Implementation of Clustering Algorithms for real datasets in Medical Diagnostics using MATLAB

B. Venkataramana¹, L. Padmasree², M. Srinivasa Rao³, G. Ganesan⁴*, K. Rama Krishna⁵

(1) Department of Computer Science & Engineering, Holy Mary Institute of Technology, Bogaram, Kesara, Medcharla, Telengana, India
(2) Department of Electronics & Communications Engineering, VNR Vignana Jyothi Institute of Engineering & Technology, Bachupalli, Nizampet, Telengana, India
(3) School of Information Technology, Jawaharlal Nehru Technological University Hyderabad, Telangana, India
(4) Department of Mathematics, Adikavi Nannaya University, Rajahmundry, Andhra Pradesh, India
(5) Department of Electrical & Electronics Engineering, Holy Mary Institute of Technology, Bogaram, Kesara, Telengana, India

Abstract
As in the medical field, for one disease there require samples given by diagnosis. The samples will be analyzed by a doctor or a pharmacist. As the no. of patients increases their samples also increases, there require more time to analyze samples for deciding the stage of the disease. To analyze the sample every time requires a skilled person. The samples can be classified by applying them to clustering algorithms. Data clustering has been considered as the most important raw data analysis method used in data mining technology. 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 Fuzzy Possibilistic c-Means (FPCM) clustering algorithms. To analyze these algorithms three known data sets from UCI machine learning repository are taken such as thyroid data, liver and wine. The efficiency of clustering output is compared with the classification performance, percentage of correctness. The experimental results show that K-means and FCM give same performance for liver data. And FCM and FPCM are giving same performance for thyroid and wine data. FPCM has more efficient classification performance in all the given data sets.

**Keywords:** Classification, K-means, Fuzzy c-means, Fuzzy Possibilistic c-means.

1 Introduction

Data mining techniques plays 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 or Classification is the key element. Clustering is an organization of 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]. For one disease there require samples given by diagnosis in the medical field. The doctor or pharmacist will be analyzed the stage of disease by these samples. As the no. of patients increases their samples also increases, there require more time to analyze samples for deciding the stage of the disease. An assistance of Technical person is needed to analyze the sample for every patient. The samples can be classified by applying them to clustering algorithms.

The main task in clustering algorithm 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, FPCM 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, K-means (KM), Fuzzy C-means and Fuzzy Possibilistic c-Means (FPCM) algorithms to popular real data set namely Liver disorder, thyroid and wine data. The comparative analysis of performance of the three algorithms is presented in this paper.
2 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 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, thyroid and wine data and the algorithms KM, FCM and FPCM are presented.

2.1. The Dataset

To evaluate K-means (KM), Fuzzy c-means, Fuzzy Possibilistic c-Means (FPCM) algorithms, the real world data sets Liver disorder data set donated by Richard [7] and Wine data set donated by Forina et al. [8] and Thyroid data set donated by from the UCI Machine Learning Repository have been considered.

The dataset contains 215 samples with 5 attributes or lab measurements each. The samples are classified into three different classes according to the Thyroid functions: Normal (150 samples), Hyperthyroid (35 samples) and Hypothyroid (30 samples). The 5 attributes are the lab tests to measure the thyroid function. These attributes are T3-resin uptake test (A percentage), value of total serum thyroxin given by the isotopic displacement method, total serum tri-iodothyronine value given by radioimmunoassay, value of basal thyroid stimulating hormone (TSH) given by radioimmunoassay and after injection of 200 micro grams of thyrotrophic-releasing hormone the maximal absolute difference of TSH value as compared to the basal value.

The real world data sets Liver Disorder is obtained from the UCI Machine Learning Repository donated by Richard [7]. 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.

The Wine data set contains 178 samples and each sample has 13 attributes, they are chemical analysis of the wine derived from three different cultivars but grown in the same region in Italy. The samples are grouped into three different classes according to the cultivars: Cultivar 1 containing 59 samples, Cultivar 2 containing 71 samples and Cultivar 3 containing 48 samples. The attributes are the values of chemical analysis of Alcohol, Malic acid, Ash, Alkalinity of ash, Magnesium, Total phenols, Flavonoids, Nonflavonoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines and Proline.

2.2. Methods

2.2.1. K-means clustering

MacQueen [9] 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

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

(2.1)

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:

Step 1: Select K points as initial centroids.
Step 2: Repeat.
Step 3: Form k clusters by assigning all points to the closest centroid.
Step 4: Re-compute the centroid of each cluster.
Step 5: 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.

2.2.2. 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. \( z_k = [z_{k1}, z_{k2}, ..., z_{kn}] \in \mathbb{R}^n \). 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 ≤ c ≤ N) clusters is represented by the fuzzy partition matrix \( U = [\mu_{ik}]_{c \times N} \). The fuzzy partitioning space for \( Z \) is the set

\[ M_f = \{ U \in \mathbb{R}^{c \times N} / \mu_{ik} \epsilon [0,1], \forall i, k; \sum_{i=1}^{c} \mu_{ik} = 1, \forall k; 0 < \sum_{k=1}^{N} \mu_{ik}, \forall i \} \]  

(2.2)

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

\[ \min_{U,V} \{ J(Z; U, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m \| z_k - v_i \|^2 \} \]  

Where \( U = [\mu_{ik}] \epsilon M_k \)  

(2.3)

Here \( m \epsilon [1, \infty) \) is a parameter that determines the degree of fuzziness, \( V = [v_1, v_2, ..., v_c] \) where \( v_i \epsilon \mathbb{R}^n \) is a vector of (unknown) cluster prototypes (centers). The prototypes, the membership functions and the Euclidian distance metric are calculated by the equations (2.4), (2.5), (2.6) respectively.

\[ v_i = \frac{\sum_{k=1}^{N} (\mu_{ik})^m z_k}{\sum_{k=1}^{N} (\mu_{ik})^m}, 1 \leq i \leq c \]  

(2.4)

\[ \mu_{ik} = \left( \sum_{j=1}^{c} \left( \frac{D_{jkA}}{D_{kkA}} \right)^{\frac{2}{m-1}} \right)^{-1}, 1 \leq i \leq c, 1 \leq k \leq N \]  

(2.5)

\[ D_{kkA}^2 = \| z_k - v_i \|_A^2 = (z_k - v_i)^T A (z_k - v_i), 1 \leq i \leq c, 1 \leq k \leq N \]  

(2.6)
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 \( \varepsilon > 0 \).

Step 2: Determine the fuzzy cluster prototypes by using the equation (2.4).

Step 3: Update the membership matrix by using the equation (2.5).

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.

2.2.3. Fuzzy Possibilistic c-mean algorithm

Traditional clustering approaches the partition whereby each object can only belong to one cluster at any one time. Fuzzy clustering extends this notion to each object can belong to more than one cluster at a time with different membership values using a membership function. These membership values ranged from 0 to 1. FPCM was developed based on fuzzy theory by Pal and Bedzek \[11\]. The concept of typicality and membership functions was introduced in FPCM model to overcome the drawbacks occurring in FCM model proposed by Bezdek et al. \[12\]. The partition of the dataset \( Z \) into \( c \) clusters is represented by the fuzzy partition matrix \( N \). The fuzzy partitioning space for \( Z \) is the set

\[
M_c = \left\{ U \in \mathbb{R}^{c \times N} / \mu_{i,k} \in [0,1], \forall i,k; \sum_{i=1}^{c} \mu_{i,k} = 1, \forall k; 0 < \sum_{k=1}^{N} \mu_{i,k}, \forall i \right\}
\]  

Fuzzy Possibilistic c-Mean model achieves its partitioning by optimizing its iterative objective function defined as where

\[
\min_{U,T,Y} \left\{ J_{m,\eta}(Z; U, T, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} \left( \mu_{i,k}^{m} + t_{i,k}^{\frac{1}{\eta}} \right) \times \| z_k - v_i \|_{A}^{2} \right\}
\]  

\( T \in \left\{ U \in \mathbb{R}^{c \times N} / \mu_{i,k} \in [0,1], \forall i,k; \forall k \exists i; 0 < \sum_{k=1}^{N} \mu_{i,k} < N, \forall i \exists \mu_{i,k} > 0 \right\}

Subject to the constraints \( 0 \leq \mu_{i,k}, t_{i,k} \leq 1, m > 1, \eta > 1 \) and \( \forall i, t_{i,k} \in T \) such that \( \sum_{k=1}^{N} t_{i,k} = 1 \)

Here \( V = [v_1, v_2, ..., v_c] \) where \( v_i \in \mathbb{R}^n \) denotes a vector of (unknown) cluster prototypes (centers) and the degree of fuzziness determined by a weighting parameter,

\[
v_i = \frac{\sum_{k=1}^{N} \left( \mu_{i,k}^{m} + t_{i,k}^{\frac{1}{\eta}} \right) z_k}{\sum_{k=1}^{N} \left( \mu_{i,k}^{m} + t_{i,k}^{\frac{1}{\eta}} \right)}, 1 \leq i \leq c
\]  

\[
\mu_{i,k} = \left( \sum_{j=1}^{c} \left( \frac{D_{i,kA}}{D_{j,kA}} \right)^{\frac{2}{m-1}} \right)^{-1}, 1 \leq i \leq c, 1 \leq k \leq N
\]  

\[
t_{i,k} = \left( \sum_{j=1}^{N} \left( \frac{D_{i,kA}}{D_{j,kA}} \right)^{\frac{2}{\eta-2}} \right)^{-1}, 1 \leq i \leq c, 1 \leq k \leq N
\]

\[
D_{i,kA}^2 = \| z_k - v_i \|_{A}^2 = (z_k - v_i)^T A (z_k - v_i), 1 \leq i \leq c, 1 \leq k \leq N
\]

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

Step 2: Centroid calculation: Determine the fuzzy cluster prototypes by using the equation (2.9).

Step 3: Classification: update the membership matrix by using the equation (2.10) and the typically matrix by using the equation (2.11).

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

### 3 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 $\varepsilon = 0.00001$ and the weighting exponent in FCM is $m = 1.1$ and for FPCM $m=1.1$.

#### 3.1. Liver Data

The clustering results of three methods are listed below. The clustering results obtained by the algorithms K-means, FCM and FPCM clusters for the liver disorder data set.

<table>
<thead>
<tr>
<th></th>
<th>k-means</th>
<th>FCM($m=1.1$)</th>
<th>FPCM($m=1.1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class1</td>
<td>Class2</td>
<td>Class1</td>
</tr>
<tr>
<td>Correct</td>
<td>14</td>
<td>176</td>
<td>14</td>
</tr>
<tr>
<td>Incorrect</td>
<td>23</td>
<td>128</td>
<td>23</td>
</tr>
<tr>
<td>Total</td>
<td>37</td>
<td>304</td>
<td>37</td>
</tr>
<tr>
<td>Percentage of Correctness</td>
<td>9.85%</td>
<td>88.44%</td>
<td>9.85%</td>
</tr>
</tbody>
</table>

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 and FPCM are applied to generate two clusters.

The method KM generates two clusters containing 37 and 304 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.

FCM generates two clusters corresponding to class 1 and class 2 containing 37 and 304 samples respectively. 23 samples that belong to class 2 are wrongly grouped into class 1 and 128 samples that belong to class 1 are wrongly grouped into class 2.

The FPCM generated two clusters that contain 63 and 278 samples corresponding to class 1 and class 2 respectively. 19 samples that belongs to class 2 are wrongly grouped into class 1 and 115 samples that belongs to class 1 are wrongly grouped into class 2.
3.1.1. Results

**Figure 1: K-Means result for liver data**

In the Fig 1 shows the results of K-means clustering algorithm applied to liver data. In this Green colored line indicated class 2, Red line indicated class 1.

**Figure 2: FCM result for liver data**

In the Fig 2 red line indicated class 2 green line indicated class 1.
In the Fig 3 shows the results FPCM clustering algorithm applied to liver data. In this red colored line indicated class 2, green line indicated class 1.

3.2. Thyroid Data
The clustering results obtained by the algorithms K-means, FCM and FPCM clusters for the thyroid data set

Table 2: Comparisons of performance of clustering results

<table>
<thead>
<tr>
<th></th>
<th>k-means</th>
<th>FCM(m=1.1)</th>
<th>FPCM(m=1.1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal</td>
<td>Hyperthyroid</td>
<td>Hypothyroid</td>
</tr>
<tr>
<td>correct</td>
<td>137</td>
<td>23.</td>
<td>24</td>
</tr>
<tr>
<td>incorrect</td>
<td>15</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>152</td>
<td>29</td>
<td>34</td>
</tr>
<tr>
<td>% of correctness</td>
<td>91.33</td>
<td>65.71</td>
<td>80.00</td>
</tr>
</tbody>
</table>

K-means generates three clusters corresponding to Normal, Hyperthyroid and Hypothyroid containing 152, 29 and 34 samples respectively. The cluster which is associated with Normal contains 9 samples that belong to Hyperthyroid and 6 samples that belong to hypothyroid clusters are wrongly grouped. The cluster which is associated hyperthyroid contains 6 samples that belong to Normal clusters are wrongly assigned, and 7 samples that belong to Normal and 3 samples that belong to Hyperthyroid clusters are wrongly classified in to the cluster associated with Hypothyroid.

FCM generates three clusters corresponding to Normal, Hyperthyroid and Hypothyroid containing 153, 39 and 23 samples respectively. The cluster which is associated with Normal contains 10 samples that belong to Hyperthyroid and 6 samples that belong to hypothyroid clusters are wrongly grouped. The cluster which is associated hyperthyroid contains 13 samples that belong to Normal and 1 sample that belongs to Hypothyroid clusters are wrongly assigned with Hyperthyroid, and no such other sample that belong to Normal and hyperthyroid clusters are wrongly classified in to the cluster associated with Hypothyroid.

FPCM generates three clusters corresponding to Normal, Hyperthyroid and Hypothyroid containing 153, 39 and 23 samples respectively. The cluster which is associated with Normal contains 10 samples that
belong to Hyperthyroid and 6 samples that belong to hypothyroid clusters are wrongly grouped. The cluster which is associated hyperthyroid contains 13 samples that belong to Normal and 1 sample that belongs to Hypothyroid clusters are wrongly assigned with Hyperthyroid, and no such other sample that belong to Normal and hyperthyroid clusters are wrongly classified in to the cluster associated with Hypothyroid.

3.2.1. Results

![Figure 4: K-Means result for Thyroid](image)

In the Fig 4 shows the results of K-means clustering algorithm applied to thyroid data. In this Green colored line indicated normal, blue line indicated hyperthyroid and red line indicated hypothyroid.

![Figure 5: FCM result for Thyroid data](image)
In the Fig 5 blue line connects normal, red line connects hypothyroid and green line connects hyperthyroid samples.

![Figure 6: FPCM result for Thyroid](image)

In the Fig 6 blue line connects Normal, red line connects hyperthyroid and green line connects hypothyroid samples.

### 3.3. Wine Data

The clustering results obtained by the algorithms K-means, FCM and FPCM clusters for the wine data set.

<table>
<thead>
<tr>
<th></th>
<th>k-means</th>
<th>FCM(m=1.1)</th>
<th>FPCM(m=1.1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cultivar1</td>
<td>Cultivar2</td>
<td>Cultivar3</td>
</tr>
<tr>
<td>Correct</td>
<td>30 62 11</td>
<td>46 50 29</td>
<td>46 50 29</td>
</tr>
<tr>
<td>Incorrect</td>
<td>29 9 37</td>
<td>1 19 33</td>
<td>1 19 33</td>
</tr>
<tr>
<td>Total</td>
<td>59 71 48</td>
<td>47 69 62</td>
<td>47 69 62</td>
</tr>
<tr>
<td>% of Correctness</td>
<td>50.84 87.32 22.91</td>
<td>77.96 70.42 60.41</td>
<td>77.96 70.42 60.41</td>
</tr>
</tbody>
</table>

K-means generates three clusters corresponding to cultivar1, cultivar2 and cultivar3 containing 59, 71 and 48 samples respectively. The cluster which is associated with cultivar1 contains 1 sample that belongs to cultivar2 and 28 samples that belong to cultivar3 are wrongly grouped. The cluster which is associated with cultivar2 contains 9 samples that belong to cultivar3 are wrongly assigned, and 37 samples that belong to cultivar2 are wrongly classified in to the cluster associated with cultivar3.

FCM generates three clusters corresponding to cultivar1, cultivar2 and cultivar3 containing 47, 69 and 62 samples respectively. The cluster which is associated with cultivar1 contains 1 sample that belongs to cultivar2 is wrongly grouped. The cluster which is associated cultivar2 contains 19 samples that belong to cultivar3 are wrongly grouped, and 13 samples that belong to cultivar1 and 20 samples that belong to cultivar2 are wrongly classified in to the cluster associated with Cultivar3.

FPCM generates three clusters corresponding to cultivar1, cultivar2 and cultivar3 containing 47, 69 and 62 samples respectively. The cluster which is associated with cultivar1 contains 1 sample that belongs to
cultivar2 is wrongly grouped. The cluster which is associated cultivar2 contains 19 samples that belong to cultivar3 are wrongly grouped, and 13 samples that belong to cultivar1 and 20 samples that belong to cultivar2 are wrongly classified in to the cluster associated with Cultivar3.

### 3.3.1. Results

![Figure 7: K-Means result for wine data](image)

In the Fig 7 blue line connects cultivar2, green line cultivar1 normal and red line cultivar3.

![Figure 8: FCM result for wine data](image)

In the Fig 8 red line connects cultivar3, green line connects cultivar2 and blue line connects cultivar1 samples.
In the Fig 9 green line connects cultivar1, red line connects cultivar2 and blue line connects cultivar3 samples.

4 Comparison of percentage of correctness and classification performance of three methods

Here, we provide the comparative analysis of correctness and classification performance of three methods on three data sets.

<table>
<thead>
<tr>
<th>Clustering method</th>
<th>Liver data set (2 clusters)</th>
<th>Thyroid data set (3 clusters)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percentage of correctness</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Class 1</td>
<td>Class2</td>
</tr>
<tr>
<td>K-means (KM)</td>
<td>9.85%</td>
<td>88.44%</td>
</tr>
<tr>
<td>Fuzzy c-Means (FCM)</td>
<td>9.85%</td>
<td>88.44%</td>
</tr>
<tr>
<td>FPCM</td>
<td>30.98%</td>
<td>86.93%</td>
</tr>
</tbody>
</table>

Table 4: Comparisons of classification performance and percentage of correctness performance

Table 5: Comparisons of classification performance and percentage of correctness performance
Table 6: Comparisons of classification performance and percentage of correctness performance

<table>
<thead>
<tr>
<th>Clustering method</th>
<th>Wine data set (3 clusters)</th>
<th>Percentage of correctness</th>
<th>classification performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Class 1</td>
<td>Class 2</td>
</tr>
<tr>
<td>K-means (KM)</td>
<td>50.84%</td>
<td>87.32%</td>
<td>22.91%</td>
</tr>
<tr>
<td>Fuzzy c-Means (FCM)</td>
<td>77.96%</td>
<td>70.42%</td>
<td>60.41%</td>
</tr>
<tr>
<td>FPCM</td>
<td>77.96%</td>
<td>70.42%</td>
<td>60.41%</td>
</tr>
</tbody>
</table>

4 Conclusion

In this work, authors examined those classification algorithms of various clustering methods in medical diagnostics. The authors implemented the fuzzy clustering algorithms such as Fuzzy c-Means (FCM) and FPCM and a non-fuzzy clustering algorithm K-means algorithms are applied to real time data sets such as liver, thyroid and wine data and discussed the results. Among all clustering algorithms FPCM is performing better than K-Means and FCM. For liver data, FPCM has a highest classification performance of 63.63% and K-Means and FCM has similar classification performance of 55.71%. For thyroid data, FCM and FPCM give similar and highest classification performance of 86.06%. For wine data also, FCM and FPCM give similar and highest classification performance of 70.22%. Among all the methods the best one is FPCM because it has high percentage of correctness and classification performance.

References

   https://doi.org/10.1145/331499.331504

   https://doi.org/10.3923/itj.2011.478.484


   https://doi.org/10.1016/j.scient.2011.04.009

   https://doi.org/10.17950/ijer/v3i3/310

http://archive.ics.uci.edu/ml

http://archive.ics.uci.edu/ml


https://doi.org/10.1109/TFUZZ.2004.840099

https://doi.org/10.1109/FUZZY.1997.616338

https://doi.org/10.1016/0098-3004(84)90020-7