A Comparative Study of Hard and Fuzzy Data Clustering Algorithms with Cluster Validity Indices

O. A. Mohamed Jafar\textsuperscript{1,}\textsuperscript{*} and R. Sivakumar\textsuperscript{2}

\textsuperscript{1}Research Scholar, Department of Computer Science, A.V.V.M. Sri Pushpam College (Autonomous), Poondi 613 503, Thanjavur, Tamil Nadu, India.

\textsuperscript{2}Associate Professor, Department of Computer Science, A.V.V.M. Sri Pushpam College (Autonomous), Poondi 613 503, Thanjavur, Tamil Nadu, India.

\textit{e-mail: mohamedjafaroa@gmail.com}

Abstract. Data clustering is one of the important data mining methods. It is a process of finding classes of a data set with most similarity in the same class and most dissimilarity between different classes. The well known hard clustering algorithm ($K$-means) and Fuzzy clustering algorithm (FCM) are mostly based on Euclidean distance measure. In this paper, a comparative study of these algorithms with different distance measures such as Chebyshev and Chi-square is proposed. The new algorithms are tested on the four well known data sets such as Contraceptive Method Choice (CMC), Diabetes, Liver Disorders and Statlog (Heart) from the UCI repository. Experimental results show that FCM based on Chi-square distance measure gives better result than Chebyshev distance measure. We also propose the FCM algorithm based on $\sigma$-distance measure. The FCM algorithm is also tested with cluster validity indices such as partition coefficient and partition entropy. The results show that Chebyshev distance measure is reported maximum partition coefficient and minimum partition entropy than the other distance measures. This paper also provides a brief review of applications of $K$-means and Fuzzy $c$-means algorithms.

Keywords: Data clustering, $K$-means, Fuzzy $c$-means, Distance measures, Cluster validity.

1. Introduction

Most enterprises generate unlimited amounts of data and store them in databases. This large volume of data contains valuable hidden knowledge. The techniques that allow the enterprise to extract the most valuable information from databases are known as knowledge discovery in databases (KDD) or Data mining. Data mining is the process of extracting previously unknown, valid and useful information from large databases. It is a collection of techniques and tools for handling large amount of information [1].

Data clustering is a popular technique in data analysis. It is a method for finding classes or groups of a data set with most similarities in the same class and most dissimilarities between different classes. Clustering techniques have been widely used in many areas such as data mining, artificial intelligence, pattern recognition, bioinformatics, segmentation and machine learning [2].

Many clustering algorithms have been proposed by researchers. Partitioning clustering and hierarchical clustering are two main approaches to clustering. Some of the clustering algorithms in the literature are $K$-means, $K$-medoid, FCM, PAM, CLARA, CLARANS, BIRCH, CURE, ROCK and CHAELON. Among the above mentioned clustering techniques, $K$-means and FCM algorithms are widely used partitioning techniques by the researchers in many real-world applications.

$K$-means [3] is the most popular classical hard clustering technique. Each data point is from only one cluster. It requires the previous knowledge about the number of clusters. This method is not suitable for real world data sets in which there are no definite boundaries between the clusters.

FCM algorithm is one of the most important fuzzy clustering methods, initially proposed by Dunn [4], and then generalized by Bezdek [5]. FCM algorithm is a technique of clustering which permits one piece of data to belong to

© Elsevier Publications 2013.
two or more clusters. The aim of the FCM algorithm is the assignment of data points into clusters with varying degrees of membership values. Membership values lie between 0 and 1. This membership value reflects the degree to which the point is more representative of one cluster than the other.

The well known K-means and FCM algorithms are mostly based on Euclidean distance measure. In this paper, we have compared these algorithms with different distance measures such as Chebyshev, Chi-square and $\sigma$-distance measures. The well known real world data sets such as CMC, Diabetes, Liver Disorders, and Statlog (Heart) from UCI repository are applied to test the performance of K-means and FCM algorithms with cluster validity indices.

The rest of this paper is organized as follows: In section 2, clustering algorithms are presented. Section 3 describes the different distance measures and cluster validity indices. The methodology is given in section 4. A brief review of applications of K-means and FCM is provided in section 5. Experimental results are discussed in section 6. Finally, conclusions are presented in section 7.

2. Clustering Algorithms

Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects. The main objective of clustering is to group the data set into clusters such that similar objects are placed in the same cluster while dissimilar objects are placed in different clusters. The desired features of clustering techniques are scalability, ability to stop and resume, robustness, ability to discover different cluster shapes, ability to handle different data types, and ability to deal with noise and outliers.

Partitioning algorithms construct partitions of a data set on $n$ objects into a set of $c$ clusters. They attempt to determine $c$ partitions that optimize a certain objective function. The principle of this method is that objects in the same cluster are similar to each other, whereas objects of different clusters are very different. Hierarchical algorithms create a hierarchical decomposition of the data set. These methods either start with one cluster and then split into smaller clusters or start with each object forming a separate group and then try to merge similar clusters into larger clusters.

2.1 K-means algorithm

The K-means algorithm is described as follows:

Let $X = \{x_1, x_2, \ldots, x_n\}$ be the set of data points and $v = \{v_1, v_2, \ldots, v_c\}$ be the set of cluster centers.

Step 1. Select the number of cluster centers.
Step 2. Randomly initialize the $c$ cluster centroid vector.
Step 3. Compute the distance of each object in the data set from each of the cluster centroids.
Step 4. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
Step 5. i) Recalculate the new cluster center using

$$v_i = \frac{\sum_{j=1}^{c} x_i}{c_i}$$

where $c_i$ represents the number of data points in the $i$-th cluster.

ii) Recalculate the distance between each data point and new obtained cluster centers.
Step 6. If the stopping criterion has been met then stop otherwise go to step 4.

2.2 Fuzzy c-means algorithm

Given an unlabeled data set $X = \{x_1, x_2, \ldots, x_n\}$ in $R^n$ dimensional space, the FCM partitions data set into $c(1 < c < n)$ fuzzy clusters with $v = \{v_1, v_2, \ldots, v_c\}$ cluster centers by minimizing the objective function which is the weighted sum of squared errors within groups and is defined as follows:

$$J_m = \sum_{j=1}^{c} \sum_{i=1}^{n} u_{ij}^m d_{ij}^2$$

© Elsevier Publications 2013.
where \( d_{ij} \) is the distance from the object \( x_i \) to the cluster centers \( v_j \); \( u_{ij} \) is the membership degree of data point \( x_i \) to the \( j \)-th cluster; \( m \in [1, \infty) \) is a weighting exponent which determines the fuzziness of the resulting clusters; \( v_j \) is obtained using

\[
v_j = \frac{\sum_{i=1}^{n} u_{ij}^m x_i}{\sum_{i=1}^{n} u_{ij}^m}
\]  

(3)

The fuzzy partition matrix satisfies:

\[
u_{ij} \in [0, 1], \forall i = 1, 2, \ldots, n; \quad \forall j = 1, 2, \ldots, c
\]  

(4)

\[
\sum_{j=1}^{c} u_{ij} = 1, \forall i = 1, 2, \ldots, n
\]  

(5)

\[
0 < \sum_{i=1}^{n} u_{ij} < n, \forall j = 1, 2, \ldots, c \quad \text{and} \quad 1 < n < \infty
\]  

(6)

The Fuzzy \( c \)-means algorithm is stated as follows:

Step 1. Initialize the membership function values \( U_{t-1} = [u_{ij}^{(t-1)}] \) of \( x_i \) belonging to cluster \( v_j \) for \( 1 \leq i \leq n \), \( 1 \leq j \leq c \) (initially \( t \leftarrow 1 \)) and select \( m (m > 1) \)

Step 2. Calculate the cluster centroids \( v_j = [v_1^{(t)}, v_2^{(t)}, \ldots, v_c^{(t)}] \) for \( 1 \leq j \leq c \) using (3)

Step 3. Compute the distances between \( x_i \) and \( v_j^{(t)} \) for \( 1 \leq i \leq n \), \( 1 \leq j \leq c \)

Step 4. Update the membership of \( x_i \) at \( t \) by

\[
u_{ij}^{(t)} = \frac{1}{\sum_{k=1}^{c} \left( \frac{d_{ij}}{d_{ik}} \right)^{\frac{1}{m-1}}}
\]  

(7)

Step 5. If \( \|U_t - U_{t-1}\| < \epsilon \), then stop; otherwise \( t \leftarrow t + 1 \) and go to Step 2.

3. Distance Measures and Cluster Validity Indices

3.1 Distance measures

Clustering techniques are based on measuring similarity between data vectors by calculating the distance between each pair. There is no common distance measure which can be best suited for all clustering applications.

The following points are a few important characteristics of distance measure:

i. Distance is always positive.

ii. Distance from point \( a \) to itself is always zero.

iii. Distance from point \( a \) to point \( b \) cannot be greater than the sum of the distance from \( a \) to some other point \( c \) and distance from \( c \) to \( b \).

iv. Distance from \( a \) to \( b \) is always the same as from \( b \) to \( a \).

Chebyshev distance

This distance measure is based on the maximum attribute difference. It is also called Chessboard distance or Chebyshev norm or \( L_{\infty} \) norm. It is named after Pafnuty Lvovich Chebyshev. It calculates the absolute magnitude of the differences between coordinates of a pair of data vectors [6]. The Chebyshev distance between data vector \( x \) and centroid \( v \) is given by

\[
d(x, v) = \text{Max}_{i=1,2,\ldots,n} |x_i - v_i| \quad \text{(8)}
\]

Chi-square distance

The distance between data vector \( x \) and centroid \( v \) is computed as

\[
d(x, v) = \sqrt{n \sum_{i=1}^{n} \frac{(x_i - v_i)^2}{x_i + v_i}} \quad \text{(9)}
\]
Table 1. Cluster validity indices for fuzzy clustering.

<table>
<thead>
<tr>
<th>Validity Index</th>
<th>Function Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partition coefficient</td>
<td>$\frac{1}{n} \sum_{j=1}^{c} \sum_{i=1}^{n} u_{ij}^2$</td>
</tr>
<tr>
<td>Partition entropy</td>
<td>$-\frac{1}{n} \left( \sum_{j=1}^{c} \sum_{i=1}^{n} u_{ij} \log u_{ij} \right)$</td>
</tr>
</tbody>
</table>

| \(\sigma\)-distance |

Tsai and Lin [7] proposed a new distance metric that takes the distance variation in each clusters as the regularization of the Euclidean distance. The distance between data vector \(x\) and centroid \(v\) [8] is computed as

\[
d(x, v) = \frac{\|x - v\|}{\sigma_i}
\]

where \(\sigma_i\) is the weighed mean distance in cluster \(i\) and is given by

\[
\sigma_i = \left\{ \frac{\sum_{j=1}^{n} u_{ij}^m \|x - v\|^2}{\sum_{j=1}^{n} u_{ij}^m} \right\}^{1/2}
\]

3.2 Cluster validity indices

Cluster validity indices are used to evaluate the performance of fuzzy clustering. Many cluster validity indices were proposed in literature. Among them, the two measures such as partition coefficient [9,10] and partition entropy [9,11] are used in this paper. For good clustering results, partition coefficient should have the maximum value and partition entropy should have the minimum value. Table 1 summarizes the two cluster validity indices.

4. Methodology

The methodology used in this paper for data clustering is that the comparison of \(K\)-means and FCM algorithms has been studied using different distance measures such as Chebyshev, Chi-square and \(\sigma\)-distance measures. The algorithms are evaluated using the cluster validity indices such as partition coefficient and partition entropy.

The following distance measures are used in \(K\)-means and FCM algorithms:

**Chebyshev distance:** The distance between data vector \(x\) and centroid \(v\) is computed as

\[
d(x, v) = \text{Max}_{i=1,2,\ldots,n} |x_i - v_i|
\]

**Chi-square distance:** The distance between data vector \(x\) and centroid \(v\) is computed as

\[
d(x, v) = \sqrt{\sum_{i=1}^{n} \frac{(x_i - v_i)^2}{x_i + v_i}}
\]

**\(\sigma\)-distance:** The distance between data vector \(x\) and centroid \(v\) is computed as

\[
d(x, v) = \frac{\|x - v\|}{\sigma_i},
\]

where \(\sigma_i\) is the weighed mean distance in cluster \(i\) and is given by

\[
\sigma_i = \left\{ \frac{\sum_{j=1}^{n} u_{ij}^m \|x - v\|^2}{\sum_{j=1}^{n} u_{ij}^m} \right\}^{1/2}
\]
The objective function of FCM algorithm is calculated by
\[ \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} \left\| x - v_{i} \right\|^{2} / \sigma_{i} \] (12)

The following cluster validity indices are applied for FCM algorithm:

**Partition coefficient:** For each iteration of FCM algorithm, the partition coefficient is calculated by
\[ \frac{1}{n} \sum_{j=1}^{c} \sum_{i=1}^{n} u_{ij}^{2} \]

**Partition entropy:** For each iteration of FCM algorithm, the partition entropy is calculated by
\[ -\frac{1}{n} \left\{ \sum_{j=1}^{c} \sum_{i=1}^{n} u_{ij} \log u_{ij} \right\} \]

5. Applications

K-means and FCM algorithms have been used in many real-world application areas. They are applied for image segmentation, patterns in biological sequences, face recognition, cellular manufacturing, network intrusion detection system, prediction of students academic performance, meteorological data application, application in university libraries, evaluation of success of software reuse and so on.


Castro, Boveda and Arcay [15] have analyzed the Fuzzy \( c \)-means algorithms for the segmentation of burn wounds photographs. Ye Qian [16] presented \( K \)-means algorithm which is based on the historical financial ratios, by making use of the cluster analysis technology to analyze the listed companies.


Hong Liu and Xiaohong Yu [21] presented an application of \( k \)-means clustering algorithm to image retrieval system. Meng Jianliang, Shang Haikun, Bian and Ling [22] used \( K \)-means algorithm to cluster and analyze the data on application on network intrusion detection system. Wuling Ren, Jinzhu Cao and Xianjie Wu [23] applied Fuzzy \( c \)-means clustering algorithm to the detection of network intrusion. Guangquan Liu, Gan Huang, Jianjun Meng, Dingguo Zhang and Xiangyang Zhu, [24] proposed an unsupervised approach based on FCM algorithm for the online adaptation of the LDA classifier for electroencephalogram (EEG) based BCI.

Ovelade, Oladipupo, and Obagbuwa [25] implemented \( k \) Means clustering algorithm for predicting the academic performance of students. This method is a good bench mark to monitor the progress of students academic performance. Hongwei Xei, Li Zhang, Jingyu Sun, and Xueli Yu [26] discussed the \( K \)-means clustering algorithm and how to cluster news comments in order to obtain types of a special news comments. Yuan Sun, Shiqiu Song, Guoqing Wang, Zhenxing Li, Zhengbo Xu and Lifang Wei [27] proposed fuzzy \( c \)-means algorithm for classification of tobaccos based on their rare earth elements. Zhiye Sun, Li Gao, Shuang Wei and Shijue Zheng [28] applied Fuzzy \( c \)-means algorithm in the meteorological data application. They compensated effectively the deficiencies of the algorithm and achieved a relatively good clustering effect of temperature observation data.

Table 2. Real world well-known UCI data sets with their characteristics.

<table>
<thead>
<tr>
<th>Data set</th>
<th>No. of objects</th>
<th>No. of attributes</th>
<th>No. of clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMC</td>
<td>1473</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Diabetes</td>
<td>768</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Liver Disorders</td>
<td>365</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Statlog (Heart)</td>
<td>270</td>
<td>13</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3. Comparative performance of K-means and FCM using Chebyshev and Chi-square distance measures for various data sets.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Distance Measure</th>
<th>K-means</th>
<th>FCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMC</td>
<td>Chebyshev</td>
<td>15445.258</td>
<td>11545.637</td>
</tr>
<tr>
<td></td>
<td>Chi-square</td>
<td>2630.398</td>
<td>23121.580</td>
</tr>
<tr>
<td>Diabetes</td>
<td>Chebyshev</td>
<td>4272303.420</td>
<td>3317486.399</td>
</tr>
<tr>
<td></td>
<td>Chi-square</td>
<td>29411.859</td>
<td>3154.535</td>
</tr>
<tr>
<td>Liver Disorders</td>
<td>Chebyshev</td>
<td>305492.100</td>
<td>239373.648</td>
</tr>
<tr>
<td></td>
<td>Chi-square</td>
<td>4159.772</td>
<td>3154.535</td>
</tr>
<tr>
<td>Statlog (Heart)</td>
<td>Chebyshev</td>
<td>408446.992</td>
<td>305718.390</td>
</tr>
<tr>
<td></td>
<td>Chi-square</td>
<td>2431.845</td>
<td>1627.913</td>
</tr>
</tbody>
</table>


6. Experimental Results

The main objective of this paper is to compare the K-means and Fuzzy C-means algorithm to data clustering problem using distance measures such as Chebyshev, Chi-square and σ-distance measures. The Fuzzy C-means algorithm is measured using cluster validity indices such as partition coefficient and partition entropy.

A. Data sets

For evaluating K-means and Fuzzy C-means algorithms, the following four popular real-world data sets from the UCI machine learning repository have been considered:

- Contraceptive Method Choice (CMC), which consists of 1473 objects and 3 different types characterized by 9 attributes.
- Diabetes, which consists of 768 objects and 2 different types characterized by 8 attributes.
- Liver Disorders, which consists of 345 objects and 2 different types characterized by 6 attributes.
- Statlog (Heart), which consists of 270 objects and 2 different types characterized by 13 attributes.

Table 2 summarizes these four data sets. For each data set, it records the number of objects, number of attributes and number of clusters.

B. Results

The algorithms are implemented in Java. For our experimental tests, we used a PC Pentium IV (CPU 3.06 GHZ and 1.97 GB RAM) by considering the maximum of 100 iterations and 10 independent test runs, the stop criterion for iteration $C = 0.00001$ and weighting exponent in FCM $m = 2$.

The objective function values of K-means and Fuzzy c-means algorithms are shown in table 3. The Fuzzy c-means algorithm based on Chi-square distance measure is better value than K-means algorithm for all the data sets. The objective function value of Fuzzy c-means algorithm for σ-distance measure for all the data sets is given in table 4. The cluster validity indices such as partition coefficient and partition entropy for different distance measures are recorded in table 5. The Chebyshev distance measure shows the maximum partition coefficient value and minimum partition entropy than other distance measures for various data sets.
Table 4. Objective function value by FCM using $\sigma$-distance measure.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Objective Function Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMC</td>
<td>5389.081</td>
</tr>
<tr>
<td>Diabetes</td>
<td>67551.115</td>
</tr>
<tr>
<td>Liver Disorders</td>
<td>12673.465</td>
</tr>
<tr>
<td>Statlog (Heart)</td>
<td>8776.705</td>
</tr>
</tbody>
</table>

Table 5. Cluster validity indices for FCM algorithm.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Distance Measure</th>
<th>Partition Coefficient</th>
<th>Partition Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMC</td>
<td>Chebyshev</td>
<td>0.7636</td>
<td>0.4347</td>
</tr>
<tr>
<td></td>
<td>Chi-square</td>
<td>0.4251</td>
<td>0.9610</td>
</tr>
<tr>
<td></td>
<td>$\sigma$-distance measure</td>
<td>0.6230</td>
<td>0.6635</td>
</tr>
<tr>
<td>Diabetes</td>
<td>Chebyshev</td>
<td>0.8494</td>
<td>0.2543</td>
</tr>
<tr>
<td></td>
<td>Chi-square</td>
<td>0.7293</td>
<td>0.4326</td>
</tr>
<tr>
<td></td>
<td>$\sigma$-distance measure</td>
<td>0.7490</td>
<td>0.3832</td>
</tr>
<tr>
<td>Liver Disorders</td>
<td>Chebyshev</td>
<td>0.8503</td>
<td>0.2555</td>
</tr>
<tr>
<td></td>
<td>Chi-square</td>
<td>0.7640</td>
<td>0.3846</td>
</tr>
<tr>
<td></td>
<td>$\sigma$-distance measure</td>
<td>0.7552</td>
<td>0.3848</td>
</tr>
<tr>
<td>Statlog (Heart)</td>
<td>Chebyshev</td>
<td>0.7503</td>
<td>0.3989</td>
</tr>
<tr>
<td></td>
<td>Chi-square</td>
<td>0.5506</td>
<td>0.6412</td>
</tr>
<tr>
<td></td>
<td>$\sigma$-distance measure</td>
<td>0.6413</td>
<td>0.5380</td>
</tr>
</tbody>
</table>

7. Conclusions

Data clustering is a useful technique for the discovery of knowledge from a data set. The popular clustering techniques such as $K$-means and FCM algorithms are mostly based on Euclidean distance. In this paper, we have made a comparative study of these algorithms with different distance measures such as Chebyshev, Chi-square and $\sigma$-distance measures. The new algorithms have been tested on the four well known data sets such as Contraceptive Method Choice (CMC), Diabetes, Liver Disorders and Statlog (Heart) from the UCI repository. The experimental results showed that FCM based on Chi-square distance measure had better result than Chebyshev distance measure. We have also proposed FCM algorithm based on $\sigma$-distance measure. The Chebyshev distance measure produced the maximum partition coefficient and minimum partition entropy than the other distance measures. A brief review of applications of $K$-means and Fuzzy $c$-means algorithms were also presented.

Recently, hybrid evolutionary algorithms used to improve the performance of clustering. As a future study, the $K$-means algorithm and FCM algorithms can be integrated with evolutionary algorithms to produce the good clustering results.

References


© Elsevier Publications 2013.