Extended K-Modes with Probability Measure

Aranganayagi.S[1] and Thangavel. K[2]

Abstract— **Clustering Categorical data is more complicated process than numerical clustering. In this paper the traditional K-Modes algorithm is extended with the weighted measure based on the probability of respective matching attribute value in the data set. The proposed method is experimented with the data sets obtained from UCI data repository. Results prove that the proposed weighted measure is superior to K-Modes.**

***Index Terms*—Clustering, Categorical Data, K-Modes, Probability, Weighted measure**

I. Introduction

Clustering can be defined as the process of organizing objects in a database in to clusters or groups so that the objects within the same cluster have a high degree of similarity, while the objects belonging to different clusters have a high degree of dissimilarity. Clustering algorithms are increasingly required to deal with large scale data sets, particularly in data mining. Most of the earlier work focused on numerical clustering, where geometric properties can be exploited to naturally to define distance functions between data points[1],[4],[5],[7]. Huang developed the K-Modes algorithm by extending the K-Means algorithm with a simple mismatching dissimilarity measure for categorical data, and a frequency based method to update modes during clustering. Huang proposed a K-Prototypes algorithm which is the combination of K-Modes and K-Means to cluster the mixed numerical and categorical attributes. K-Modes algorithm is unstable due to non-uniqueness of the modes[14], [15]. In all centroid based algorithms the resultant clusters are influenced by the selection of initial modes.

The simple mismatching measure does not use the inherent relationship between the attributes. Attributes with few values and attribute with more values are considered with the same weight in K-Modes thus we propose a new weighted measure based on the probability of attribute value. Out of n objects less frequent values are given less weight than the more frequent values. "Majority Vote" concept is used in the proposed measure. Few variations of K-Modes exist with relative frequency measure as a weight.

O.M.San et al proposed an algorithm based on the "cluster centers" called representatives for categorical objects[6]. As arithmetic operation is completely absent in the setting of categorical objects, the Cartesian products and the union operations are used for the formation of "cluster centers". Representative of C is defined by

$$Q = \{q_1, q_2, \ldots, \mathrm{q}_\mathrm{k}\} \tag{1}$$

with $q_j = \left\{ \left(c_j, f_{c_j} \right) \middle| c_j \in D_j \right\}$ where f_{c_j} is the relative frequency of category c_j within C. $f_{c_j} = {n_{c_j}}/{p}$ where n_{c_j} is the number of objects in C having category c_j at attribute A_j. The dissimilarity between object X and representative Q is defined by

$$d(X,Q) = \sum_{j=1}^{\mathrm{m}} f_{c_j} (1 - \partial(x_j, q_j)) \tag{2}$$

Zengyou He et al proposed a variant of K-Means called K-Histogram[14]. Each cluster is assigned with one histogram. Where the histogram H is defined as

$$H = \{h_1, \mathrm{h}_2, \ldots, \mathrm{h}_\mathrm{k}\} \tag{3}$$

and each h_i is defined as

$$h_i = \{(v_1, f_1), (v_2, f_2), \;.\;.\;.\;, (v_k, f_{p_k})\} \tag{4}$$

v_j refers to the attribute value and f_j refers to the frequency of the attribute value. Each object is compared with the histogram and placed in the cluster with minimum distance. Histogram is updated during each allocation. Instead of using crisp clusters, fuzzy clusters are formed using fuzzy K-Modes[3].

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 [1].

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], [12].

Clustering Categorical Data Using Summaries (CACTUS)

Manuscript received September 18, 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, IACSIT: Corresponding author, phone: 0424-2230855, 9842723085; e-mail: arangbas@gmail.com.

Dr.K.Thangavel is with the Periyar University, Salem, Tamilnadu, India as Professor in Computer Science. (drktvelu@yahoo.com)).

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 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 [2]. 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 [8]. 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[12].

Section-2 describes the definitions used Section-3 briefs the K-Modes algorithms Section-4 explains the proposed method Section-5 discuss on experiments and results and Section-6 concludes the paper.

II. Definitions and notations

Let $\mathrm{D} = \{x_1, x_2, \ldots, x_n\}$ be the data set with *m* categorical attributes. Each object is characterized with the attributes $\{A_1, A_2, \ldots, A_m\}$. Define the Attribute Value Set

$$Avs(A_i) = \{(v_{i_1}, f_{i_1}), (v_{i_2}, f_{i_2}), \; . \; . \; . \; , (v_{i_p}, f_{i_p})\} \quad (5)$$

Where the $Domain\,(A_i) = \{v_{i_1}, v_{i_2}, \; . \; . \; . \; , v_{i_p}\}$ contain *'p'* distinct values and f_{i_1} refers the frequency of the attribute value v_{i_1} in the data set D.

The dissimilarity between the object x_i and the centroid Q is defined as

$$d(x_i, q_j) = \sum_{\mathrm{i=1}}^{\mathrm{m}} w_l (1 - \partial(x_{i,l}, q_{j,l}) \quad (6)$$

$$\partial(x_{i,l}, q_{j,l}) = \begin{cases} 0 & x_{i,l} \; = \; q_{j,l} \\ 1 & x_{i,l} \; \neq \; q_{j,l} \end{cases} \quad (7)$$

$$w_l = \frac{f_{l_t}}{n} \quad if \;\; x_{i,l} = v_{l_t} \quad (8)$$

w_l is the probability of having the attribute value v_{l_t} in the data set D.

The objective function is similar to K-Modes and it is defined as

$$P(U, Q) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} u_{i,l} d(x_{ij}, q_{lj}) \quad (9)$$

$$\text{Subject to} \;\; \sum_{l=1}^{k} u_{i,l} = 1, \qquad 1 \leq i \leq n$$

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

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 Q 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.

III. Overview of K-Modes

K-Modes is the scalable, partitional clustering method, which partitions the data set in to *K* homogeneous groups based on the simple mismatching measure. *K* distinct records are selected at random and assumed as the initial modes. The process is started with initial modes. Each object is read in sequence and compared with the modes of *K* clusters. The object is placed in the cluster which results in the minimum distance. Modes are updated with most frequently occurring value of each attribute as an entry. The detailed algorithm is given below.

(5)

Steps of K-Modes algorithm:

Step 1: Select K initial modes.

Step 2: Allocate an object to the cluster whose mode is nearest to the object, using simple mismatching measure. Update the mode of the cluster after each allocation.

Step 3: After all objects have been allocated to the respective cluster, retest the objects with new modes and update the clusters.

Step 4: Repeat (2) and (3) until there is no change in clusters.[16]

IV. K-Modes with probability measure

In general the objective of the clustering is to group the most similar objects. When compared to numerical clustering the process is more complicated. K distinct records or most frequently occurring values are selected and assumed as initial modes or centroids. Each object is compared with the modes of the cluster using the definition (6). Object is placed in the cluster, which results in minimum dissimilarity 'd'. Update the mode of the cluster. Repeat the process with new updated modes, until there is no change in the cluster membership. The proposed algorithm is given below.

Algorithm : K-Modes with probability measure

Input : Data set D and the number of clusters 'K'
Output : 'K' groups

Step 1: Initialize Modes of K clusters.

Step 2: Compute the dissimilarity between the object and the modes of the clusters. Place the object in the cluster which results in minimum dissimilarity. Update the mode of the cluster.

Step 3: After all objects have been allocated to the respective cluster, retest the objects with new modes and update the clusters.

Step 4: Repeat steps 2 and 3 until there is no changes in clusters.

The proposed measure is experimented with the data sets from UCI data repository[11]. K distinct records are selected at random and considered as initial centroids or modes.

V. Experiments

This section contains the experimental results on the performance of the proposed Extended K-Modes with Probability measure. The Proposed method is similar to K-Modes, thus the results are compared with the traditional K-Modes. Five different sets of modes are selected at random and the same is applied to both K-Modes and the proposed method.

A. *Quality Measure*

To evaluate the quality of the clustering results we need cluster validation methods. For numerical data, the clustering structure is usually validated by the geometry and density distribution of the clusters. Because of the special property of the categorical data the geometric based quality measures are inapplicable or inefficient. We have taken the datasets with class labels from UCI for analysis[12]. Thus the external quality measures based on the class label such as purity and F-measure are used. The F-measure favors the coarser clustering.

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,

$$r = 1/n\sum_{i=1}^{k} a_i \tag{10}$$

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 [16].

F-Measure: Given g categories or classes K_h $(h \in \{1, 2, \ldots, g\})$. Let $\mathrm{n}^{(\mathrm{h})}$ be the number of objects in category K_h, n_l be the number of objects in the cluster C_l and $n_l^{(h)}$ denote the number of objects in cluster l and in the category h.

Precision and recall are the standard measure. Precision is the fraction of correctly retrieved objects.

$$P(c_l, k_h) = n_l^{(h)} / n_l \tag{11}$$

Recall is the fraction of correctly retrieved objects out of all matching objects in the database.

$$R(c_l, k_h) = n_l^{(h)} / n^h \tag{12}$$

The F-measure combines the precision and recall into a single number given a weighting factor. This is a combination of precision and recall with equal weights. The F-measure for entire clustering is defined as

$$F = \frac{1}{n}\sum_{h=1}^{g} n^h \max of\ l \left(\frac{2n_l^{(h)}}{n_l + n_h}\right) \tag{13}$$

Unlike purity, F-measure is not biased towards a large number of clusters.

B. *Data Set*

K-Modes and the extended K-Modes are executed with the data sets from UCI such as soybean small, breast cancer, Zoo, voting congress, mushroom, and lymbhography.

Michalski Soybean disease 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.

Wisconsin Breast cancer Data set:

The breast cancer data set consists of 699 cases with 10 attributes. The last attribute is a decision attribute, which determines the object as either benign or malign. The data set contains 458 benign cases and 241 malign cases.

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, 8. The data set consists of 7 different categories of animals.

Congressional Vote 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 mans votes 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.

Mushroom data set:

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.

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

Table-I shows the average purity measure of K-Modes and the proposed K-Modes with probability measure. Table-II depicts the average F-Measure of the data sets for both K-Modes and Extended K-Modes.

Fig-1 shows that the extended K-Modes is superior to the traditional K-Modes. The same set of modes is applied for both the algorithms. Extended K-Modes perform better than the original K-Modes, F-measure value of all the data sets is higher than the original K-Modes. The Extended K-Modes is similar to K-Modes thus it is scalable. The computational complexity of the proposed Extended K-Modes O(tkn + n). As in K-Modes resultant clusters of the proposed method depend on the initial selection of modes. Better initialization yields better clusters.

VI. CONCLUSION

Experimentation results prove that the proposed probability measure is efficient than the K-Modes. The same set of modes results in better quality clusters when the Extended K-Modes is used. In real world most of the data sets are very large, thus the proposed method is applicable to this. All variations of K-modes depend on the initial mode and the ordering of the data set. In Future we plan to extend this to produce efficient clusters with better initial seeds.

TABLE I. COMPARATIVE RESULTS OF K-MODES AND EXTENDED K-MODES FOR PURITY

Data set	Number of Clusters	K-Modes	Extended K-Modes
Mushroom	2	0.54	**0.60**
Congressional Votes	2	0.61	**0.89**
Soybean small	4	0.37	**0.83**
Breast cancer	2	0.65	**0.83**
Lymbhography	4	0.64	**0.70**
Zoo	7	0.42	**0.81**

TABLE II. COMPARATIVE RESULTS OF K-MODES AND EXTENDED K-MODES FOR F-MEASURE

Data set	Number of Clusters	K-Modes	Extended K-Modes
Mushroom	2	0.55	**0.62**
Congressional Votes	2	0.62	**0.87**
Soybean small	4	0.31	**0.77**
Breast cancer	2	0.69	**0.79**
Lymbhography	4	0.54	**0.54**
Zoo	7	0.34	**0.71**

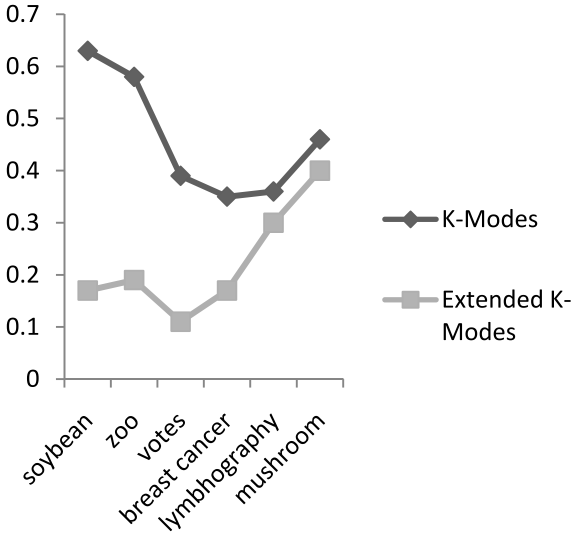

Fig.1. Error rate of data sets

REFERENCES

[1] Arun.K.Pujari , Data Mining Techniques, Universities Press, pp 114-147, 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, 582 - 589 , 2002

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

[4] George Karypis, Eui-Hong (Sam) Han, and 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, 2nd edition, pp 83,94,383-433, 2001.

[6] Ohm Mar San, Van-Nam Huynh, Yoshiteru Nakamori, An Alternative Extension Of The K-Means algorithm For Clustering Categorical Data, International Journal of Applied Mathematics & Computer Science Vol. 14, No. 2, 241–247, 2004

[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, 123-146, EDBT 2004.

[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 & Data Mining, San Diego, CA USA, 1999.

[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 Categorical 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, 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.

.