Review of K Mean and Fuzzy Clustering Algorithm

Harmeet Kaur[1], Manjit Kaur[2]

[1]M. Tech CSE, Lovely Professional University, Phagwara, Punjab, India
[2]Assistant Professor, Lovely Professional University, Phagwara, Punjab, India

Abstract*—Data mining is the process of discovering patterns in large data sets. It has attracted a lot of attention from various industries and researchers for critical decision making and development. Researchers has proposed a number of algorithms have been proposed in past for clustering of data as data mining techniques. There are multiple techniques of mining relevant information from existing databases. K-Means is most common used algorithm for clustering. Multiple versions of K-Means have been proposed with different improvements. In this paper, we will review few data clustering techniques.*

Keywords*— *Data mining, data mining techniques, K-means, Fuzzy k means.

I. INTRODUCTION

The world around us contains large volume of raw data. With the advancement of information technology data is growing enormously. Data from different aspects of behavior is collected and saved. Data warehouses are built to manage large amount of historical and incoming data. Data mining finds out the interesting patterns and hidden relationships among large volume of raw data. This extracted information is used for making critical decisions for development and future activities of the businesses. Clustering is used to identify underlying structure in data. It can be defined as: Given a representation of n objects, find K groups based on a measure of similarity such that the items in one group are more similar to each other than items in other groups. One of the most important clustering algorithms is K-Means algorithm. A review was provided on data clustering, major issues and challenges in designing clustering algorithms [1].

II. DATA MINING TECHNIQUES

Classification: It is a data mining technique which involves classifying the newly introduced objects in to predefined classes based on multiple attributes. For example,classifying credit applicants as low, medium or high risk, classifying engine faults by their symptoms, classifying web attacks in to intrusions, anamolies. The given data set is divided in to training set and testing set. The input data, also called the training set, consists of multiple records each having multiple attributes or features. Each record is tagged with a class label. The objective of classification is to analyze the input data and to develop an accurate descriptor or model for each class using the attributes present in the data. This model is used to classify test data for which the class descriptions are not known. Accuracy rate is defined as the percentage of test set samples that are correctly classified by the model. Various classification techniques are used for example; decision tree based, Rule based, neural network based and memory based reasoning etc.

Association Rule mining: It is a data mining technique which finds out association or correlation relationships among large set of data items. Association rules are used to uncover relationships between unrelated dataset. In data mining, it is used to analyze and predict customer purchasing pattern which helps in business development applications like marketing promotions, inventory management and customer relationship. Association rules are created by analyzing data for frequent if/then patterns. Association rules should satisfy both a minimum support threshold and minimum confidence threshold. [6] A transaction t contains X, a set of items (itemset) in I, if $X \subseteq t$. An association rule is an implication of the form:

$X \rightarrow Y$, where $X, Y \subset I$, *and* $X \cap Y = \varnothing$

The rule $X \rightarrow Y$ holds in the transaction set t with support s, where s is the percentage of transactions in T (transaction data set) that contain $X \cup Y$. The rule $X \rightarrow Y$ has confidence c. in the transaction set T if c is the percentage of transactions in T containing X

which also contain Y. That is, support$(X \rightarrow Y)$ = Prob$\{X \cup Y\}$ and confidence$(X \rightarrow Y)$ = Prob$\{Y/X\}$.

An association rule is a pattern that states when X occurs, Y occurs with certain probability.

Clustering: Clustering technique is used in search engines for grouping similar objects in to one group and dissimilar objects in to other group hence getting required result on front page, in academics to group students according to

their level of performance, in biology to find group of genes having similar functions, detecting intrusions where malicious data and normal data are grouped into different cluster and many other applications [5].

In image processing applications, clustering is used to find image as similar as provided by query image. Images are grouped in to given number osf clusters based on features like color, texture, shape contained in images in form of pixels.

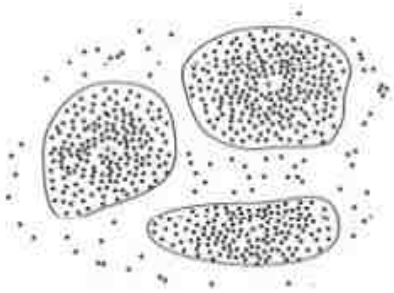

Fig.1: Three clusters are formed based on similarities

III. K MEAN ALGORITHMs

According to study various clustering algorithms have been designed after K mean algorithm but still K-mean is widely used. It is one of the oldest and most popular clustering algorithm developed till date [2].

The k-means algorithm is a simple iterative method to partition a given dataset into set of K clusters, C= {C_k, k=1…K}. The algorithm operates on a set of n d-dimensional vectors, D = {xi | i = 1. . . N}, where xi □ _R_d denotes the ith data point. The algorithm is initialized by picking k points in R_d as the initial k cluster representatives or "centroids".

The basic K-Mean algorithm is:

1. Select k points as initial centriods
2. **Repeat**
3. Form k clusters by assigning each point to its closest centriod
4. Recompute the centriod of each cluster
5. **Until** the centriods do not change

Step 1: Data Assignment. Each data point is assigned to its closest centroid, with ties broken arbitrarily. This results in a partitioning of the data. Centroid is typically the mean of the points in the cluster. 'Closeness' is measured by Euclidean distance, cosine similarity, correlation, etc.

Step 2: Re-computation of "centroid". Each cluster representative is relocated to the center (mean) of all data points assigned to it.

The algorithm terminates at convergence condition that is k means reaches a state where no points are shifting from one cluster to another; hence no change in centriods. Each iteration needs N × k comparisons, which determines the time complexity of one iteration. The number of iterations required for convergence varies and may depend on N.

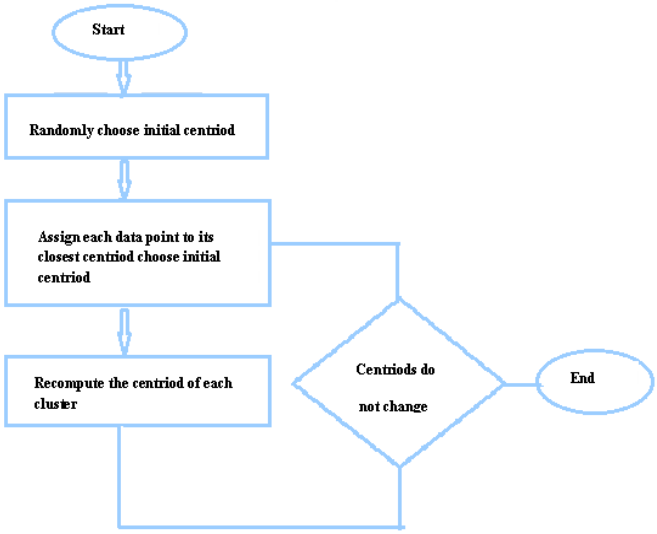

Fig.2: K means algorithm flowchart

K-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized.

Let µk be the mean of cluster C_k. The squared error between µk and the points in cluster C_k is defined as

$$J(c_k) = \sum_{x_i \in c_k} ||x_i - \mu_k||^2.$$

The goal of K-means is to minimize the sum of the squared error over all K clusters,

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} ||x_i - \mu_k||^2.$$

Complexity is O (n*k*l*d) where

n = points, k= number of clusters,

I = number of iterations, d = number of attributes

Problems with k-means algorithm

- The standard k-means algorithm needs to calculate the distance from the each data object to all the centers of k clusters when it executes the iteration each time, which takes up a lot of execution time especially for large-capacity databases.
- Choosing K is another problem with K mean algorithm.
- K-means has problems when the data contains outliers.

- K means algorithm has problem of empty clusters i.e. no data point may be assigned to a cluster during execution.

Various algorithms were proposed in past to solve these problems. An improved k-mean clustering algorithm for improved clustering with reduced complexity. This algorithm combines a systematic method for finding initial centroids and an efficient way for assigning data points to clusters. This method ensures the entire process of clustering in O (n2) time without sacrificing the accuracy of clusters [3].The problem of calculating distance from each data object to all the centers of k clusters in each iteration was taken into consideration by using an improved k mean clustering algorithm [4]. This improved algorithm uses two simple data structures to retain the labels of cluster and distance of all data objects to the nearest cluster in each iteration that can be used in next iteration. The distance between current data object and new cluster centre is computed, if it is less than or equal to the distance to the old centre, the data object remains in same cluster as assigned in the previous iteration. This reduces the number of distance calculation which effectively improves the speed of clustering and time complexity.

Fuzzy K-mean

It is a variation of K means algorithm in which soft clustering is used i.e. relationship between data point and cluster centre is fuzzy. A data point can belong to one or more cluster with membership grades between zero and one rather than belonging to just one cluster as in traditional k-means where hard clustering is used.

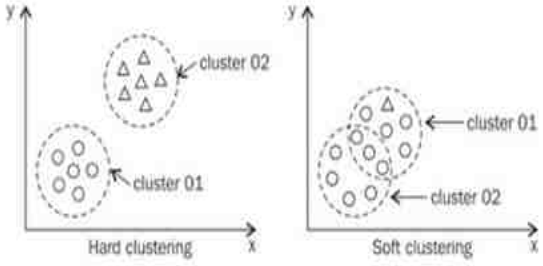

Fig.3: Hard clustering and soft clustering

Fuzzy clustering is used in real applications where there are no sharp boundaries between clusters. One of the examples is market segmentation where each customer is assigned a fuzzy score which provides precise measure to delivers values to customers and profit to the company. Data points are partitioned in to k clusters S*p* (*p*=1, 2 …k) and clusters S*p* is associated with cluster centers C*p*.

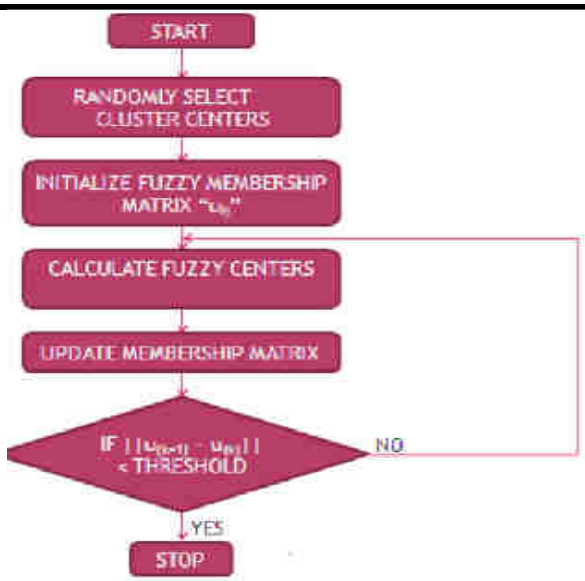

Fig.4: Fuzzy k algorithm Flowchart

Fuzzy K means algorithm is as follows:

- Choose a fixed number of clusters
- Calculate the fuzzy membership and cluster centers
- Update the membership function and cluster centers in each iteration

  Equation to update membership is

$$u_{i,j} = \left((d_{ij})^{1/m-1} \sum_{l=1}^{k} \left(\frac{1}{d_{il}} \right)^{1/m-1} \right)^{-1}$$

Here, d_{ij} is calculated for i=1 to N and j= 1 to k

Whereas new cluster centers are computed as

$$\mathbf{C}_j(p) = \frac{\sum_{i=1}^{N} u_{ij}^m \mathbf{X}_i}{\sum_{i=1}^{N} u_{ij}^m}$$

- Iterate till convergence is met or objective function is minimized or $||U^{(k+1)} - U^{(k)}|| < \beta$. Where,

  'k' is the iteration step.

  'β' is the termination criterion between [0, 1].

  'U = $(\mu_{ij})_{n*c}$' is the fuzzy membership matrix.

  Objective function is defined as

$$J = \sum_{j=1}^{k} \sum_{i=1}^{N} u_{i,j}^{m} d_{ij}$$

Where, u_{ij} is membership function representing the degree of belongingness between data point x_i and cluster center c_j
d_{ij} is squared Euclidean distance between data point x_i and cluster center c_j
N is number of data points
m is fuzzifier parameter
k is number of clusters

Problems with Fuzzy k-means

- Long Computational time
- With lower value of β we get the better result but at the expense of more number of iteration.
- Euclidean distance measures can unequally weight underlying factors.

IV. CONCLUSION

K-Means partitioning based clustering algorithm required to define the number of final cluster (k) beforehand. Such algorithms are also having problems like susceptibility to local optima, sensitivity to outliers, memory space and unknown number of iteration steps that are required to cluster. The time complexity of the K-Means algorithm is O (nkdi) and the time complexity of Fuzzy k algorithm is $O(ndk^2i)$. From the obtained results we may conclude that K-Means algorithm is better than Fuzzy k algorithm. Fuzzy k means produces close results to K-Means clustering but it still requires more computation time than K-Means because of the fuzzy measures calculations involvement in the algorithm. Infact, Fuzzy k means clustering which constitute the oldest component of software computing, are really suitable for handling the issues related to understand ability of patterns, incomplete/noisy data, mixed media information, human interaction and it can provide approximate solutions faster. They have been mainly used for discovering association rules and functional dependencies as well as image retrieval. So, it is concluded that K-Means algorithm is superior to Fuzzy k-Means algorithm.

REFERENCES

[1] Anil K. Jain ,"data clustering 50 years beyond K-means", Pattern Recognition Letters 31(2010) 651-666
[2] Mac Queen, J., 1967.Some methods for classification and analysis of multivariate observations (pp. 281–297).
[3] K. A. Abdul Nazeer & M. P. Sebastian" Improving the Accuracy and Efficiency of the K-Means Clustering Algorithm" .Proceedings of the World Congress on Engineering 2009 Vol. I WCE 2009, London, U.K, July 1 - 3, 2009.s
[4] Shi Na, Liu Xumin and Guan Yong," research on K mean, Clustering Algorithm," Third International Symposium on Intelligent Information Technology and Security Informatics, 2010
[5] Shraddha Shukla and Naganna S.," A Review ON K-means DATA Clustering APPROACH" International Journal of Information & Computation Technology,Volume 4, Number 17 (2014), pp. 1847-1860
[6] Irina Tudor, "association rule mining as a data mining technique", Vol. LX, No. 1/2008, 49-56, buletinul, Universitatea Petrol-Gaze din Ploieşt.