IJIAET International Journal of Innovative Analyses and Emerging Technology

| e-ISSN: 2792-4025 | http://openaccessjournals.eu | Volume: 1 Issue: 7

Classification of Liver Samples by Fuzzy Clustering Algorithms

D. Latha

Department of Computer Science and Engineering, AdikaviNannaya University, Andhra Pradesh, India

B. Venkataramana

Department of Computer Science & Engineering, Holy Mary Institute of Technology, Hyderabad, India

Abstract: Partition based clustering algorithms are widely used in data clustering. The most popular methods are fuzzy algorithm, Fuzzy c-Means (FCM), and non-fuzzy algorithm, k-means (KM) methods. K-means and Fuzzy c-Means use centroid distance measure and standard Euclidian distance measure respectively. In this work, a comparative study of these algorithms with liver disorder data set from the UCI repository is presented. Repository results were compared with these results. Based on the clustering output criteria the performance of these two algorithms is analyzed in terms of percentage of correctness and classification performance. The experimental results demonstrate that k-means outperforms the Fuzzy c-Means algorithm. Thus the efficiency of k-means is better than that of Fuzzy c-Means.

Keywords: Fuzzy C-Means, K Means, Fuzzy Clustering

Introduction:

Organizations store and maintain huge volume of data in their databases. Knowledge discovery or Data mining is the method to extract the most useful knowledge from the databases. Data mining is an analytic process of discovering valid, unsuspected relationships among datasets and transforms the data into a structure that are both understandable and useful to the users. Data analysis contains several techniques and tools for handling the data. Cluster is a collection of data objects that are similar to one are dissimilar to the objects in other clusters. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering.

Classification or clustering is well known method in data analysis. In a dataset to partition the dataset into group classification or clustering analysis is needed. It is multivariate analysis technique to partition the dataset into groups so that the most indiscernible objects belong to the same group while the discernible objects in different groups. In many fields using clustering methods generate data streams have become more popular such as medical diagnostics and bioinformatics researches, machine learning, pattern recognition, image segmentation,. Clustering is also often called as Classification. Clustering is an important tool in medical diagnostics, pattern recognition, data analysis, data mining, image processing and etc [1].

Liver is the largest organ in the body. It contributes about 2% of total body weight or 1.5 kg in the average adult human. The basic unit of liver is liver lobule. The human liver contains 50,000 to 1,00,000 individual lobules. The lobule consists of liver cellular plates that radiated from the central vein like

spokes in a wheel. In between the lobules there is a portal triad; it consists of bile duct, hepatic artery, portal vein.

Different methods in clustering are partition based clustering and hierarchical-based clustering. The clusters are formed to optimize an objective Partition criterion, such as a dissimilarity function based on distance. Partioning clustering algorithms have the capable of discovering underlying structures of clusters by using appropriate objective function [2]. Partition-based clustering algorithms are widely used in k-means (KM), Fuzzy c-Means (FCM) .The algorithms k-means and Fuzzy c-Means are proposed based on Euclidean distance measure

Several comparisons are carried out by the following researchers: Jaindong, Hongzan, Jaiwen, Qiyong [3] analyzed the performance of k-means and Fuzzy c-Means algorithms and reported that the k-means method is preferable to FCM for Arterial Input Function (AIF) detection using both clinical and simulated data. Velmurugun [4] has compared the clustering performance of k-means and Fuzzy c-Means algorithms using different shapes of arbitrary distributed data points and reported that the k-means performs better than FCM method. Wang and Garibaldi [5] have compared the performance of k-means and Fuzzy c-Means algorithms on Infrared spectra collected from auxiliary lymph node tissue section. . Soumi Gosh and Sanjay Kumar Dubey [6] evaluated the clustering performance of k-means and Fuzzy c-Means algorithms on the basis of the efficiency of the clustering output and the computational time and reported that k-means is superior to FCM. Bharati and Gohokar [7] compared the color image segmentation performance between k-means and Fuzzy c-Means algorithms. The work in this paper aimed to compare the performance of the two clustering techniques, k-means (KM), Fuzzy c-Means (FCM). The most popular real world date sets such as Liver Disorders is applied to test the performance of these algorithms and a comparative analysis is presented in this work.

2. Materials and methods:

Clustering is an unsupervised data analysis which is used to partition a set of records or objects into clusters or classes with similar characteristics. The partition is done in such a fashion that most similar (or related) objects are placed together, while dissimilar (or unrelated) objects are placed in different classes or groups. The desired characteristics of clustering methods are ability to deal with different types of attributes with high dimensionality, effective handling of outliers and noise with minimum knowledge, ability to discover the underlying shapes and structures of the data, scalability, usability and interpretability. Five different Clustering methods are categorized: partitioning method, hierarchical method, density based method, grid based method and model based or soft computing methods. Among these five methods partition based methods, k-means (KM), Fuzzy c-Means (FCM) clustering algorithms are implemented using two well known data sets liver disorders and wine to generate two clusters .

2.1. The dataset:

The real world data sets Liver Disorder is obtained from the UCI Machine Learning Repository donated by Richard [8]. The Liver data set contains 341 samples with 6 attributes or blood tests each. These blood tests are capable of detecting liver disorders which might arise due to excessive alcohol consumption. The attributes are the measurements of the blood tests namely mean corpuscular volume

(mcv), alkaline phosphatase (alkphos), alanine aminotransferase (sgpt), aspartate aminotransferase (sgot), gamma-glutamyl transpeptidase (gammagt) and the number of half-pint equivalents of alcoholic beverages drunk per day (drinks). The 341 samples are clustered into two different classes according to the liver disorders: Class 1 containing 142 samples and Class 2 containing 199 samples.

2.2. k-means clustering

MacQueen [9] introduced the k-means algorithm in 1967. It is a partitioning algorithm. Partitioning method first creates an initial set of k partitions, where parameter k is the number of partitions to construct. It then uses an iterative relocation technique that attempts to improve the portioning by moving objects from one group to another. Typical partitioning methods include k-means and their improvements.

Given D, a data set of n objects, and k, the number of clusters to form, a partitioning algorithm organizes the objects into k partitions, where each partition represents a cluster. The clusters are formed to optimize an objective partition criterion, such as dissimilarity function based on distance, so that the objects within a cluster are similar, whereas the objects of different clusters are dissimilar in terms of the data set attributes. K-means algorithm is an iterative method. This algorithm can be run several times to reduce the sensitivity caused by initial random selection of centroids.

Method:

1. Arbitrarily choose k objects from D as the initial cluster centers;
2. Repeat
3. (re)assign each object to the cluster to which the object is most similar, based on the mean value of the objects in the cluster
4. Update the cluster mean i.e., calculate the mean value of the objects for each cluster;
5. Until no change;

2.3. Fuzzy c-Mean clustering

Fuzzy c-Means algorithm (FCM) is one of the most popular fuzzy clustering methods. FCM is developed based on fuzzy theory. In this method it uses membership function to assign membership values ranged from 0 to 1 to each object. The feature in FCM is that every object belongs to every cluster with different membership values. The partition of the dataset Z into c clusters is represented by the fuzzy partition matrix $U = [\mu_{ik}]_{cxN}$

. The fuzzy partitioning space for Z is the set

$$M_{fc} = \left\{ U \epsilon \Re^{cxN} /_{\mu_{ik} \epsilon [0,1]}, \forall i,k; \sum_{i=1}^{c} \mu_{ik} = 1, \forall k; 0 < \sum_{k=1}^{N} \mu_{ik}, \forall i \right\}$$

IJIAET International Journal of Innovative Analyses and Emerging Technology
| e-ISSN: 2792-4025 | http://openaccessjournals.eu | Volume: 1 Issue: 7

Fuzzy c-Mean model achieves its partitioning by the iterative optimization of its objective function given as

$$\underbrace{min}_{U,V}\left\{J(Z;U,V)=\sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^{m}\|z_k-v_i\|^2{}_A\right\}$$

Here $m\epsilon[1,\infty)$ is a weighting parameter that determines thedegree of fuzziness, $V=[v_1,v_2,\ldots.v_c]$ where $v_i\epsilon\Re^n$ is a vector of (unknown) cluster prototypes (centers). The prototypes, the membership functions and the distance metric are calculated by the Eqs. (3)–(5) respectively.

$$v_i=\frac{\sum_{k=1}^{N}(\mu_{ik})^{m}z_k}{\sum_{k=1}^{N}(\mu_{ik})^{m}}$$

$$\mu_{ik}=\left(\sum_{j=1}^{c}\left(\frac{D_{ikA}}{D_{jkA}}\right)^{\frac{2}{m-1}}\right)^{-1}$$

$$D_{ikA}^2=\|z_k-v_i\|_A^2=(z_k-v_i)^TA(z_k-v_i),$$

$$where\ 1\le i\le c,\qquad 1\le k\le N$$

When the objective function converges to a local minimum the iteration terminates. Bezdek et al. [9] proposed the detailed algorithm. The algorithm comprises of the following basic steps:

Step1: Ramdomly initialize $U^{(0)}$, number of clusters c,weighting parameter m and the termination tolerance $\epsilon>0$

Step2: With $U^{(k)}$ determine the centroids vector $V=[v_1,v_2,\ldots.v_c]$ by using Eq.(4)

Step3: Update U$^{(k)}$, $U^{(k+1)}$ by using Ew.(5)

Step4: If$\left\|U^{(k+1)}-U^{(k)}\right\|<\epsilon$ then stop, else repeat from step2 by increasing the step value k

Although FCM is a popular clustering method it has some drawbacks. For example, it creates noise points when the method is applied to partition two clusters with an object having equidistance from two cluster's centers. FCM uses standard Euclidean distance norm.

3. Results and discussion:

3.1. Results

The algorithms were implemented in MATLAB version R2012a. To achieve good clustering results authors considered the maximum of 100 iterations and 15 independent test runs. The threshold value is e= 0:00001 and the weighting exponent in FCM is m = 1. The liver disorder data set contains 341 samples classified as two different classes. Each sample is characterized by 6 attributes and all the

samples are labeled by numbers 1 to 341. The samples from 1 to 142 are classified as class 1 and from 143 to 341 are classified as class 2. The algorithms KM, FCM are applied to generate two clusters. FCM generates two clusters corresponding to class 1 and class 2 containing 53 and 288 samples respectively. 36 samples that belong to class 2 are wrongly grouped into class 1 and 125 samples that belong to class 1 are wrongly grouped into class 2. The method KM generates two clusters containing 38 and 303 samples corresponding to class 1 and class 2 respectively.

The clustering results of the two fuzzy methods

The clustering results obtained by the algorithms k-means, FCM clusters for the liver disorder data set

	k-means		FCM(m=1)	
	Class1	class2	class1	class2
Correct	14	176	17	163
Incorrect	23	128	36	125
Total	37	304	53	288
Percentage of Correctness	9.85%	88.44%	11.97%	81.90%

Table 1 Comparison of performance of the clustering results obtained by the algorithms KM, FCM for the liver disorder

	percentage of correctness		classification performance
	Class 1	Class 2	
k-means (KM)	9.85%	88.44%	55.71%
Fuzzy c-Means (FCM)	11.97%	81.90%	52.78%

IJIAET International Journal of Innovative Analyses and Emerging Technology
| e-ISSN: 2792-4025 | http://openaccessjournals.eu | Volume: 1 Issue: 7

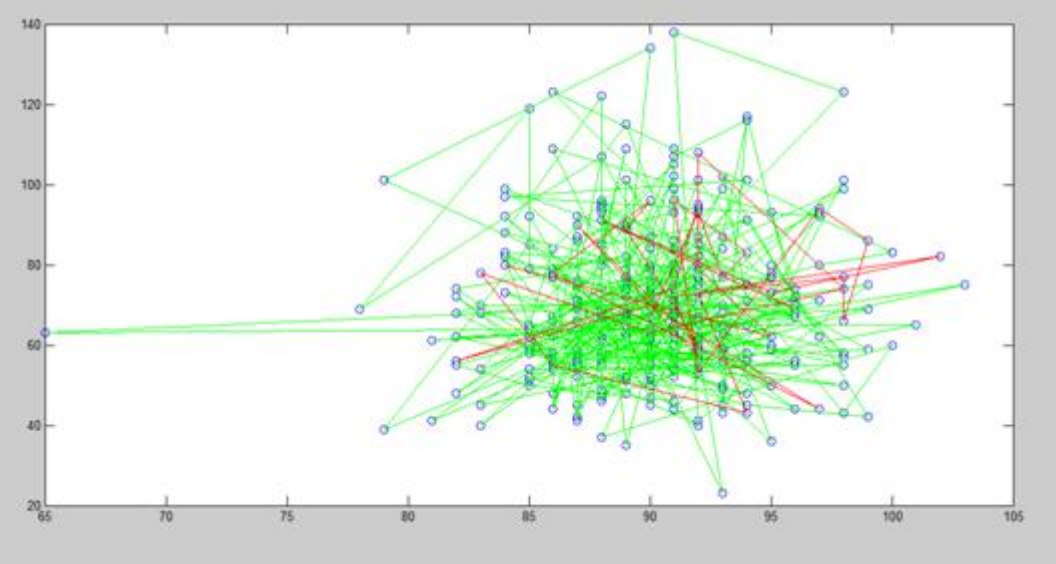

Figure 1: K-Means result

In the above graph shows the results of K-means clustering algorithm applied to liver data. In this Green colored line indicated class2, Red line indicated class 1

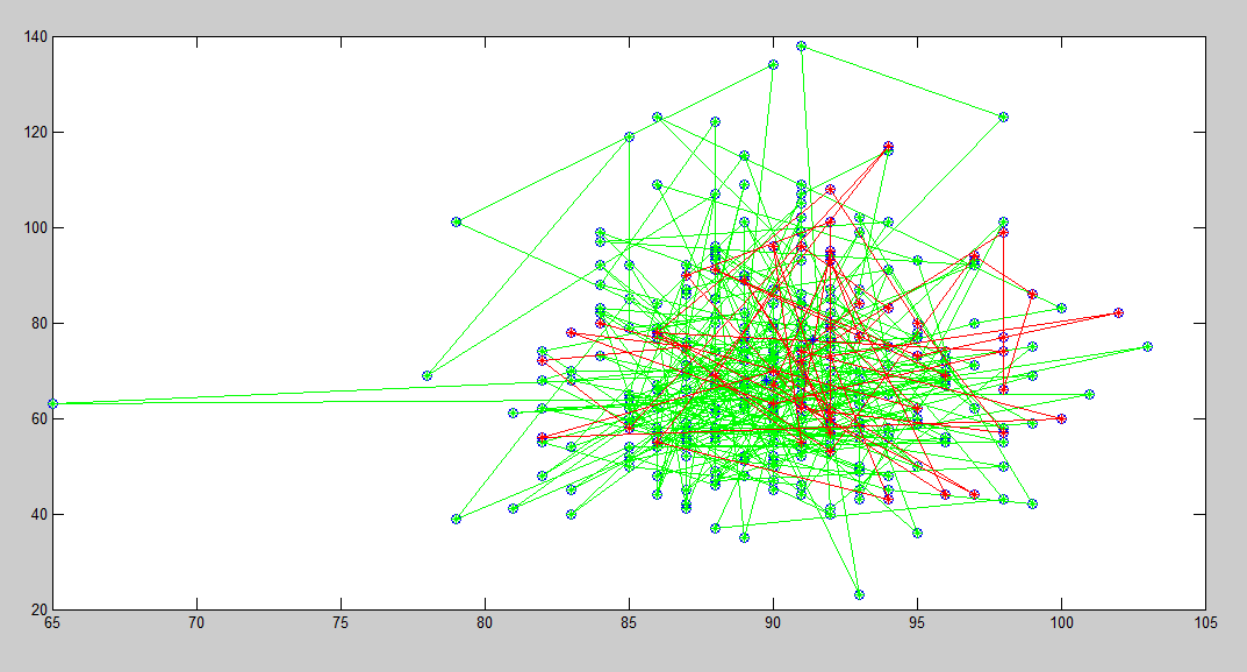

Figure 2: FCM result

In the above results green line indicated class 2 red line indicated class 1

4. Conclusion:

In this work, authors examined that classification production of various clustering method in medical diagnosis. The authors implemented the fuzzy clustering algorithm, Fuzzy c-Means (FCM), and k-means algorithms and discussed the results. The two algorithms performed well and the performance

and correctness were comparared. As well as in percentage of correctness and classification performance k-means is better than fuzzy c-means.

References:

1. J.C.Bezdek(1981): "pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York
2. Velmurugan T, Santhanam T. A survey of partition based clustering algorithms in data mining: an experimental approach. Inform Technol J 2011;10(3):478–84.
3. Yin J, Sun H, Yang J, Guo Q. Comparison of k-means and fuzzy c-means algorithm performance for automated determination of the arterial input function. PLoS ONE 2014;9(2):1–8. http://dx. doi.org/10.1371/journal.pone.0085884.
4. Velmurugun T. Performance comparison between k-means and fuzzy c-means algorithms using arbitrary data points. Wulfenia J 2012;9(8):234–41.
5. Wang XY, Garibaldi JM. A comparison of fuzzy and non-fuzzy clustering techniques in caner diagnosis. In: Proceedings of second international conference in computational intelligence in medicine and healthcare the biopattern conference, Costa da Caparica
6. Gosh Soumi, Dubey Sanjay Kumar. Comparative analysis of kmeans and fuzzy c means algorithms. Int J Adv Comput Sci Appl 2013;4(4):35–9.
7. Jipkate Bharati R, Gohokar VV. A comparative analysis of fuzzy c-means clustering and k means clustering algorithms. Int J Comput Eng Res 2012;2(3):737–9.
8. Forsyth Richard S. UCI machine learning repository [http:// archive.ics.uci.edu/ml]. Mapperley Park, Nottingham NG3 5DX, England; 1990.
9. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Berkeley, vol 1: statistics; 1967. p. 281–97.
10. Bezdek James C, Ehrlich Robert, Full William. FCM: the fuzzy c-means clustering algorithm. Comput Geosci 1984;10(2–3): 191–203.