DOI:10.21276/sjeat.2016.1.4.6

Saudi Journal of Engineering and Technology
Scholars Middle East Publishers
Dubai, United Arab Emirates
Website: http://scholarsmepub.com/

ISSN 2415-6272 (Print)
ISSN 2415-6264 (Online)

Research Article

Study on Liver Data using Clustering Algorithms

B. Venkataramana[1], L. Padmasree[2], M. Srinivasa Rao[3], D. Latha[4], G. Ganesan[5]

[1]Department of Computer Science & Engineering, Holy Mary Institute of Technology, Bogaram, Telengana, India
[2]Department of Electronics & Communications Engineering, VNR Vignana Jyothi Institute of Engineering & Technology, Bachupalli, Telengana, India
[3]School of Information Technology, Jawaharlal Nehru Technological University Kukatpally, Telangana, India
[4]Department of Computer Science, Adikavi Nannaya University, Rajahmundry, Andhra Pradesh, India
[5]Department of Mathematics, Adikavi Nannaya University, Rajahmundry, Andhra Pradesh, India

***Corresponding Author:**
G. Ganesan
Email: prof.ganesan@yahoo.com

Abstract: Data clustering has been considered as the most important raw data analysis method used in data mining technology. To extract the unknown valuable information from the large volume of data for so many real time applications are used in data classification. Most of the clustering techniques proved their efficiency in many applications such as decision making systems, medical sciences, earth sciences etc. Partition based clustering is one of the main approach in clustering. There are various algorithms of data clustering, every algorithm has its own advantages and disadvantages. This work reports the results of classification performance of three such widely used algorithms namely K-means (KM), Fuzzy c-means and Possibilistic Fuzzy c-Means (PFCM) clustering algorithms. To analyze these algorithms two known data sets from UCI machine learning repository are taken. From the repository the efficiency of clustering output is compared with the classification performance, percentage of correctness and no. of iterations taken to converge objective function. The experimental results prove that PFCM produces poor results compared to FCM and K-means algorithm yields more accurate results than the FCM and PFCM algorithms for liver data.

Keywords: Possibilistic fuzzy c-means, classification, K-means, fuzzy c-means.

INTRODUCTION

Data mining techniques play an important role to help the decision makers in making predictions that impact people and enterprises in the field of data analysis. In data analysis Clustering and Classification are the key elements. Clustering is the procedure of organizing the data into groups of similar objects called as clusters or classes. Many clustering algorithms which are having applications in different fields, such as medical sciences, image processing, earth science; decision making systems etc. have been developed by Researchers [1]. Among the different clustering algorithms partition based clustering algorithms have the more advantage of being accuracy in decision making by using suitable objective function based on similarity measures [2].

The objective function main task is to locate the cluster prototypes or centroids by optimizing the objective function. So that most identical objects with respect to the centroid create a cluster. K-means (KM) and Fuzzy c-Means (FCM) algorithms are widely used iterative algorithms in partition based clustering. Although FCM is a popular clustering algorithm it has some drawbacks such as creating noise points etc. To overcome the drawbacks occurring in FCM researchers developed different clustering algorithms. Among various clustering algorithms, Fuzzy Possibilistic c-Mean (FPCM) and Possibilistic Fuzzy c-Mean (PFCM) algorithms are popular. The membership function which assigns a number called membership value ranged between 0 and 1 to each object in the dataset is employed by Fuzzy cluster analysis. Many researchers analyzed the clustering performance of these techniques in their literature. The clustering performance of FCM, k-means and PFCM on medical diagnostics and reported that the efficiency of PFCM is better than FCM method evaluated by Simhachalam and Ganesan [3].

J.Quintanilla-Dominguez et al.[4] compared the advantages and drawbacks of KM , FCM and PFCM algorithms for detection of micro calcifications in image segmentation. Nidhi Grover [5] studied the advantages and drawbacks of FCM and PFCM algorithms. Rajendran and Dhanasekaran [6] analyzed FCM and PFCM methods on MRI brain image tissue segmentation and reported that the PFCM achieved better clustering results than FCM.

In this work, the authors aim to present the application of the three unsupervised clustering algorithms namely K-means (KM), Fuzzy C-means and Possibilistic Fuzzy c-Mean (PFCM) algorithms to popular real data set namely Liver disorder. The comparative analysis of performance of the three algorithms is presented in this work.

MATERIALS AND METHODS

In data analysis clustering is a discipline devoted to investigating and describing the clusters with similar objects. The efficiency and robustness of clustering algorithms could be investigated by clustering output. The performance of clustering algorithms can be improved by defining suitable objective function. The partition based clustering algorithms FCM and KM were developed by introducing memberships and distance measures in its objective functions respectively. The algorithms FPCM and PFCM were developed by implementing memberships and introducing typicalities to improve the performance of FCM. In this section the brief details of data sets, liver disorder and the algorithms KM, FCM and PFCM are presented.

The Dataset

To evaluate K-means (KM), Fuzzy c-means and Possibilistic Fuzzy c-Mean (PFCM) algorithms, the real world data sets Liver disorder data set donated by Richard [7] from the UCI Machine Learning Repository have been considered. Liver disorder data set contains 341 samples with 6 attributes each. These attributes are the measurements of the blood tests that are sensitive to liver disorders which might arise due to excessive alcohol consumption. These blood tests are mcv-mean corpuscular volume, alkphos-alkaline phosphotase, sgpt-alamine aminotransferase, sgot-aspartate aminotransferase, gammagt-gamma-glutamyl transpeptidase and drinks-the number of half-pint equivalents of alcoholic beverages drunk per day. The samples in the Liver disorder data set are classified into two different classes according to the liver disorders: 142 samples belong to class 1 and 199 samples belong to class 2.

Methods

K-means clustering

MacQueen [8] introduced the k-means algorithm in 1967. It is a partitioning algorithm. It takes the input parameter k, the number of clusters, and partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high but the inter-cluster similarity is low. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different results. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to recalculate k new centroids. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2 \qquad (1)$$

Where $\left\|x_i^{(j)} - c_j\right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center, c_j is an indicator of the distance of an n data points from their respective cluster centers.

The Algorithm is:

Step1: Select K points as initial centroids.

Step2: Repeat.

Step3: Form k clusters by assigning all points to the closest centroid.

Step4: Re-compute the centroid of each cluster.

Step5: Until the centroids do not change.

K-means algorithm is significantly sensitive to the initial randomly selected cluster centers. The algorithm can be run multiple times to reduce this effect. The K-Means is a simple algorithm that has been adapted to many problem domains and it is a good candidate to work for a randomly generated data Repeat 2 and 3 until no change in each cluster center

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. Consider a dataset Z with N observations is an n-dimensional row vector. $z_{k=}[z_{k1}, z_{k2}, \dots\dots, z_{kn}] \in \Re^n$. The dataset Z is represented as N x n matrices. In medical diagnostics the rows of Z represents patients and the columns are symptoms or laboratory measurements for these patients. The partition of the dataset Z into c $(1 \leq c \leq N)$ clusters is represented by the fuzzy partition matrix $U = [\mu_{ik}]_{cxN}$. The fuzzy partitioning space for Zistheset

$$M_{fc} = \left\{ U\epsilon\Re^{cxN} \big/ \mu_{ik}\epsilon[0,1], \forall i,k; \sum_{i=1}^{c}\mu_{ik} = 1, \forall k; 0 < \sum_{k=1}^{N}\mu_{ik}, \forall i \right\} \qquad (2)$$

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

$$\underbrace{min}_{U,V}\{J(Z;U,V) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^m\|z_k - v_i\|^2{}_A\} \text{ Where } U = [\mu_{ik}]\epsilon M_k \quad (3)$$

Here $m\epsilon[1,\infty)$ is a parameter that determines the degree of fuzziness,$V = [v_1, v_2, \dots . v_c]$ where $v_i \epsilon \Re^n$ is a vector of (unknown) cluster prototypes (centers). The prototypes, the membership functions and the Euclidian distance metric are calculated by the equations (4),(5),(6) respectively.

$$v_i = \frac{\sum_{k=1}^{N}(\mu_{ik})^m z_k}{\sum_{k=1}^{N}(\mu_{ik})^m}, 1 \le i \le c \quad (4)$$

$$\mu_{ik} = \left(\sum_{j=1}^{c}\left(\frac{D_{ikA}}{D_{jkA}}\right)^{\frac{2}{m-1}}\right)^{-1}, \quad 1 \le i \le c, \ 1 \le k \le N \quad (5)$$

$$D_{ikA}^2 = \|z_k - v_i\|_A^2 = (z_k - v_i)^T A(z_k - v_i), \ 1 \le i \le c, \ 1 \le k \le N \quad (6)$$

When the objective function converges to a local minimum the iteration terminates. Detailed algorithm was proposed [9] is given below.

The algorithm is given by the following basic steps:
Step 1: Randomly initialize partition matrix U, number of clusters c, weighting parameter m and the termination tolerance ε > 0.

Step 2: Determine the fuzzy cluster prototypes by using the equation (4).

Step 3: update the membership matrix by using the equation (5).

Step 4: Compare the membership matrices of previous and after the iteration and repeat from step 2 until it meets the convergence criteria.

In fuzzy clustering, FCM is a popular clustering method but it has also some drawbacks. For example, if the method is used to partition two clusters and there is an object which is equidistance from two centers then according to the constraint on the membership value it assigns equal membership value regardless of the actual belonging to a cluster. These points are called as noise points.

Possibilistic Fuzzy c-Mean Clustering

In order to achieve good clustering results the memberships and typicalities are both important. Nikhil et al. [10] proposed Possibilistic Fuzzy c-Mean (PFCM) model. In this proposed model the constraint in the FPCM model that the sum of the typicalities of all data points in a cluster is equal to 1 is relaxed and retains the constraint on memberships. Possibilistic Fuzzy c-Mean model achieves its partitioning by optimizing its iterative objective function defined as

$$\underbrace{min}_{U,T,Y}\{J(Z;U,T,V) = \sum_{i=1}^{c}\sum_{k=1}^{N}(au_{ik}^m + bt_{ik}^n)\times \|z_k - v_i\|^2{}_A + \sum_{i=1}^{c}\gamma_i \sum_{i=1}^{c}(1-t_k)^\eta\} \quad (7)$$

$0 \le \mu_{ik}, t_{ik} \le 1, m, \eta > 1, a, b > 0 \ and \ \gamma_i \ > 0$. The typicality matrix and the prototypes are calculated by the equation (8) and equation (9) respectively.

$$t_{ik} = \left(1 + \left(\frac{b}{\gamma_i}D_{ika}^2\right)^{\frac{1}{\eta-1}}\right)^{-1}, \quad 1 \le i \le c, \ 1 \le k \le N \ (8)$$

$$v_i = \frac{\sum_{k=1}^{N}(au_{ik}^m + bt_{ik}^n)^m z_k}{\sum_{k=1}^{N}(au_{ik}^m + bt_{ik}^n)^m}, 1 \le i \le c \quad (9)$$

The basic steps of the PFCM algorithm are described as follows:
Step 1: Initialization: Randomly initialize partition matrix U, and typicality matrix , number of clusters c, parameters m, η, a, b and the termination tolerance ε > 0

Step 2: Centroid calculation: Calculate the fuzzy cluster prototypes by using the equation (9).

Step 3: Classification: Update the membership matrix by using the equation (5) and the typicality matrix by using the equation (8).

Step 4: Convergence criteria: Compare the membership matrices of previous and after the iteration. If the comparison value is less than the termination tolerance, then stop, else repeat from step 2.

This model has the potential that is either it can influence the prototypes by means of memberships (when a >b) or by typicalities (when b > a). If the values of a and b are restricted as a =1 and b =0 then the PFCM model performs as FCM model. The effect of outliers can be reduce by considering high value of b (m) than a (η).

RESULTS AND DISCUSSION

The algorithms were implemented in MATLAB version R2012a. To achieve good clustering results authors considered the maximum of 100 iterations. The threshold value is e= 0.00001 and the weighting exponent in FCM is m = 2 and for PFCM m=1.5. The liver disorder data set contains 341 samples classified into 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. 23 samples that belongs to class 2 are wrongly grouped into class 1 and 128 samples that belongs to class 1 are wrongly belongs to class 2.

The PFCM generated two clusters that contain 110 and 231 samples corresponding to class 1 and class 2 respectively. 76 samples that belongs to class 2 are wrongly grouped into class 1 and 123 samples that belongs to class 1 are wrongly grouped into class 2.

The clustering results of the three fuzzy methods. The clustering results obtained by the algorithms k-means, FCM and PFCM clusters for the liver disorder data set.

Table 1: comparisons of performance of clustering results

	k-means		FCM(m=2)		PFCM(m=1.5)	
	Class1	class2	class1	class2	Class1	Class2
Correct	14	176	17	163	34	123
Incorrect	23	128	36	125	76	108
Total	37	304	53	288	110	231
Percentage of Correctness	9.85%	88.44%	11.97%	81.90%	23.94%	61.81%

Table 2: Comparisons of classification performance and percentage of correctness performance

Clustering method	Liver data set (2 clusters)		
	Percentage of correctness		Classification performance
	Class 1	Class 1	
K-means (KM)	9.85%	88.44%	55.71%
Fuzzy c-Means (FCM)	11.97%	81.90%	52.78%
PFCM	23.94%	61.81%	46.04%

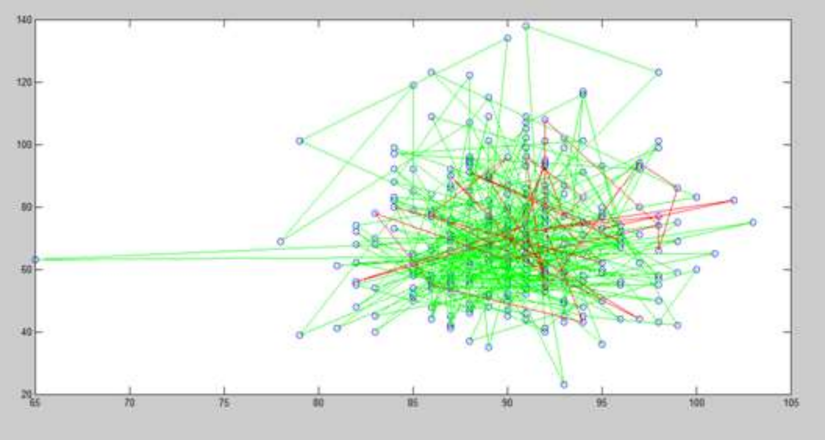

Fig-1: K-Means result

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

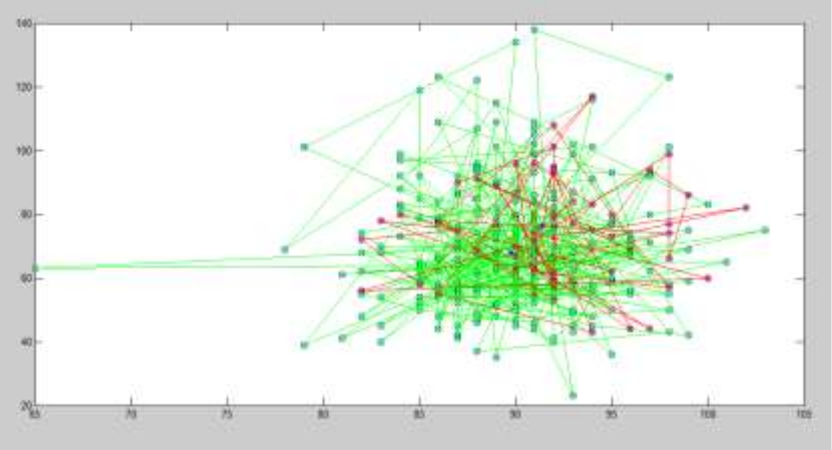

Fig-2: FCM result

In the above results, green line indicated class 2 and red line indicated class 1.

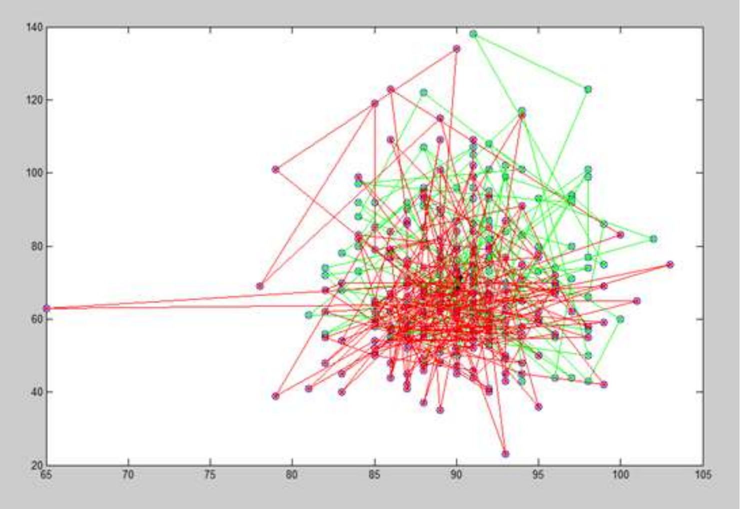

Fig-2: PFCM result

In the above result, green line indicated class 1 red line indicated class 2.

CONCLUSION

In this work, authors examined that classification production of various clustering method in medical diagnostics. The authors implemented the fuzzy clustering algorithms Fuzzy c-Means (FCM) and PFCM and a non-fuzzy clustering algorithm k-means algorithms and discussed the results. Comparing to the FCM and PFCM, FCM giving better results than PFCM. Among all the methods K-means is the best one as well as in percentage of correctness and classification performance.

REFERENCES

1. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264-323.
2. Velmurugan, T., & Santhanam, T. (2011). A comparative analysis between k-medoids and fuzzy c-means clustering algorithms for statistically distributed data points. *Journal of Theoretical & Applied Information Technology*, *27*(1).
3. Simhachalam, B., & Ganesan, G. (2015). Performance comparison of fuzzy and non-fuzzy classification methods. *Egyptian Informatics Journal*.
4. Quintanilla-Dominguez, J., Ojeda-Magaña, B., Cortina-Januchs, M. G., Ruelas, R., Vega-Corona, A., & Andina, D. (2011). Image segmentation by fuzzy and possibilistic clustering algorithms for the identification of microcalcifications. *Scientia Iranica*, *18*(3), 580-589.
5. Grover, N. (2014). A study of various Fuzzy Clustering Algorithms. *International Journal of Engineering Research (IJER)*, *3*(3), 177-181.
6. Rajendran, A., & Dhanasekaran, R. (2011). MRI brain image tissue segmentation analysis using

possibilistic fuzzy C-means method. *International Journal on Computer Science and Engineering*, *3*(12), 3832-3836.

7. Richard, S., & Forsyth. (1990). UCI Machine Learning Repository. Mapperley Park, Nottingham NG3 5DX, England. From http://archive.ics.uci.edu/ml
8. MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, *1*(14), 281-297.
9. Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2-3), 191-203.
10. Pal, N. R., Pal, K., Keller, J. M., & Bezdek, J. C. (2005). A possibilistic fuzzy c-means clustering algorithm. *IEEE transactions on fuzzy systems*, *13*(4), 517-530.