A Data Labeling Algorithm Used For Clustering Mixed Data Set To Drifting Categorical Data

AMARNATH ANNADASU[#1], K.Ramesh[#2],A.Siddhartha Reddy[#3]

[1,2#] (Kakinada Institute of Engineering and Technology, JNTU Kakinada, A.P, India)

ABSTRACT:

Clustering is become an major issue in data mining applications. There are several clustering algorithms are available to cluster datasets that contain either numeric or categorical attributes. The actual life database consists of numeric, categorical and mixed type of attributes. Now, clustering data with categorical attributes, whose attribute values do not have a natural ordering, has received some notice that, it is an crucial task to cluster these data sets to extract major knowledge from the current database or to obtain statistical information about the database. Clustering large database is a time taken process. Sampling is a process of gaining a small set of data from the large database. Applying sampling procedure would not cluster all the data points. Labeling non- clustered data point is an issue in data mining process. This paper mainly focuses on clustering mixed data set using modified MARDL (MAximal Resemblance Data Labeling) method and to allocate each unlabeled data point into the corresponding appropriate cluster based on the novel clustering representative namely, N-Nodeset Importance Representative (NNIR). Accuracy and Error rate are considered as the metrics for evaluating the performance of the existing and proposed algorithm for mixed data set. The experimental result shows that MARDL for mixed data set algorithm performs better than the existing enhanced algorithms.

Keywords*:* Data mining, Data labeling, Categorical Clustering, Sampling

Corresponding Author: K.Ramesh (Assistant Professor, M.tech CSE Department)

I. INTRODUCTION:

The clustering problem has been deemed an important issue in the data mining, statistical pattern recognition, machine learning, and information retrieval because of its use in a wide range of applications [1]. Given a set of data points, the goal of clustering is to partition those data points into several groups of similar points according to the predefined similarity measurement [2]. However, finding the optimal clustering result has been proved to be an NP-hard problem [3]. As the size of data grows at rapid pace, clustering a very large database inevitably involves a very time consuming process.

To improve the efficiency, sampling is usually used to scale down the size of the database In particular, sampling has been employed to speed up clustering algorithms in [4]. A typical way to utilize sampling techniques on clustering is to randomly choose a small set from the original database, and then the clustering algorithm is executed on the small sampled set. The clustering result which is expected to be similar to that obtained from the original database can hence be efficiently obtained.

However, the problem of how to allocate the unclustered data into appropriate clusters has not been fully explored in the previous works. This can be explained by the reason that in the numerical domain, there is a common solution to measure the similarity between an unclustered data point and a cluster based on the distance between the unclustered data point and the centroid of that cluster [1]. Each unclustered data point can be allocated to the cluster with the minimal distance. Previous works usually deal with such a issue by this straightforward method. However, much of the data in the existing database is categorical. In the categorical domain, the above procedure is infeasible because the centroid of cluster is difficult to define. Without loss of generality, the goal of clustering is to allocate every data point into an appropriate cluster. A partial clustering result obtained from the sampled database is usually not what the user really wants. Therefore, in the categorical domain, the problem of how to allocate the unclustered data remains as a challenging issue..

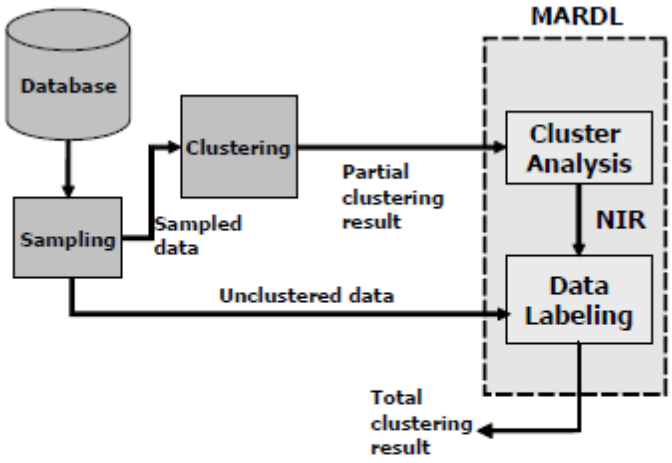

Figure 1. The framework of clustering a categorical very large database with sampling and MARDL.

As a result, we propose in this paper a mechanism, named **MA**ximal **R**esemblance **D**ata **L**abeling (abbreviated as *MARDL*), to allocate each categorical unclustered data point into the corresponding proper cluster. The allocating process is referred to as *Data Labeling: to give each* unclustered data point a cluster label. The unclustered data points are also called unlabeled data points. Figure 1 shows the entire framework on clustering a very large database based on sampling and MARDL. In particular, MARDL is independent of clustering algorithms, and any categorical clustering algorithm can in fact be utilized in this framework.

II. Related Work:

Data labeling is used to allocate an unlabeled data point into the corresponding appropriate cluster. The technique of data labeling is available in CURE [5]. However, CURE is a special numerical clustering algorithm to find non-spherical clusters. A specific data labeling algorithm is defined to assign each unlabeled data point into the cluster which contains the representative point closest to the unlabeled data point. CURE is robust to outliers and identifies clusters with non-spherical shapes, and wide variances in size. Each cluster is represented by a fixed number of well scattered points.

To allocate each categorical unclustered data point into the corresponding proper cluster MARDL [6], mechanism is proposed. It is a framework of clustering large categorical database with sampling and data labeling techniques. MARDL is independent of clustering algorithms, and any categorical clustering algorithm can be utilized in this framework.

Clusters are represented by several representative points. ROCK [7], is an adaptation of an agglomerative hierarchical categorical clustering algorithm. This algorithm assigns data point to a separated cluster, and then merges the clusters repeatedly according to the closeness between clusters. The closeness between clusters is defined as the sum of the number of "links" between all pairs of representative points in the clusters. However, this representative utilizes several representative points and moreover it does not provide a summary of cluster, and thus cannot be efficiently used for the post processing.

For example, in the data labeling, the similarity between unclustered data points and clusters is needed to be measured. It is time consuming to measure the similarity between unclustered data points and each representative point, especially when a large amount of representative points is needed for the better representability.

Squeezer algorithm [8], produces high-quality cluster in high-dimensional categorical datasets. This algorithm has been extended for the domains with mixed numeric and categorical attributes algorithm namely dsqueezer and usm-squeezer. Since the Squeezer algorithm has been demonstrated to be very effective for clustering categorical datasets, in the squeezer algorithm, we adopt a simple strategy of transforming the original dataset into categorical dataset by discretizing numeric attributes. Then, the Squeezer algorithm is used to cluster the transformed dataset. For the usm-squeezer algorithm, a unified similarity measure for mixed-type attributes, in which both numeric and categorical attributes could be handled equally in the framework of Squeezer algorithm.

III.PROJECTED FRAMEWORK FOR CLUSTERING USING MODIFIED MARDL

In MARDL, those unlabeled data points will be allocated into clusters via two phases, namely, the Cluster Analysis phase and the Data Labeling phase. The work doing in each phase is described below and shown in above fig 1.

Cluster Analysis phase:

In the cluster analysis phase, a cluster representative is generated to characterize the clustering result. In this paper, a cluster representative, named NIR is devised. NIR represents clusters by the attribute values, and the importance of an attribute value is measured by the following two concepts: 1) the attribute value is important in the cluster when the frequency of the attribute value is high in this cluster and 2) the attribute value is important in the cluster if the attribute value appears prevalently in this cluster rather than in other clusters. To measure the importance of attribute values, NIR considers both the intracluster similarity and the intercluster similarity to represent the cluster. Moreover, we extend NIR to represent clusters by multivariate attribute values

Data Labelling Phase:

In the data labeling phase, each unlabeled data point is given a label of appropriate cluster according to NIR/NNIR. The similarity between the unlabeled data point and the cluster is designed based on the NIR/NNIR. Based on this similarity measurement, MARDL allocates each unlabeled data point into the cluster which possesses the maximal resemblance. Therefore, NNIR is able to be utilized in the clustering visualization, and tries to represent the clustering result in an effective way.

Explanation (*Node*)

A *node*, dt, is defined as *attribute name + attribute value* .The term *node* which is defined to represent attribute value in this paper avoids the ambiguity which might be caused by identical attribute values. If there are two different attributes with the same attribute value, e.g., the age is in the range *50~59* and the weight is in the range *50~59*, the attribute value *50~59* is confusing when we separate the attribute value from the attribute name. *Nodes [age=50~ 59]* and [w*eight=50~59]* avoid this ambiguity. Note that if the attribute name and the attribute value are both the same in the nodes d1 and d2, d1 and d2 are said to be equal.

Node Importance Representative:

NIR is used to represent a cluster as the distribution of the attribute values. A node Ir , is defined as attribute name plus attribute value. NIR considers both the intracluster and intercluster similarity. The importance of the node in a cluster is measured making use of the two concepts that figure below:

(i) The node is important in the cluster when the frequency of the node is high in this cluster.

(ii) The node is important in the cluster if the node appears predominantly in this cluster rather than in other clusters.

The idea of NNIR is to represent a cluster as the distribution of the n node sets, which are already defined in this section. NNIR is an extension of NIR where each attribute value combinations are considered to characterize the clustering results.

Based on the above two concepts, we define the n-node set: I_{ir}^n in equation (1)

$$w(c_i, I_{ir}^n) = \frac{|I_{ir}^n|}{m_i} * f(I_r^n) \quad (1)$$

$$f(I_r^n) = 1 - \frac{-1}{\log k} * \sum_{y=1}^{k} p\left(I_{yr}^n\right) \log\left(p\left(I_{yr}^n\right)\right)$$

Where

$$p(I_{yr}^n) = \frac{|I_{yr}^n|}{\sum_{z=1}^{k^t} |I_{zr}^n|}$$

Where mi is the number of data points in cluster Ci. , $|I_{ir}^n|$ is the frequency of the node set I_{ir}^n , and k is number of clusters, since this is a product of two factors. The first factor is the probability of I_{ir} being in Ci using rule I_{ir}

(i) which aims to maximize the intra cluster similarity and the second factor is the weighting function arrived at using rule

(ii) which minimizes the inter cluster similarity

IV. MARDL Algorithm

The goal of MARDL is to decide the most appropriate cluster label c for the unlabeled data point. Specifically, when an unlabeled data point is given MARDL computes the similarity between and cluster and finds the cluster which has Max In order to calculate the similarity between and referred to as resemblance

The algorithm MARDL is outlined below, where MARDL can be divided into two phases, the cluster analysis phase and the data labeling phase.

MARDL(C, U):

Clustering result C, unclustered data set U.

Procedure

main ():

The main procedure of MARDL

1. N Table=cluster analysis(C);
2. Data Labeling (N Table, U);

Procedure Cluster Analysis(C): analyze input clustering result and return the NIR hash table

Luster analysis(C)

1. While (C[next]! ='\0') {
2. p [i][j]=C[next];
3. divide Nodes (p [i][j]);
4. Update NF(c[i]);
5. }
6. For ($N=d_{i1}$; $N<=d_{it}$; N++)
7. Compute Weight f (d_{ix});
8. For ($C = c_1$ & $C<=c_n$; C++) {
9. For ($N=d_{i1}$&$N<=d_{it}$; N++) {
10. Calculate n (w_i, d_{ix})
11. Add NIR table NTable (d_{ix}, w_i, d_{ix})}}
12. Return NTable;

Procedure Data Labeling (NTable,U): give each unclustered data point a cluster label

13. While (U[next]! ='\0') {
14. U [u}[j]=U[next};
15. Divide nodes (p[u][j]);
16. For ($N=c_1$; $N \leq c_n$; $N++$)
17. Calculate Resemblance(C[m])
18. Give label c[m] to p[u][j] ;}.

The main purpose of the cluster analysis phase is to represent the prior clustering result with NIR. NIR represents cluster by a table which contains all the pairs of a node and its node importance. For better execution efficiency, the technique of hash can be applied on the represented table. Since the node names are never repeated, node is suitable to be a hash key for efficient execution. The main purpose of the data labeling phase is to decide the most appropriate cluster label for each unlabeled data point. Each unlabeled data point is labeled and then classified to the cluster which attains the maximal resemblance. The resemblance value of the specific cluster is computed efficiently by the sum of each node importance through looking up the NIR hash table q times. After all the resemblance values is computed and recorded, the maximal resemblance value is found, and the unlabeled data point is labeled to the cluster which obtains the maximal resemblance value. Note that after executing the data labeling phase, the

labeled data point just obtains a cluster label but is not really added to the cluster. Therefore, NIR table will not be modified in the data labeling phase. This is because the MARDL framework does not cluster data, but rather, presents the original clustering characteristics to the incoming unlabeled data points.

V. Conclusion:

This paper stylized the definition of a cluster when the data consists of categorical features, and then initiated a fast summarization-based algorithm MARDL. To allocate each unlabeled data point into the appropriate cluster when the sampling technique is utilized to cluster a very large categorical database categorical cluster representative technique, named NIR, to represent clusters which are obtained from the sampled data set by the distribution of the nodes. This MARDL method works based on NIR. The evaluation validates our claim that MARDL is of linear time complexity with respect to the data size, and MARDL preserves clustering characteristics, high intra-cluster similarity and low inter-cluster similarity. It is shown that MARDL is significantly more efficient than prior works while attaining results of high quality.

VI. References:

1) A. K. Jain, M. N. Murthy, and P. J. Flynn. Data clustering: a review. ACM Computing Surveys, 1999.

2) P. Berkhin. Survey of clustering data mining techniques.Technical report, Accrue Software, 2002.

3) D. S. Johnson M. R. Garey and H. S. Witsenhausen. The complexity of the generalized lloyd-max problem.*IEEE Trans. Inf. Theory*, 1982

4) R. T. Ng and J. Han. CLARANS: A Method for Clustering Objects for Spatial Data Mining. IEEE Transactions on Knowledge and Data Engineering, 2002

5) S. Guha, R. Rastogi, and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases", Proc. ACM SIGMOD IEEE Trans. Knowledge and Data Eng., 1998.

6) Hung-Leng Chen, Kun-Ta Chuang, Member, IEEE, and Ming- Syan Chen, Fellow, IEEE, "On Data Labeling for Clustering Categorical Data" IEEE Trans. Knowledge and Data Eng.,2008

7) S. Guha, R. Rastogi, and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes," Proc. 15th Int'l Conf. Data Eng. (ICDE), 1999

8) Zengyou He, Xiaofei Xu, Scengchun Deng, "Scalable Algorithms for Clustering Large Datasets with Mixed Type Atributes," International journal of intelligent systems, vol. 20, 1077–1089 (2005).