	0.9802	17	12
0.35	0.9010	10	7

TABLE XI
RESULTS OF WMCD FOR ZOO DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
0.4	0.6040	2	2
0.5	0.7129	3	3
0.6	0.7822	5	3
0.7	0.8713	7	4
0.8	0.9604	14	3
0.9	0.9802	32	3

TABLE XII
RESULTS OF WMPFA FOR ZOO DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
6	0.8218	6	4
7	0.9109	8	6
8	0.9604	22	7

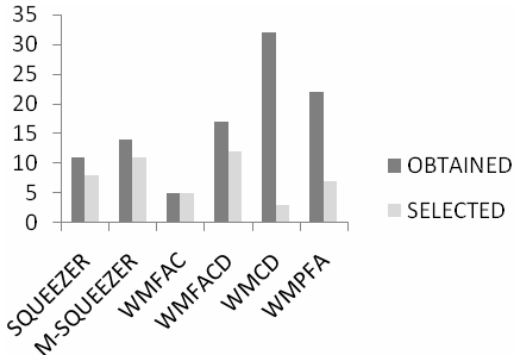

Fig. 3 Clusters obtained based on higher purity for zoo data set

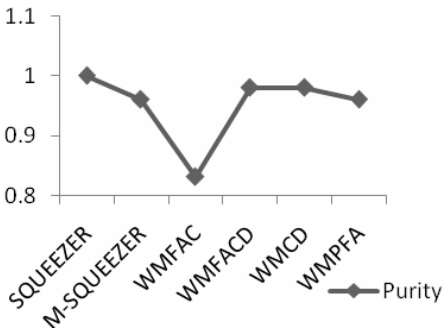

Fig. 4 Purity rate of zoo data set

D. Results of congressional votes data set

This data set is the United States Congressional voting records in 1984. Total number of records is 435. Each row corresponds to one Congress man's vote on 16 different issues (e.g., education spending, crime etc.). All attributes are Boolean with Yes (that is, 1) and No (that is, 0) values. A classification label of Republican or Democrat is provided with each data record. The data set contains records for 168 Republicans and 267 Democrats. Squeezer generates eight clusters with error rate of 0.06. But only 8 clusters are selected from 52 clusters(Table- XIII). M-Squeezer is executed for the range from 0.2 to 0.5. Table-XV to XVIII shows the results of proposed measures. Squeezer and M-Squeezer generates large number of clusters like 52 and 104 respectively. All the proposed measures generates small clusters with k = 2, where each tuple in the data set is classified into yes or no. Fig -5 and Fig -6 depicts the clusters selected and the purity rate of the clusters respectively.

TABLE XIII
RESULTS OF SQUEEZER FOR CONGRESSIONAL VOTES DATA SET

Average Similarity	Purity	Number of clusters Obtained	Number of clusters selected
10	0.94	52	8

TABLE XIV
RESULTS OF M-SQUEEZER FOR CONGRESSIONAL VOTES DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
0.2	0.9668	104	5
0.3	0.9561	30	7
0.4	0.9525	15	6
0.5	0.8851	12	5

TABLE XV
RESULTS OF WMFAC FOR CONGRESSIONAL VOTES DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
0.1	0.7143	3	2
0.2	0.8690	3	2
0.3	0.8805	4	3
0.5	0.9057	16	7
0.6	0.9425	36	7

TABLE XVI
RESULTS OF WMFACD FOR CONGRESSIONAL VOTES DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
0.1	0.8805	4	3
0.08	0.8690	3	2

TABLE XVII
RESULTS OF WMCD FOR CONGRESSIONAL VOTES DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
0.1	0.6736	3	2
0.2	0.8575	3	2
0.3	0.8598	3	2
0.4	0.8805	4	3
0.5	0.8828	12	5
0.6	0.9494	17	7

TABLE XVIII
RESULTS OF WMPFA FOR CONGRESSIONAL VOTES DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
1	0.8644	5	2
2	0.8736	5	2
3	0.8920	8	5
4	0.9057	20	7
5	0.9586	32	6

E. Results of mushroom dataset

Each object describes the physical characteristics of mushroom like color, shape, odour etc. This data set contains 8124 objects with 23 attributes. A classification of edible or poisonous is attached with each instance. The number of edible and poisonous mushrooms in the dataset is 4208 and 3916 respectively. Average similarity for mushroom data set is 16. Number of clusters obtained is 22 in Squeezer algorithm. Results are shown in Table-XIX. Most of the resultant clusters are pure clusters. M-Squeezer method generates 29 clusters with same purity(Table- XX). WMFAC generates 15 clusters with purity of 0.97, and also this algorithm produce cluster equal to actual class labels (k=2) with the purity of 60%(Table-XXI). WMFACD outputs 25 clusters with 99% of purity. Only one cluster is omitted as an outlier(Table-XXII). Table-XXIII shows the results of WMCD. 31 clusters are obtained for with purity of 99%. Whereas the WMPFA measure generates 21 clusters with the purity of 96% . But the generated clusters are 103, after removing the outliers only 21 were selected(Table–XXIV). Fig-7 depicts the clusters obtained based on higher purity. Fig-8 displays the purity rate of the selected clusters. Comparatively the ratio between the clusters obtained and selected is low in all the proposed measures except WMPFA.

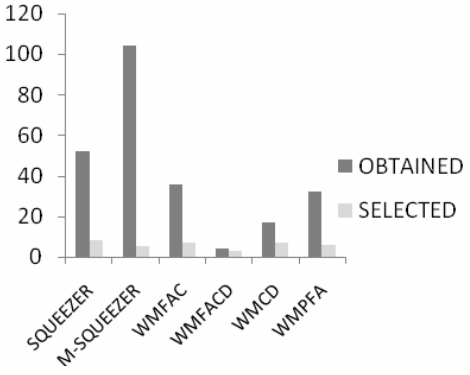

Fig. 5 Clusters obtained based on higher purity for congress data

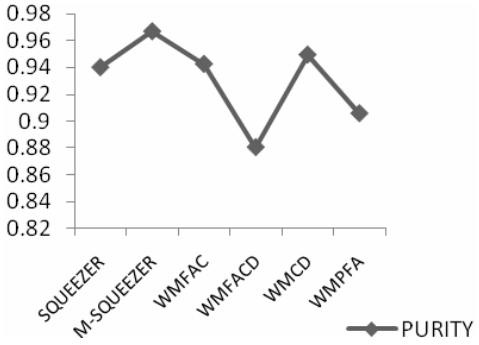

Fig. 6 Purity rate of congressional votes data set

TABLE XIX
RESULTS OF SQUEEZER FOR MUSHROOM DATA SET

Average Similarity	Purity	Number of clusters Obtained	Number of clusters selected
16	0.99	24	22

TABLE XX
RESULTS OF M-SQUEEZER FOR MUSHROOM DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
0.3	0.9896	31	29
0.4	0.9571	19	18
0.5	0.9181	13	13

TABLE XXI
RESULTS OF WMFAC FOR MUSHROOM DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
0.2	0.6073	2	2
0.3	0.7629	6	6
0.4	0.8874	9	8
0.5	0.9696	16	15

TABLE XXII
RESULTS OF WMFACD FOR MUSHROOM DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
0.1	0.7771	5	5
0.2	0.9903	27	25
0.15	0.9343	13	12

TABLE XXIII
RESULTS OF WMCD FOR MUSHROOM DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
0.3	0.7792	4	4
0.4	0.8844	9	8
0.5	0.8999	12	12
0.6	0.9515	19	17
0.7	0.9930	32	31

TABLE XXIV
RESULTS OF WMPFA FOR MUSHROOM DATA SET

Threshold value	Purity	Number of clusters Obtained	Number of clusters selected
4	0.6928	2	2
5	0.8859	8	7
7	0.9554	103	21

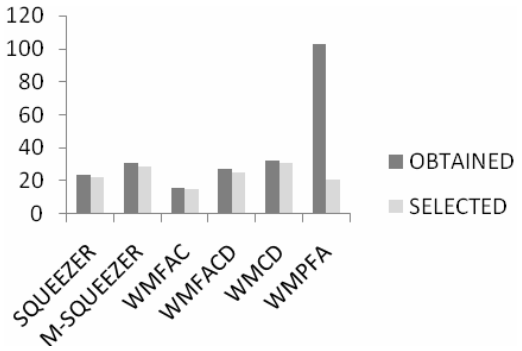

Fig.7 Clusters obtained based on higher purity for mushroom data

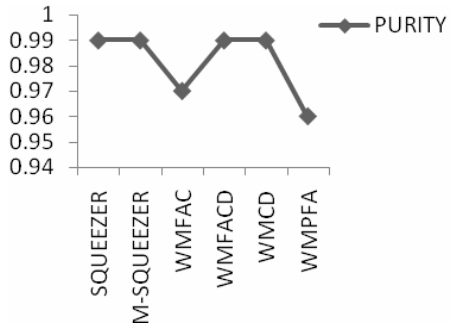

Fig. 8 Purity rate of mushroom data set

Results show that the proposed measures yield better results. The proposed measure generates the cluster which is closer to actual class labels. Instead in Squeezer and M-Squeezer we get more number of clusters with one or two element. Purity rate is computed for the number of clusters selected. Clusters with less than 10 objects are considered as small for large data sets. These objects are considered as outliers as in Squeezer. For soybean and Zoo data set the singleton clusters are omitted.

The fig-9 shows the number of clusters selected for all the six methods. Results prove that the proposed methods generate less number of clusters with better quality.

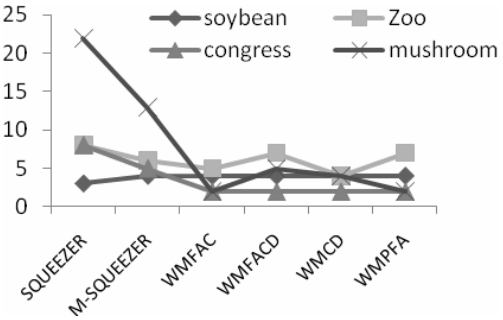

Fig. 9 Comparative chart of best clusters selected

VI. CONCLUSION

Incremental algorithm finds clusters in less computation time. Only the summarized information is stored in the memory, thus the memory and I/O constraint is avoided in the proposed method. Experimental results prove that the frequency based measure generates clusters with high purity and the number of clusters generated is also small. In general the incremental algorithms generate large number of clusters; naturally the purity is also more, whereas the proposed measures generate less number of clusters with high purity. As this is a single pass algorithm no iteration is necessary. Hence the proposed method is capable of clustering large data set.

REFERENCES

[1] Aranganayagi.S and K.Thangavel, "M-Squeezer Algorithm to Cluster the Categorical Data", Computational Mathematics, Narosa, Publishing House, New Delhi, India, 2009
[2] Aranganayagi.S and K.Thangavel, "Improved K-Modes for Categorical Clustering Using Weighted Dissimilarity Measure", International Journal of Computational Intelligence (IJCI), Vol.5, No.2, pp.182-189,WASET, spring 2009.
[3] Arun.K.Pujari, "Data Mining Techniques", University Press, 2001.
[4] Ching- San Chiang, Shu-Chuan Chu, Yi-Chih Hsin and Ming-Hui Wang, "Genetic Distance measure for K-modes Algorithm", International Journal of Innovative Computing and Information and Control, Vol.2 , 2006, pp 33 -40.
[5] 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.
[6] Dae-won kim, Kwang H.Lee, Doheon Lee, "Fuzzy clustering of categorical data using centroids", Pattern recognition letters 25, Elseveir, (2004), 1263-1271.
[7] Dutta, M. and Mahanta, A. Kakoti and Pujari, Arun K., "QROCK a quick version of the ROCK algorithm for clustering of categorical data, Pattern Recogn. Letters, volume = {26}, 2005, 2364 – 2373, Elsevier Science Inc
[8] Hsu.C.C., & Huang,Y.P., "Incremental Clustering of Mixed Data Based on the Distance Hierarchy", Expert System with Applications,(2007),doi:10.1016/j/eswa 2007.08.049
[9] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques", Harcourt India Private Limited, 2001.
[10] 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.

[11] Periklis Andristos, “Clustering Categorical Data based On Information Loss Minimization”, EDBT 2004: 123-146.
[12] 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
[13] 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.
[14] www.ics.uci.edu/~mlearn/MLRepository.html
[15] 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.
[16] Zengyou He, Xiaofei Xu, Shengchun Deng, Bin Dong, ”K-Histograms: An Efficient Algorithm for Categorical Data set”, www.citebase.org.
[17] 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.
[18] 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, 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 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, Neural Network, Data Mining, rough sets, Web mining, and fuzzy logic.