Volume 2, No. 4, July-August 2011
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

Clustering Categorical Data – Study of Mining Tools for Data Labeling

M. Sathya Narayana
Asst Prof.
VCET, Hyderabad,India
sathyam542@gmail.com

B. V.V.S Prasad*
Asst Prof.
VCET, Hyderabad,India
prasad_bvvs2004@yahoo.co.in

N. Siva Ram Babu
Software Analysts.
TCS, Hyderabad
nsivarambabu@gmail.com

B.Suresh Kumar
Asst Prof.
Jayoti Vidyapeeth Womens University
sureshkumar5656@gmail.com

Abstract: Cluster analysis sampling has been recognized as a best technique to improve the efficiency of clustering. However, with sampling applied to those points which are not sampled will not have their labels after the normal process. Although there is a straightforward approach in the numerical domain, the problem of how to allocate those unlabeled data points into proper clusters remains as a challenging issue in the categorical domain. In this paper, a mechanism named Maximal Resemblance Data Labeling (abbreviated as MARDL) is proposed to allocate each unlabeled data point into the corresponding appropriate cluster based on the novel categorical clustering technique, importance of the combinations of attribute values. MARDL has two advantages: 1) MARDL exhibits high execution efficiency and 2) MARDL can achieve high intra cluster similarity and low inter cluster similarity, which are regarded as the most important properties of clusters, thus benefiting the analysis of cluster behaviors. This article analysis the implementing the proposed system using data mining tools, the algorithm which shows the effective from Rock.

Keywords: Clustering, Rock, sampling, MARDL, Data mining tools

I. INTRODUCTION

Data mining is the process of extracting patterns from data. As more data are gathered, with the amount of data doubling every three years,[1] data mining is becoming an increasingly important tool to transform these data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery. Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details (akin to lossy data compression), but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, clustering is unsupervised learning of a hidden data concept. Data mining deals with large databases that impose on clustering analysis additional severe computational requirements. Given a set of data points, the goal of clustering is to partition the data points into different groups according to the predefined similarity measurement [3]. However, finding the optimal clustering result has been proved to be an NP-hard problem [5]. As the size of data grows at rapid pace, clustering a very large database inevitably incurs a time-consuming process. In the numerical domain, there is a common solution to measure the similarity between an un clustered data point and a cluster based on the distance between the un clustered data point and the centroid of that cluster [4]. Each un clustered data point can be allocated to the cluster with the minimal distance. In the categorical domain, the above procedure is infeasible because the centroid of cluster is difficult to define. In algorithm ROCK [19], a similar sampling strategy has been applied to speed up the entire clustering procedure, and the problem of allocating the un clustered data has been discussed. Although ROCK provides high quality on the problem, the allocating procedure is time consuming. The result can be explained by the reason that ROCK utilizes the original sampled data, not the summary of the sampled clustering result, to perform the allocating procedure. As a result, for the categorical domain, the problem of how to efficiently allocate the un clustered data into corresponding proper clusters remains as a challenging issue.

As a result, in this paper a mechanism, named MAximal Resemblance Data Labeling (abbreviated as MARDL), to allocate each categorical unclustered data

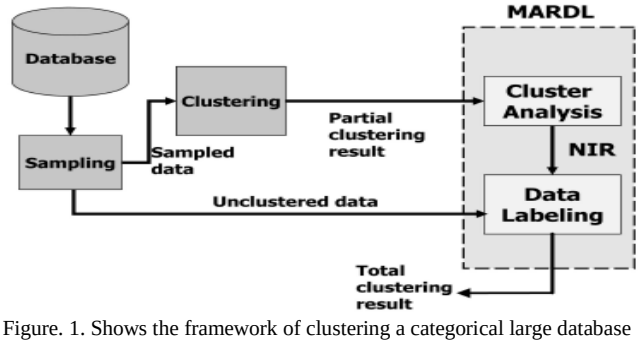

Figure. 1. Shows the framework of clustering a categorical large database

point into the corresponding proper cluster. The allocating process is referred to as Data Labeling: to give each unclustered data point a cluster label. For simplicity, we call the un clustered data points as unlabeled points in the sequel. Fig. 1 shows the entire framework on clustering a large database based on sampling and MARDL. In particular, MARDL is independent of clustering algorithms, and any categorical clustering algorithm can be utilized in this framework. In MARDL, those unlabeled data points will be allocated into clusters via two phases, namely, the Cluster Analysis phase and the Data Labeling phase.

II. SECTION

A. Clustering:

Cluster is a number of similar objects grouped together. It can also be defined as the organization of dataset into homogeneous and/or well separated groups with respect to distance or equivalently similarity measure. Cluster is an aggregation of points in test space such that the distance between any two points in cluster is less than the distance between any two points in the cluster and any point not in it. There are two types of attributes associated with clustering, numerical and categorical attributes. Numerical attributes are associated with ordered values such as height of a person and speed of a train. Categorical attributes are those with unordered values such as kind of a drink and brand of car.
Clustering is available in flavors of
Hierarchical
Partition
Grid-Based
Density-Based

a. Hierarchical Clustering:

In hierarchical clustering the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to n clusters each containing a single object. Hierarchical Clustering is subdivided into agglomerative methods, which proceed by series of fusions of the n objects into groups, and divisive methods, which separate n objects successively into finer groupings.

b. Partition Clustering:

Partition clustering technique splits the data points into k partition, where each partition represents a cluster. The partition is based on certain objective function. One such criterion functions is minimizing square error criterion which is computed as,

$$E = \Sigma \Sigma || p - m_i ||^2$$

where p is the point in a cluster and m_i is the mean of the cluster. The cluster has two properties, they are (1) each group must contain at least one object (2) each object must belong to exactly one group. The main draw back of this algorithm [7] is whenever a point is close to the center of another cluster, it gives poor result due to overlapping of data points.

c. Grid based Clustering:

Grid based algorithm quantize the object space into a finite number of cells that forms a grid structure [1].Operations are done on these grids. The advantage of this method is lower processing time. Clustering complexity is based on the number of populated grid cells and does not depend on the number of objects in the dataset. The major features of this algorithm are:

i. No distance computations.
ii. Clustering is performed on summarized data points.
iii. Shapes are limited to union of grid-cells.
iv. The complexity of the algorithm is usually O(Number of populated grid-cells)
STING [6] is an example for this algorithm.

d. Density based Clustering:

Density based algorithm continues to grow the given cluster as long as the density in the neighborhood exceeds certain threshold [6]. This algorithm is suitable for handling noise in the dataset. The following points are enumerated as the features of this algorithm.
1. Handles clusters of arbitrary shape
2. Handle noise
3. Needs only one scan of the input dataset.
4. Needs density parameters to be initialized.
DBSCAN, DENCLUE and OPTICS [6] are examples for this algorithm.

B. Categorical Data

Categorical data variables are characterized by values, which are classified into: Dichotomous, Multi-categorical. Dichotomous variables are often coded by the values zero and one. For similarity measuring it is necessary to take into account whether the variables are symmetric or asymmetric. In the first case, both categories have the same importance (male, female). In the second case, one category is more important (presence of the word in a textual document is more important than its absence). Multi-categorical variables can be classified into three types: nominal, ordinal and quantitative. Unlike the other types, categories of nominal variables cannot be ordered (from the point of view of intensity etc.). Categories of ordinal variables can be ordered but we cannot usually do the arithmetic operations with them (it depends on the relations among categories,) we can do arithmetic operations with quantitative variables (number of children). To denote nominal, ordinal and dichotomous variables as categorical. These variables are also called qualitative. Suppose that the dichotomous variables are binary with categories zero and one. The same similarity measures are used for clustering of both objects and variables in this case is binary data.

If binary variables are symmetric, apply the same measures as for quantitative data. Moreover, many specific coefficients have been proposed for this kind of data, as well as for data files with asymmetric binary variables. If there are no special means for clustering multi-categorical data in a software package, then transformation of the data file to a file with binary data is usually needed. The difference between nominal and ordinal types is necessary.

First, mention the data file with nominal variables. In comparison with classification tasks involving a target variable (regression and discriminant analyses, decision trees), the number of dummy variables must be equal to the number of categories, in the Table 1. In this way it is guaranteed that one can obtain only two possible values of similarity: one for the matched categories, and the second for unmatched categories.

Table 1 Recoding of the nominal variable School for three binary variables P1 to P3

School	P1	P2	P3
Grammar school	1	0	0
Business school	0	1	0
Technical school	0	0	1

There are two processes for transforming ordinal data. The first one consists of transformation of a data file to a binary data file. In comparison to the case with nominal variables, *k* possible values of similarity should be considered where *k* is a number of categories. It is guaranteed by the coding shown in Table 2.

Table 2 Recoding of the ordinal variable Reaction for three binary variables P1 to P3

Reaction	P1	P2	P3
No	0	0	0
Weak	1	0	0
Medium	1	1	0
Strong	1	1	1

The second process makes use of the fact that values of an ordinal variable can be ordered. Under the assumption of the same distances between categories, the arithmetic operations can be done. It is recommended to code categories from 1 to *k* and divide these codes by the maximum value. In this way, the values will be in the interval from 0 to 1. Then we can apply the techniques designed for quantitative data.

C. Data Labeling:

Here's what the Orange Book says about data labeling: "Access control labels must be associated with objects. In order to control access to information stored in a computer, according to the rules of a mandatory security policy, it must be possible to mark every object with a label that reliably identifies the object's sensitivity level (e.g., classification), and/or the modes of access accorded those subjects who may potentially access the object.

III. USAGE OF ROCK & MARDL ALGORITHMS

A. Rock:

ROCK is a Robust Clustering using links is a clustering algorithm for data with categorical and Boolean attributes. It redefines the distances between points to be the number of shared neighbors whose strength is greater than a given threshold and then uses a hierarchical clustering scheme to cluster the data.

ROCK deals primarily with market basket data. Traditional Euclidean distance measures are not appropriate for such data and instead, ROCK uses the Jaccard coefficient to measure similarity. This rules out clustering approaches such as K-means or Centroid based hierarchical clustering. (However, K-means can actually be modified to work well for some non-Euclidean data, e.g., documents.) While the Jaccard coefficient provides a reasonable measure of the distance between points, clusters are sometimes not well separated and so a new measure of similarity between points was introduced that reflects the neighborhood of a point. If *sim(pi, pj)* is the similarity between points, *pi* and *pj*, and $0 \le \theta \le 1$ is a parameter, then

$$link(p_i, p_j) = |\{q : sim(p_i, q) \ge \theta\} \cap \{q : sim(p_j, q) \ge \theta\}|$$

In words, $link(p_i, p_j)$ is the number of shared neighbors of *pi* and *pj*. The idea is that two points will be "close" only if they share a relatively large number of neighbors. Such a strategy is intended to handle the problem of "border" points, which are close to each other, but belong to different clusters. ROCK also introduces a new objective function that is to be maximized:

$$E = \sum_{i=1}^{k} n_i * \sum_{p_q, p_r \in C_i} \frac{link(p_q, p_r)}{n_i^{1+2f(\theta)}}$$

Thus, we try to minimize the sum of the "link" similarity, i.e., the number of shared neighbors, between pairs of points in a cluster, subject to some scaling by the size of the cluster. This criterion can be used to derive a criterion for merging clusters via a hierarchical agglomerative scheme by merging the two clusters that lead to the largest increase in *E*. ROCK samples the data set in the same manner as CURE in order to avoid using a hierarchical clustering algorithm on a large number of points. This is followed by an assignment step where each remaining points is assigned to a cluster. A fraction of points is selected from each cluster and a calculation is performed to determine the number of those points that are neighbors of the point to be assigned. This quantity is scaled by the expected number of neighbors (based on the size of the cluster) and the point is assigned to cluster with the maximum number of neighbors after scaling.

The basic steps of ROCK are as follows:

a) Obtain a sample of points from the data set.
b) Compute the link value for each set of points, i.e., transform the original similarities computed by the Jaccard coefficient into similarities that reflect the number of shared neighbors between points.
c) Perform an agglomerative hierarchical clustering on the data using the "number of shared neighbors" similarities and the "maximize the shared neighbors" objective function defined above.
d) Assign the remaining points to the clusters that have been found.

B. Maximal Resemblance Data Labeling

The goal of MARDL, Maximal Resemblance Data Labeling, is to decide the most appropriate cluster label c_i^* for

Table 3. Shows the of cluster c1, c2, and c3

Cluster c_1		Cluster C_2		Cluster C_3	
d_{1j}	$w(d_{1j})$	d_{2j}	$w(d_{2j})$	d_{3j}	$w(d_{3j})$

[A_1=a] 0.027	[A_1=a] 0.009	A_1=a] 0.009
[A_1=b] 0.004	[A_1=b] 0.004	[A_1=b] 0.007
[A_1=c] 0.005	[A_1=c] 0.016	[A_1=c] 0.011
[A_2=m] 0.009	[A_2=m] 0.005	[A_2=m] 0.007
[A_2=f] 0.005	[A_2=f] 0.016	[A_2=f] 0.011
[A_3=a] 0.014	[A_3=a] 0.056	[A_3=a] 0.014
[A_3=b] 0.004	[A_3=b] 0.004	[A_3=b] 0.007
[A_3=c] 0.077		[A_3=c] 0.052

The unlabeled data point. Specifically, suppose that an unlabeled data point p(U,j) is given. MARDL computes the similarity S(c_i , p(U,j)) between p(U,j) and cluster c_i, $1 \le i \le n$, and finds the cluster which has max(S(c_i, p(U,j))). The similarity between p(U,j) and ci can be obtained in light of the concept of calculating the similarity between the query string and the document in the vector-space model as mentioned before [8]. The cluster represented by NIR can be mapped to a node vector, which is similar to the term vector used in the vector-space model to describe document. Moreover, the unlabeled data point can be seen as a query string which consists of nodes. As a result, in MARDL, the similarity between p(U,j) and c_i can be deemed as the similarity between a query string and a document. In view of the above, the similarity, referred to as *resemblance* in this paper, is defined below. (*Resemblance and Maximal Resemblance*)**:** Given an unlabeled data point p(U,j) and a NIR table of cluster ci, the *resemblance* is defined by the following equation:

$$R(p_{(U,j)}, c_i) = \sum_{x=1}^{q} w(c_i, d_{ix}),$$

where d_{ix} is one entry in the NIR table of cluster c_i. The value of resemblance R(p(U,j), c_i) can be directly obtained by summing up the importance of nodes in the NIR table of the cluster c_i, where these nodes are decomposed from the unlabeled data point p(U,j). This equation which sums the nodes importance considers how much the unlabeled data point is similar to the cluster based on the nodes in the unlabeled data point. When an unlabeled data point contains nodes which are more important in the cluster c_i than the cluster cj , R(p(U,j), c_i) will be larger than R(p(U,j), c_j). Finally, an unlabeled data point p(U,j) is labeled to the cluster which obtains the maximal resemblance. The decision function is defined by Eq. (5).

$$Label = \arg\max_{c^*} R(p_{(U,j)}, c_i), \text{ where } 1 \le i \le n$$

Since we measure the similarity between the unlabeled data point p(U,j) and the cluster ci as the R(p(U,j), ci), the cluster with the maximal resemblance is the most appropriate cluster for the unlabeled data point.

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

i. NIR hash table N Table = Cluster Analysis(C);
ii. Data Labeling(N Table, U);

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

i. while has next tuple in C {
ii. read in data point p(i, j) into nodes;
iii. divided p(i, j) into nodes;
iv. update node frequency in cluster c_i ;
v. }
vi. For each node d_{i1} to d_{it}
vii. Compute weight f(d_{ix});
viii. for each cluster c_1 to c_n {
ix. for each node d_{i1} to d_{it} {
x. calculate node importance w_i , $d_{ix;}$
xi. add (d_{ix}, w_i, d_{ix}) into INR table N table;
xii. }
xiii. }
xiv. return N Table;

Procedure Data Labeling (N Table, U): give each unclustered Data point a cluster label

i. while has next tuple in U {
ii. read in data point p(u, j) from U;
iii. divided p(u, j) into nodes;
iv. for each cluster c_1 to c_n
v. calculate Resemblance R(p(u, j), c_i);
vi. find Maximal Resemblance cm
vii. give label cm to p(u, j);
viii. }

IV. SOFTWARE IMPLEMENTATION TOOL FOR THE PROPOSED SYSTEM

Better performance based on the data sampling for categorical data, has two modules cluster analysis and data labeling which we need to take any of data mining tool below. For the proposed analysis weka is best java based data mining tool

A. Rapid Miner

Formerly called as YALE (Yet another Learning Environment), is an environment for machine learning and data mining experiments that is utilized for both research and real-world data mining tasks. It enables experiments to be made up of a huge number of arbitrarily nestable operators, which are detailed in XML files and are made with the graphical user interface of Rapid Miner. Rapid Miner provides more than 500 operators for all main machine learning procedures, and it also combines learning schemes and attribute evaluators of the Weka learning environment. It is available as a stand-alone tool for data analysis and as a data-mining engine that can be integrated into your own products.

B. Weka

Written in Java, Weka (Waikato Environment for Knowledge Analysis) is a well-known suite of machine learning software that supports several typical data mining tasks, particularly data preprocessing, clustering, classification, regression, visualization, and feature selection. Its techniques are based on the hypothesis that the

data is available as a single flat file or relation, where each data point is labeled by a fixed number of attributes. Weka provides access to SQL databases utilizing Java Database Connectivity and can process the result returned by a database query. Its main user interface is the Explorer, but the same functionality can be accessed from the command line or through the component-based Knowledge Flow interface.

C. Comparative Study

Sampling has been recognized as an important technique to improve the efficiency of clustering. However, with sampling applied, those points which are not sampled will not have their labels after the normal process. Although there is a straightforward approach in the numerical domain, the problem of how to allocate those unlabeled data points into proper clusters remains as a challenging issue in the categorical domain. Numerous solutions like ROCK exists to apply clustering based on data sampling approach, they are able to perform efficiently only on numeric data and lagging or most of the times failed on categorical data. A mechanism, named Maximal Resemblance Data Labeling (abbreviated as MARDL), to allocate each categorical un-clustered data, This proposal should target to find a better solution for clustering approach based on data sampling concept for categorical data. The total process will be managed in different phases mentioned below:

a. Cluster Analysis Phase:

In the cluster analysis phase, a cluster representative is generated to characterize the clustering result.

b. Data Labeling Phase:

In the data labeling phase, each unlabeled data point is given a label of appropriate cluster

V. CONCLUSION

In this paper, the proposed MARDL to allocate each unlabeled data point into the appropriate cluster when the sampling technique is utilized to cluster a very large categorical database. In addition, developing the categorical cluster representative technique, the experimental evaluation validates 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. When the dimensionality of data is large, MARDL with NNIR improves the quality of data labeling because the combination of attribute values is considered. Consequently, MARDL has cluster analysis and data labeling, which are implemented in data mining tools like Rapid Miner, Weka is significantly more efficient than Rock works while attaining results of high quality.

VI. REFERNCES

[1] Jiawei Han, Micheline Kamber, “Data Mining Concepts and Techniques” Elsevier Publication.

[2] P. Berkhin, 2002. Survey of Clustering Data Mining Techniques. Technical report, Accrue Software, San Jose, Cailf.

[3] P. Berkhin, “Survey of Clustering Data Mining Techniques,” technical report, Accrue Software, 2002.

[4] A.K. Jain, M.N. Murty, and P.J. Flynn, “Data Clustering: A Review,” ACM Computing Surveys, 1999.

[5] D.S.J.M.R. Garey and H.S. Witsenhausen, “The Complexity of the Generalized Lloyd-Max Problem,” IEEE Trans. Information Theory, 1982.

[6] A.K. Jain, M.N. Murty and P.J. Flynn, “Data Clustering: a review”,

[7] P.J. Flynn, A.K. Jain, M.N. Murty, 1999. Data Clustering: A Review. ACM Computing Surveys, vol. 31, no. 3: 264-323.

[8] R. Baeza-Yates and B. Riberiro-Neto. Modern Information Retrieval. Addison-Wesley, 1999.

[9] Joachim Buhmann and Hans Kuhnel, (1993), Vector Quantization with Complexity Costs, IEEE Transactions on Information Theory 39:1133-1145. http://www-dbv.informatik.uni-bonn.de/papers.html

[10] C. Fraley and A. E. Raferty, (1998), How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis, Technical Report No. 329, Department of Statistics, University of Washington, Seattle,Washington. http://www.stat.washington.edu/fraley/

[11] D.H. Fisher, “Knowledge Acquisition via Incremental Conceptual Clustering,” Machine Learning, 1987.

[12] “Data Generator: Perfect Data for an Imperfect World,” http:// www.datasetgenerator.com, 2008.

[13] R. Agrawal, T. Imielinski, and A. Swami, “Mining Association Rules between Sets of Items in Large Databases,” Proc. ACM SIGMOD, 1993.

M. Sathya Narayana B. Tech and M.Tech in Computer Science Engineering. currently working as a Asst. Professor and HOD of IT department in VISVESVARAYA COLLEGE OF ENGINEERING AND TECHNOLOGY, having 4 years of teaching experience. His areas of interest include Data mining, Networks. sathyam542@gmail.com

B. V. V. S. Prasad MCA from M.Tech CS, Asst Prof in VISVESVARAYA COLLEGE OF ENGINEERING AND TECHNOLOGY, Hyderabad. Having 4 years Academic experience, an associate member of CSI and life member of ISTE. Certified by Sun JAVA professional and Oracle 10g Admin. Published one international journal, his research areas are Data Mining, Networks vvsprasad.author@gmail.com

N.Sivaram Babu B.Tech & M.Tech in Computer Science Engineering. Currently working as Software Engineer Analysts in TATA Consultancy Services having 4years of industry experience, certified in Oracle (OCA-1), Quality Center and Six Sigma Green Belt Trained (Lean SixSigma-E1). nsivarambabu@gmail.com

B.Suresh Kumar completed MCA and M.Tech Computer Science Engineering, working as Asst Prof in Jayoti Vidyapeeth Womens University, Jaipur, Rajasthan. Having 5years of Academic experience, researching on Cache Data base which is object oriented database. sureshkumar5656@gmail.com