RESEARCH ARTICLE OPEN ACCESS

Numerical and Categorical Attributes Data Clustering Using K-Modes and Fuzzy K-Modes

S.Sumathi[1] and M.M.Gowthul Alam[2]
[1]PG scholar, Department of CSE, National College Of Engineering, TamilNadu.
[2]Asisstant Professor, Department of CSE, National College Of Engineering, TamilNadu.
mathirajreshmi@gmail.com
alalme2005@yahoo.com

Abstract
Most of the existing clustering approaches are applicable to purely numerical or categorical data only, but not the both. In general, it is a nontrivial task to perform clustering on mixed data composed of numerical and categorical attributes because there exists an awkward gap between the similarity metrics for categorical and numerical data. This paper therefore presents a general clustering framework based on the concept of object-cluster similarity and gives a unified similarity metric which can be simply applied to the data with categorical, numerical, and mixed attributes. This paper proposes a novel initialization method for mixed data which is implemented using K – Modes algorithm and further and iterative fuzzy K – Modes clustering algorithm.

Keyword **–**Clustering, Similarity Metrics, Plasma compatability, Initialization,k-modes ,fuzzy k-modes,exemplers.

I. INTRODUCTION

Data Mining is a technology used to describe Knowledge discovery and to search for significant relationships such as patterns, association and changes among variables in databases. Medical science industry has huge amount of data. With the growth of the blood banks, enormous Donor Blood Information Systems (DBIS) and databases are produced. It creates a need and challenge for data mining. Data mining is just a step which is used to extract interesting patterns from data that are easy to perceive, interpret, and manipulate. Several major kinds of data mining techniques, including generalization, characterization, classification, clustering, association rule mining, evolution, pattern matching, data visualization and meta-rule guided mining will be reviewed. The explosive growth of databases makes the scalability of data mining techniques increasingly important. Data mining algorithms have the ability to rapidly mine vast amount of data.

A novel algorithm called CLICKS, that finds clusters in categorical datasets based on a search for kpartite maximal cliques. Unlike previous methods, CLICKS mines subspace clusters. outperforms previous approaches by over an order of magnitude and scales better than any of the existing method for high-dimensional datasets[1]. An attractive way to perform data clustering without knowing the exact number of clusters. However, their performance is sensitive to the preselection of the rival delearning rate[2,7]. two-phase iterative procedure, which attempts to improve the overall quality of the whole partition. In the first phase, cluster assignments are given, and a new cluster is added to the partition by identifying and splitting a low-quality cluster. In the second phase, the number of clusters is fixed, and an attempt to optimize cluster assignments is done [3]

k-modes algorithm for clustering categorical data. By modifying a simple matching dissimilarity measure for categorical object, a heuristic approach was developed. Which allows the use of the k-modes paradigm to obtain a cluster with strong intra similarity, and to efficiently cluster large[4,6]

COOLCAT, which is capable of efficiently clustering large data sets of records with categorical attributes, and data streams. In contrast with other categorical clustering algorithms published in the past, COOLCAT's clustering results are very stable for different sample sizes and parameter settings. Also, the criteria for clustering is a very intuitive one, since it is deeply rooted on the well-known notion of entropy[5]. A similarity –based agglomerative algorithm that works well for data with mixed numeric and nominal feature.A similarity measure proposed by goodall for biological taxonomy that gives greater weight to uncommon feature value matches in similarity computation and makes no assumption of the underlying distributions of the feature value ,is adopted to define the similarity measure between pairs of objects. An agglomerative algorithm is employed to construct a dendogram and a simple distinctness heuristic is used to extract a partition of the data[8].

Hierarchical clustering based on the decision tree approach. As in the case of supervised decision tree, the unsupervised decision tree is interpretable in terms of rules, i.e., each leaf node represents a cluster, and the path from the root node to a leaf node represents a rule. The branching decision at each node of the tree is made based on the clustering tendency of the data available at the node. We present four different measures for selecting the most appropriate attribute to be used for splitting the data at every branching node (or decision node), and two different algorithms for splitting the data at each decision node[9]. The two step clustering method is presented to find clusters on this kind of data. In this approach the item in categorical attributes are processed to construct the similarity among them based on the ideas of co-occurrence, then all categorical attributes can be converted into numeric attributes based on these constructed relationship. Finally, since all categorical data are converted into numeric, the existing clustering algorithm can be applied to the dataset without pain.

In this paper, We will propose a unified clustering approach that is capable of selecting the cluster number automatically for both categorical and numeric data sets. Firstly, we present a general clustering framework based on the concept of object-cluster similarity. Then, a new metric for both of numerical and categorical attributes is proposed. Under this metric, the object cluster similarity for either categorical or numerical attributes has a uniform criterion. Hence, transformation and parameter adjustment between categorical and numerical values in data clustering are circumvented. Subsequently, an iterative clustering algorithm is introduced. we propose a novel initialization method for mixed data to tackle the above two issues in application of the k modes- type clustering algorithms. In cluster centers initialization, we do not use all data points as the potential exemplars but propose a method to construct a potential exemplars set. The proposed initialization method has been used along with the k-modes algorithm and the fuzzy k-modes algorithm, respectively.

The rest of the paper is organized as follows. Section 2 and Section 3 describe system model and methodologies respectively. Section 4 presents the experimental results for the proposed approaches. Finally, section 5 gives conclusion.

II. SYSTEM MODEL

The blood donor system describes about the details of the blood donor's they should register their name by donor registration form , before they are going to donate the blood. Now the donor play a valuable and essential role in health care system. The goal of the blood donation is to provide sufficient amount of blood to the patient who need blood transfusions in all hospitals at all times. The system architecture-I shows that the blood donation system used to accessible data bank to any authorized centers will reduce the amount of time spent by the blood centers looking for more donors. To sum up blood donation services are working hard to recover from a period of social and economical uncertainty. This system might improve the performance of blood donation and encourage donors to arrange a high frequency of donations therefore maximizing blood collected.

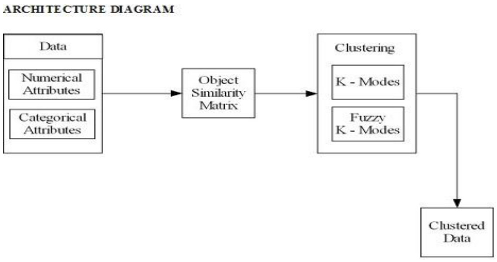

Figure1. Architecture Design

In this paper the donor details are given to the system it contains categorical and numerical data. Our important data is the blood group type of the donor. The similarity metrics are used to find and differentiates the using the cluster probability. the probability is zero and one. The Iterative approach used to fine the similarity between the cluster and object. The competitive mechanism delete the number of redundant clusters. Some objects are belong to two cluster with similar characteristics. The Blood types are AB,O,A,B under their cluster are seed points of the clustering in the iterative approach. The donor data are splitted into number of four type of blood groups using the plasma compatabillity. The Blood has fifty percentage of plasma in it. Plasma contains water (90%), proteins (albumin, fibrinogen and globulins), nutrients (glucose, fatty acids, amino acids), waste products (urea, uric acid, lactic acid, creatinine), clotting factors, minerals, immunoglobulins, hormones and carbon dioxide, i.e. all the components of blood except the red, white blood cells and thrombocytes. Components can either be dissolved (if soluble) or remain bound to proteins (if insoluble). Plasma has the density of 1025 kg/m^3. The function of plasma is maintaining the electrolytes and fluid balance of the blood. Serves as the protein reserve for the body. Aids in clotting. Immune functions. Transport of carbon dioxide, essential nutrients (organic, inorganic components and plasma proteins), hormones (bound to plasma proteins), waste (urea, uric acid and creatinine) and other substances (example drugs and alcohol) to and from the tissues.

III. METHODOLOGIES

In this section, we discuss Clustering problem and object-cluster similarity metrics, Iterative approach and automatic selection of cluster number.

3.1 object-cluster similarity metrics

The general task of clustering is to classify the given objects into several clusters such that the similarities between objects in the same group are high while the similarities between objects in Different groups are low. Therefore, clustering a setofNojects.

$$\arg_{Q}^{\max} F(Q) =$$

$$\arg_{Q}^{\max}\left[\sum_{j=1}^{k}\sum_{i=1}^{n} q_{ij}\, s(X_i, C_j)\right]$$

The distance between each vector x_{ui} can be numerically calculated, the similarity metric numerical attributes can be defined based on the measure of distance.

$$s(X_i^u, C_j) = \frac{\exp(-0.5Dis(X_i^u, C_j))}{\sum_{t=1}^{k}\exp(-0.5Dis(X_i^u, C_t))} \qquad)$$

3.2 Iterative approach

An iterative algorithm based on proposed object-cluster similarity metric to conduct clustering analysis. This paper concentrate on hard partition only.ie $q_{ij} \epsilon \{0,1\}$.Although it can be easily extended to the soft partition interms of posterior probability. Theset of N objects,the optimal, $q^*_{ij}=$

$$\begin{cases} 1 \text{ if } s(Xi, Cj) \leq 1 \leq k, \\ 0 \qquad \text{otherwise,} \end{cases}$$

TABLE 1 ITRERATIVE ALGORITHM

```
Input: data set $X=\{x_1,x_2..x_N)$ number of cluster k
Output: cluster label $Y=\{y_1,y_2..y_N\}$
1. Calculate the importance of the each categorical attributes
$H_{A_r} = -\sum_{t=1}^{m_r} P(a_{rt}) log\, P(a_{rt})$
2. Set Y={0,0,0….0} and select k initial objects ,one for each cluster
Repeat initialize
noChange=true
for i=1 to N do
$Y_i^{(new)} = \text{argmax}_{J\epsilon\{1....K\}}[S(x_i,C_j)]$
if $Y_i^{(new)} \neq Y_i^{(old)}$ then
noChange=false
3. Update the information of cluster $C_{yi}^{(new)}$ and $C_{yi}^{(old)}$
End if
End for
Until noChange is true
```

The object-cluster similarity $s(x_i\ ,C_j)$ is calculated. for mixed, categorical, or numerical data, respectively. Additionally, in order to update the cluster information conveniently in step 4, two auxiliary matrices for each cluster are maintained. One matrix is to record the frequency of each categorical value occurring in this cluster, and the other matrix stores the mean vector of the numerical parts of all objects belonging to this cluster.The random initialization method with multiple repetition to get the statistic information for clustering performance evaluation. Algorithm is applied to purely numerical data and Euclidean distance is utilized to calculate Dis(xi ,cj), according to the similarity metric defined by we can get

$$s(X_i, C_j) \geq s(X_i, C_r) \Leftrightarrow \frac{\exp(-0.5\|X_i^u, C_j\|^2)}{\sum_{t=1}^{k}\exp(-0.5\|X_i^u, C_t\|^2)}$$

Where ‘‘3’’ means ‘‘equivalent to’’. Then, the clustering criterion formulated by can be simplified as

$$\mathbf{q^*_{ij}} = \begin{cases} 1 \text{ if } \|(Xi, Cj)\|2 \leq \||(Xi, Cr)\|2 \;\; 1 \leq r \leq k, \\ 0 \qquad \text{otherwise,} \end{cases}$$

That is, each object will be assigned to the cluster whose centroid is closest to it. The time complexity analysis of OCIL algorithm. the computation cost of step 1 is $O(mNd_c)$. For each iteration, the cost of the ‘‘for’’ statement is $o(mNkdc+Nkd_u)$, where m is the average number of different values that can be chosen by each categorical attribute. Therefore, the total time cost of this algorithm is $O(t(mNkd_{c+}Nkd_u))$, where t is the number of iterations, we often have k«N, m«N and t«N. subsequently, the time complexity of this algorithm is O(dN). Hence, the proposed algorithm is efficient for data clustering, particularly for a large data set.

3.3 Initial clustering k--modes

We propose a novel initialization method for categorical data to tackle the above two issues in application of the kmodes- type clustering algorithms. In cluster centers initialization,we do not use all data points as the potential exemplars but propose a method to construct a potential exemplars set. We define a new density measure to reflect the cohesiveness of potential exemplars. Based on the density measure and the distance measure, we select the initial cluster centers from the potential exemplars set. In determination of the number of clusters, we propose a criterion to find the candidates for the number of clusters. The proposed initialization method has been used along with the k-modes algorithm and the fuzzy k-modes algorithm, respectively. The time complexity of the proposed method has been analyzed.

Categorical data is different from continuous or discretized numerical data. Due to the lack of an inherent order on the domains of the categorical attributes, the clustering techniques for numerical data cannot be applicable to categorical data. A new initialization method is proposed to simultaneously find the initial cluster centers and the number of clusters for categorical objects. In this paper, we took account of clustering not attributes but objects. We applied three widely used external evaluation measures to evaluate the effectiveness of the k-modes type algorithms with the proposed initialization method on several real data setsfrom UCI.

TABLE 1 Initialization method

Input: IS = (U,A,V, f) and k, where k is the number of clusters desired.
Output: Centers.
Step 1: Centers =Ø;
Step 2: Construct the potential exemplars set S;
Step 3: For each c ε S, calculate the Dens(c) and choose the densest potential exemplar c as the first cluster center. Set Centers = Centers U{c} and i = 1;
Step 4: If i < k, then let i = i + 1 and choose the most probable exemplar c from S as the i + 1 cluster center, Centers = CentersU{c} where c satisfies, $\text{Possibility}_i(c)=\max_{c'\varepsilon S} \text{Possibility}_i(c')$;and goto Step 4, otherwise goto Step 5;
Step 5:End

TABLE 2 k-modes algorithm

Step 1: Choose an initial point set Z(1) # R^m, where $R^m = V_{a1}x$
V_{a2} x..V_{am} . Determine $W^{(1)}$ such that $F(W,Z^{(1)})$ is minimized.
Set t = 1.
Step 2: Determine $Z^{(t+1)}$ such that $F(W^{(t),}Z(t^{+1}))$ is minimized. If
$F(W^{(t),}Z^{(t+1)}) = F(W^{(t),}Z^{(t)})$, then stop; otherwise goto Step 3.
Step 3: Determine $W^{(t+1)}$ such that F(W(t+1),Z(t+1)) is minimized. If $F(W^{(t+1),}Z^{(t+1)}) = F(W^{(t),}Z^{(t+1)})$, then stop; otherwise set t = t + 1 and goto Step 2.

This procedure removes the numeric-only limitation of the k means- type algorithm. To cluster categorical data, the k-modes-type algorithms apply the simple matching dissimilarity measure to compute the distance between a cluster center and a categorical data point, use modes as cluster centers instead of means for clusters, and update
the cluster centers at each iteration to minimize the Clustering cost function.The simple matching dissimilarity measure **$d(z_l,x_i)$** between a center **z_l** and a categorical data point x_i is defined as

$$d(z_l,x_i)=\sum_{j=1}^{m}\delta_{a_j}(z_l,x_i)$$

D defines a metric space on the set of categorical data points.d is also a kind of hamming distance.

3.4.Iterative clustering fuzzy k-modes

The fuzzy k-modes procedure removes the numeric-only limitation of the kmeans- type algorithm. Moreover, the fuzzy partition matrix provides more information to help the user to determine the final clustering and to identify the boundary data points. Such information is extremely useful in applications such as data mining in which the uncertain boundary data points are sometimes more interesting than data points which can be clustered with certainty. However, similar to the k-means-type algorithms, the k-modes type algorithms are sensitive to initial cluster centers and need to give the number of clusters in advance. The fuzzy k-modes algorithm are presents the over all blood donors with each category in the succeeding years.

IV. PERFORMANCE EVALUATIONS

To evaluate the performance of clustering algorithms, three evaluation measures are introduced in . If data set contains k classes for a given clustering, let a_i denote the number of data points that are correctly assigned to class C_i, let bi denote the number of data points that are incorrectly assigned to the class C_i, and let ci denote the number of data points that are incorrectly rejected from the class Ci. The accuracy, precision and recall are defined as follows

$$AC=\frac{\sum_{i=1}^{k} ai}{N},$$

$$PR=\frac{\sum_{i=1}^{k}\left(\frac{ai}{ai-bi}\right)}{k}$$

The performance analysis of the proposed method consists of two parts. The one is to evaluate the effectiveness of the initial cluster centers obtained by the proposed method. There are two types of clustering validation techniques which are based on external and internal criteria, respectively. The focus of this paper is on the evaluation of external clustering validation measures. In this part, we first introduce three commonly used external evaluation measures which are used to compare a clustering result with the true class distribution on a given data set. As external criteria, these measures use external information-class labels and the number of clusters. If the cluster result is close to the true class distribution, then the values of these evaluation measures are high. To ensure that the comparisons are in a uniform environmental condition, we set that the number of clusters is equal to the true number of clusters for each of the given data sets

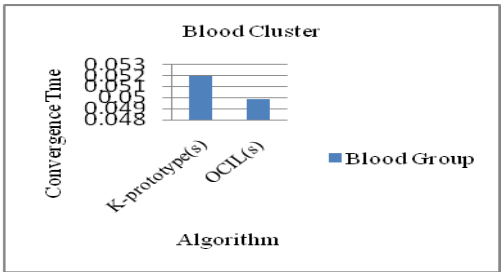

Figure 2 Convergence time between algorithms

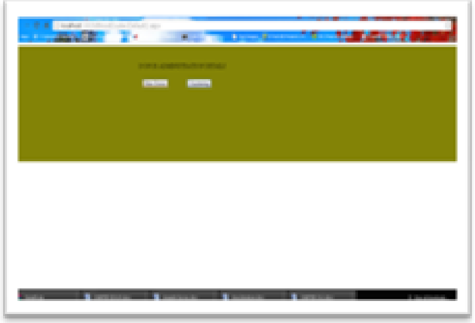

Fig 2 Administrative form

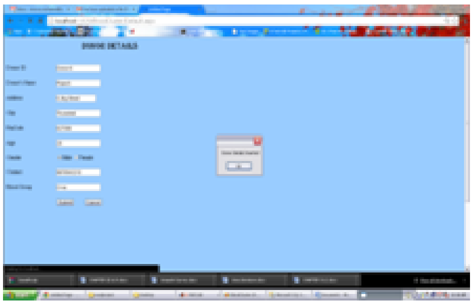

Fig 4 Donor form

Donors with Plasma Compatability 'A'

Donors with Plasma Compatability 'B'

Donors with Plasma Compatability 'O'

Donors with Plasma Compatability 'AB'

Fig 5 Classes of clusters

Donors with Plasma Compatability 'A'

Donors with Plasma Compatability 'B'

Donors with Plasma Compatability 'O'

Donors with Plasma Compatability 'AB'

Fig 6 Compatability

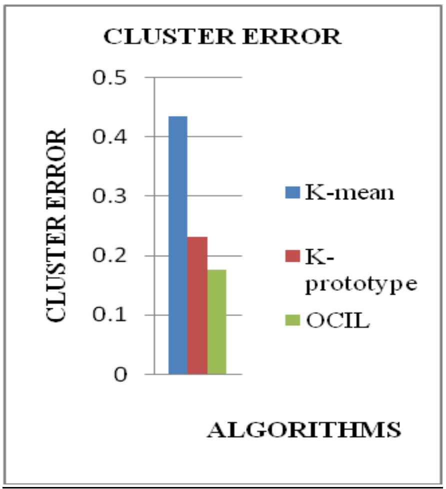

Fig 6 Cluster Error on multiple algorithms

V. CONCLUSION

Clustering framework based on object-cluster similarity, through which a unified similarity metric for both categorical and numerical attributes has been presented. Under this new metric, the object-cluster similarities for categorical and numerical attributes are with the same scale, which is beneficial to clustering analysis on various data types. Subsequently, an iterative algorithm has been introduced to implement the data clustering. The advantages of the proposed method have been experimentally demonstrated in comparison with the existing counterparts. The development of the k-modes-type algorithms was motivated to solve this problem. However, the performance of these algorithms strongly depends on two parameters, an initial set of cluster centers and the number of clusters. When the prior information about setting the two parameters for a data set is not available, it is difficult for users to implement these algorithms to effectively cluster the data set. In this paper, a new initialization method for categorical data clustering has been proposed. The proposed method can simultaneously obtain the good initial cluster centers and the candidates for the number of clusters. Furthermore, the time complexity of the proposed method has been analyzed.

References

[1] A.P. Dempster, N.M. Laird, D.B. Rubin,” Maximum likelihood from incomplete data via the EM algorithm”, Journal of the Royal Statistical Society, Series B (Methodological)pp1–38,1977..

[2] A.W.-C. Liew, H. Yan, M. Yang, “Pattern recognition techniques for the emerging field of bioinformatics”: a review, Pattern Recognition. Pp2055–2073, 2005.

[3] Cen Li and Gautam biswas, “Unsupervised Learning with mixed numeric and nominal data”, IEEE Transactions on Knowledge and Data Engineering 2002.

[4] C.C. Hsu, “Generalizing self-organizing map for categorical data”, IEEE Transactions on Neural Networks. pp 294–304,2006

[5] D.Barbara,j,Couto,Y,Li,“COOLCAT: An entropy-based algorithm for categorical clustering”, IEEE Transactions on Knowledge and Data Engineering pp 582-589,2002.

[6] E.Cesario,G.Manco,R.Ortale,“Top-Down Parameter-Free Clustering of High-Dimensional Categorical Data”, IEEE Transactions on Knowledge and Data Engineering pp 1607-1624,2007.

[7] J.Bask,R. Krishnapuram,“Interpretable Hierarchical Clustering by Constructing an Unsupervised Decision Tree”, IEEE Transactions on Knowledge and Data Engineering pp 121-132,2005.

[8] J.B. MacQueen, “Some methods for classification and analysis of multivariate observations”, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, pp 281–297,1967.

[9] L. Xu, A. Krzyzak, E. Oja, “On Rival Penalization Controlled Competitive Learning for Clustering with Automatic Cluster Number Selection”, IEEE Transactions on Knowledge and Data Engineering pp 636-648,1993.

[10] Ming-Yi Sinh,Jar-Wen Jheng and Lien- Fu Lai, “A two step method for Clustering mixed categorical and numeric data”, IEEE Transactions on Knowledge and Data Engineering pp 11-19,2010.

[11] M.J. Zaki, M. Peter,”CLICKS: Mining Subspace Clusters in Categorical Data”, IEEE Transactions on Knowledge and Data Engineering pp. 355–356,2005.

[12] M.K.Ng,M.Li.J.Z.Huang,”On the Impact of dissimilarity measure in K-modes Clustering Algorithm”,pp 503-507,2007.

[13] L. Xu, A. Krzyzak, E. Oja, “Rival penalized competitive learning for clustering analysis”, RBF net, and curve detection, IEEE Transactions on Neural Networks 4 (4) (1993) 636–648.

[14] P. Andritsos, P. Tsaparas, R.J. Miller, K.C. Sevcik, LIMBO: “scalable clustering of categorical data”, in: Proceedings of the 9th International Conference on Extending Database Technology, pp. 123–146,2004.

[15] R.S. Michalski, I. Bratko, M. Kubat, Machine Learning and Data Mining: Methods and applications, Wiley, New York, 1998.

[16] S. Guha, R. Rastogi, K. Shim, “ROCK: a robust clustering algorithm for categorical attributes”, Information Systems 25 (5)pp 345–366,2001..

[17] W. Cai, S. Chen, D. Zhang, “Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation”, Pattern Recognition pp 825–838,2007.

[18] Yiu-Ming Cheng, “On Revial penalization controlled competitive Learning for clustering with Automatic Cluster Number Selection” IEEE Transactions on Knowledge and Data Engineering,2005.

[19] Y.M. Cheung, “Maximum weighted likelihood via rival penalized EM for density mixture clustering with automatic model selection”, IEEE Transactions on Knowledge and Data Engineering .pp 750–761,2005.

[20] Z. Huang, “Clustering large data sets with mixed numeric and categorical values”, in: Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining, , pp 21–34,1997.

[21] Z. He, X. Xu, S. Deng, “Scalable algorithms for clustering large datasets with mixed type, International Journal of Intelligence Systems 20 (2005) 1077–1089.