ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)

International Conference on Recent Trends in Computing and Communication (ICRTCC 2015)

Cochin College of Engineering & Technology

Vol. 4, Special Issue 1, June 2015

Survey on Data Preprocessing Techniques for Fuzzy Association Rule Mining

ViniVijayan[1], Dr. M. Sudheep Elayidom[2]

Research Scholar, School of Engineering, Cochin University of Science and Technology, Kerala, India [1]

Associate Professor, School of Engineering, Cochin University of Science and Technology, Kerala, India [2]

Abstract: Data mining is the technique to find previously unknown and useful knowledge from a large database. Many association rule mining algorithms are introduced for data mining process. If we integrate the interesting computational logic of fuzzy set with association rule mining, it is called as Fuzzy Association Rule Mining (FARM).FARM will provide a better result for association rule mining strategy. It needs some preprocessing of data set before apply it to FARM. The main data preprocessing task in data mining is the numerical attribute management. . Most of the algorithms require the discretization of numerical attributes by splitting the continuous range of values into intervals. The data preprocessing techniques classify the entire data set in to fuzzy clusters. In this paper, we are summarized a number of suggested techniques for data preprocessing. And sort out an algorithm which seems to be the optimal in the group.

Keywords: Data mining, Apriori algorithm, association rule, fuzzy set, fuzzy rule mining.

I. INTRODUCTION

Data Mining, which is also known as, Knowledge Discovery in Databases (KDD), is a process of finding new, interesting, previously unknown, potentially useful, and ultimately understandable patterns from very large volumes of data. There are a number of techniques are available to perform association rule mining. Apriori algorithm is considered as the best known algorithm for the association rule mining. The rules generated by association rule mining are written in the form of A➔B. That means the probability to occur A and B together (both are the data items) meets a specified threshold value. The threshold value is calculated for support and confidence.

Fuzzy association rule mining is a new thought in the area of association rule mining. In this technique divide the entire data set in to different fuzzy clusters, and each data item have a degree of membership in every cluster instances. Dividing the data set in to different clusters are termed as data preprocessing. There are different techniques implemented for data preprocessing. The mainly used techniques are:

1. Hierarchical clustering
2. fuzzy k-means and C-means clustering
3. CLARANS clustering algorithm

The goal of clustering is to group data items that are similar to each other and to identify such clusters in an unsupervised manner. In ***Hierarchical Clustering,*** a set of N items to be clustered, by the help of a N*N distance matrix. At the initial stage of Hierarchical clustering, If you have N items in the given set, you now have N clusters, because, this technique assign each item to a cluster each. Thus each clusters containing just one item. Then the distances (similarities) between the clusters is same as the distances (similarities) between the items they contain .The next step is to find the most similar pair of clusters and merge them into a single cluster, so the number of clusters will be reduced by one. Then compute distances (similarities) between the new cluster and each of the old clusters. Repeat the process until all items are clustered into a single cluster of size N.

Hierarchical clustering is again classified as, Single-linkage, complete-linkage and average-linkage clustering. In single-linkage clustering, the distance between one cluster and another cluster is calculated as equal to the shortest distance from any member of one cluster to any member of the other cluster. The main disadvantage is the similarity between one cluster and another cluster will be equal to the greatest similarity from any member of one cluster to any member of the other cluster. In complete-linkage clustering, the distance between one cluster and another cluster is calculated as equal to the greatest distance from any member of one cluster to any member of the other cluster. The main disadvantage is its always trying to project the dissimilarity. In average-linkage clustering, the distance between one cluster and another cluster will be equal to the average distance from any member of one cluster to any member of the othercluster. As compared to others average-linkage clustering is more effective.

Hierarchical clustering is sometimes called agglomerative because it merges clusters iteratively. There is also a divisive hierarchical clustering which does the reverse by starting with all objects in one cluster and subdividing them into smaller pieces. Divisive methods are not generally available, and rarely have been applied. In hierarchical clustering, we can't make the N items grouped in a single cluster but, if we create the complete

DOI 10.17148/IJARCCE

hierarchical tree, for k clusters you just have to cut the k-1 longest links. Fuzzy c-means (FCM), is a method of clustering which allows one piece of data to belong to two or more clusters. It is mainly used in pattern recognition applications. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2 \qquad 1 \le m < \infty$$

x_i is the i^{th}entry of measured data, c_j is the center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center.where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j. Fuzzyupdate of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right)^{\frac{2}{m-1}}} \quad , \quad c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m}\cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

Where there are k iteration steps, the iteration will stop when,$\max_{ij}\{|U_{ij}^{(k+1)}-U_{ij}^{k}|\} < \varepsilon$ where ε is a termination criterion between 0 and 1. This procedure converges to a local minimum or a saddle point ofJ_m.

The algorithm is composed of the following steps:

Initialize U=[uij] matrix, U(0)

At k-step: calculate the centers vectors C(k)=[cj] with U(k)

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m}\cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right)^{\frac{2}{m-1}}}$$

Update U(k) , U(k+1)

If || U(k+1) - U(k)||< ε then STOP; otherwise return to step 2.

Data are bound to each cluster by means of a membership function, it produce the fuzzy behavior of this algorithm. For implementing this algorithm, we need to create a matrix named U which contents are numbers between 0 and 1, and it represents the represent the degree of membership between data and centers of clusters. Given a certain data set, suppose to represent it as distributed on an axis. The figure below shows this:

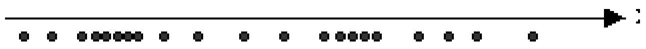


Fig 1: Data cluster example

In the above example, we can identify two clusters in proximity of the two data concentrations. We will refer to them using 'A' and 'B'.

In the fuzzy k-means algorithm - each datum attached to a specific centroid. Then the membership function looked like this:

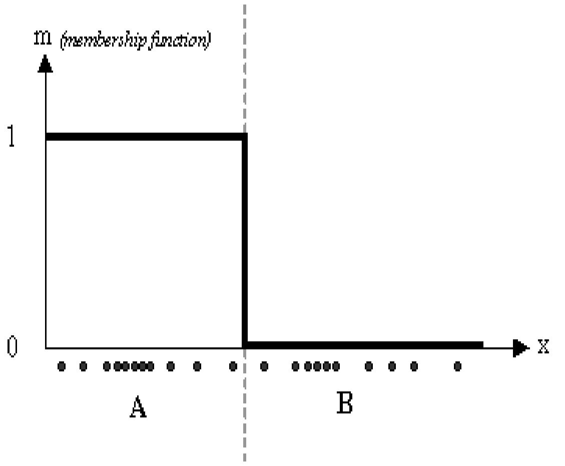


Fig2: Graphical representation of Fuzzy k-means clusters

The clusters produced by the k-means procedure are called "hard" or "crisp" clusters. In this any feature vector **x** either is or is not a member of a particular cluster. In the FCM approach, the given datum does not belong exclusively to a single cluster. So it is called "soft" or "fuzzy" clusters. In this case, the membership function follows a smoother line to indicate that every datum may belong to several clusters with different values of the membership coefficient.

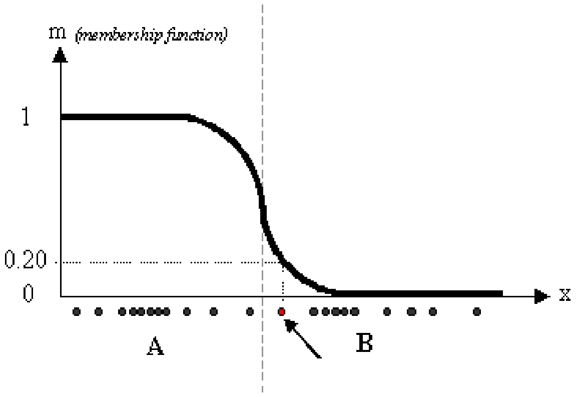


Fig 2: Graphical representation of Fuzzy c-means clusters

The datum shown as a red marked spot belongs more to the B cluster rather than the A cluster. The value 0.2 of 'm' indicates the degree of membership to A for such datum. Then we can convert the graphical representation to matrix

$$U_{MxC} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ .. & .. \\ 0 & 1 \end{bmatrix} \qquad U_{MxC} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \\ 0.6 & 0.4 \\ .. & .. \\ 0.9 & 0.1 \end{bmatrix}$$

The number of rows and columns depends on how many data and clusters we are considering. That means C is the total number of clusters and N is the total number of data. In the above examples first is the k-meansand next is the FCM.In the first case the coefficients are always unitary. It is because each datum can belong only to one cluster. Other properties are shown below:

$$u_{ij} \in [0,1] \qquad \forall i,j$$

$$\sum_{j=1}^{C} u_{ik} = 1 \qquad \forall i$$

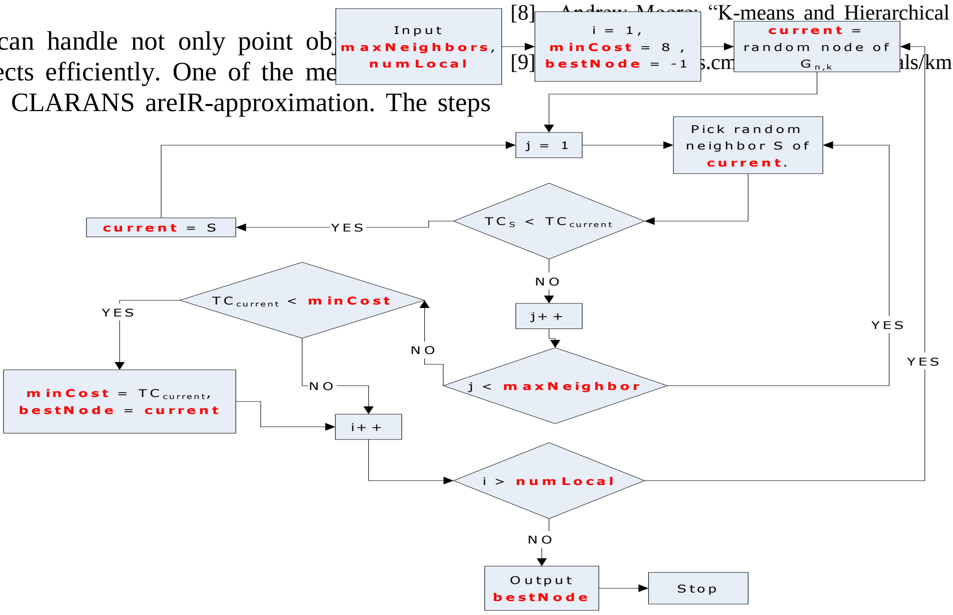
$$0 < \sum_{i=1}^{N} u_{ij} < N \quad \forall N$$

FCM gives best result for overlapped data set and comparatively better than k-means algorithm.Unlike k-means where data point must exclusively belong to one cluster center, here data point is assigned membership to each cluster center as a result of which data point may belong to more than one cluster center.

Spatial data mining is the discovery of interesting relationships and characteristics that may exist implicitly in spatial databases. **CLARANS(Clustering Large Applications based on RANdomized Search)** is a clustering method, which is used to identify spatial structures that may be present in the data. Experimental results indicate that, when compared with existing clustering methods, CLARANS is very efficient and effective.

CLARANS can handle not only point obj polygon objects efficiently. One of the me applicable to CLARANS areIR-approximation. The steps of CLARANS are given in Fig 3:

Fig 3: Steps of CLARANS algorithm.



II. CONCLUSION

In this paper we compared four mainly used data preprocessing techniques for fuzzy association rule mining. Hierarchical clustering is least effective, because it does not accurately find the clusters. It complexity is O(n2), where n is the number of total objects. It has no option to undo what was one previously. CLARANS is only applicable for special data. On comparing all the four technique , we find out that fuzzy C- means clustering is the most effective technique for fuzzy association rule mining, because, it distribute a single data to all the featured clusters with a coefficient of membership function. In fuzzy association rule mining we can use these coefficients for calculating support and confidence.

REFERENCES

[1] J. C. Dunn (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact

[2] Well-Separated Clusters", Journal of Cybernetics 3: 32-57

[3] J. C. Bezdek (1981): "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York

[4] S. C. Johnson (1967): "Hierarchical Clustering Schemes" Psychometrika, 2:241-254

[5] R. D'andrade (1978): "U-Statistic Hierarchical Clustering" Psychometrika, 4:58-67

[6] R. Agrawal, J. Gehrke, D. Gunopulos and P. Raghavan, "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications," Proc. 1998 ACM-SIGMOD, pp. 94105, 1998.

[7] M. Ankerst, M. Breunig, H.-P. Kriegel and J. Sander, "OPTICS: Ordering Points To Identify the Clustering Structure," Proc. 1999 ACM Special Interest Group on Management of Data, pp. 4960, 1999.

[8] Andrew Moore: "K-means and Hierarchical Clustering - Tutorial

[9] s.cm als/kmeans.html