Proportional explorations and recital analysis of FCM and MFPCM algorithms on Iris data

Rashid Husain
Department of Computer Science and Information Technology Kebbi State University of Science and Technology, Aliero, Nigeria. E-mail: rashid65_its@yahoo.com.
Accepted 18 December, 2011

Data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. Data mining is a computational intelligence discipline that contributes tools for data analysis, discovery of new knowledge, and autonomous decision making. Clustering is a primary data description method in data mining which groups most similar data. The data clustering is an important problem in a wide variety of fields including data mining, pattern recognition, and bioinformatics. It aims to organize a collection of data items into clusters, such that items within a cluster are more similar to each other than they are in the other clusters. There are various algorithms used to solve this problem. In this paper, we use FCM (Fuzzy C-mean) clustering algorithm and MFPCM (Modified Fuzzy Possibilistic C-mean) clustering algorithm. In this paper we compare the performance analysis of Fuzzy C-mean (FCM) clustering algorithm and compare it with Modified Fuzzy possibilistic C-mean algorithm. In this we compared FCM and MFPCM algorithm on different data sets. We measure complexity of FCM and MFPCM at different data sets. FCM clustering is a clustering technique which is separated from Modified Fuzzy Possibilistic C-Mean that employs Possibilistic partitioning.

Key words: Data clustering Algorithm, Portioning, Data Mining, Fuzzy C-Mean, Modified Fuzzy Possibilistic C-Mean.

INTRODUCTION

Data analysis is considered as a very important science in the real world. Data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. Data mining is a computational intelligence discipline that contributes tools for data analysis, discovery of new knowledge, and autonomous decision making [Ude et al., 2009 & Sorin Istrail, 2010]. The task of processing large volume of data has accelerated the interest in this field. As mentioned in Mosley (2005) data mining is the analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Data mining discovers description through clustering visualization, association, sequential analysis. Clustering is a primary data description method in data mining which group’s most similar data. Data clustering is a common technique for data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Cluster analysis is a technique for classifying data; it is a method for finding clusters of a data set with most similarity in the same cluster and most dissimilarity between different clusters. The conventional clustering methods put each point of the data set to exactly one cluster. Since 1965, Zadeh proposed fuzzy sets in order to come closer of the physical world [Agrawal Deepal, 2008]. Zadeh introduced the idea of partial memberships described by membership functions. Clustering algorithm partitions an unlabeled set of data into groups according to the similarity. Compared with the data classification, the data clustering is an unsupervised learning process, it does not need a labeled data set as training data, but the performance of the data clustering algorithm is often much poorer. Although the data classification has better performance, it needs a labeled data set as training data and labeled data for the classification is often very difficult and expensive to obtain. So there are many algorithms are proposed to improve the clustering performance. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each
subset shares one common trait [Agrawal Deepal, 2008 and Brej and Sonla, 2000]. Clustering technique is used for combining observed objects into clusters (groups), which satisfy two main criteria:

i. Each group or cluster should be homogeneous objects that belong to the same group are similar to each other.

ii. Each group of cluster should be different from other clusters, that is, objects that belong to one cluster should be different from the objects of other clusters.

Clustering can be considered the most important unsupervised learning problem. So, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects, which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. There are many clustering methods available, and each of them may give a different grouping of a data set. The choice of a particular method will depend on the type of output desired, the known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset.

**FUZZYC-MEANALGORITHM**

Fuzzy C-Mean (FCM) is a data clustering [6, 9] technique in which a data set is grouped into $c$ clusters with every data point in the data set belonging to every cluster will have a high degree of belonging or membership to that cluster and another data point that lies far away from the center of a cluster will have a low degree of belonging or membership to that cluster.

The steps of FCM algorithm are given below.

Fix $c$ and $c (2 <= c <= n)$ and select a value for parameter. Initialize the partition matrix $U(0)$. Each step in this algorithm will be labeled as $r$, where $r = 0, 1, 2, ...$

1. Calculate the $c$ center vector $V_{ij}$ for each step

$$V_{ij} = \frac{\sum_{k=1}^{n} u_{ik}^m x_{ij}}{\sum_{k=1}^{n} u_{ik}^m}$$

2. Calculate the distance matrix

$$D_{c,n} = \left[ \sum_{j=1}^{n} \left( x_{kj} - v_{ij} \right)^2 \right]^{1/2}$$

3. Update the partition matrix for $r^{th}$ step. $U^r$ as follow:

$$u_{ik}^{r+1} = \left[ \frac{1}{\sum_{j=1}^{c} \left( \frac{d(x_i, v_j)}{d(x_i, v_k)} \right)^{2/m_i} } \right]^{1/(m_i-1)}$$

If $\left| U^{k+1} - U^k \right| < \delta$ then stop. Otherwise return to step 2 by iteratively updating the cluster centers and the membership grades for data point. FCM iteratively moves the cluster to the “right” location within a dataset.

**Modified Fuzzy Possibilistic C-Mean Algorithm**

The FPCM algorithm attempts to partition a finite collection of elements $X = \{x_1, x_2, x_3, ... x_n\}$ into a collection of $c$ fuzzy clusters with respect to some given criterion [6, 7]. Given a finite set of data, the algorithm returns a list of $c$ cluster centers $V$, such that $V = v_l l = 1, 2, 3, ... c$, and a partition matrix $U$ such that $U = u_{lj} l = 1, 2, 3, ... c$ and $j = 1, 2, 3, ... n$. Where $u_{lj}$ is a numerical value in $[0, 1]$ that tells the degree to which the elements $x_j$ belongs to the $lth$ cluster. Defines a family of fuzzy sets $\{\mu_i l = 1, 2, 3, ... c\}$ as a fuzzy c partition on a universe of data points $X$.

1. Fuzzy set allows for degree of membership
2. A single point can have partial membership in more than one class.
3. There can be no empty classes and no class that contains no data points

The steps of MFPCM algorithm given below:

1. The objective function of the MFPCM can be formulated as follows:

$$I_{MFPCM} = \sum_{l=1}^{c} \sum_{j=1}^{n} \mu_{ij}^{m} \left( \frac{d_{ij}}{d_{ij}} + t_{ij} \mu_{ij} \right) (x_j, v_l(j))$$

2. Calculate $U = \{u_{lj}\}$ represents a fuzzy partition matrix, is defined as:

$$u_{lj} = \left[ \left( \frac{\sum_{l=1}^{c} \left( \frac{d(x_i, v_j)}{d(x_i, v_k)} \right)^{2/m_i} }{d(x_i, v_k)} \right)^{-1/(m_i-1)} \right]^{1/n}$$
Table 1: Time Complexity when Number of cluster varying

<table>
<thead>
<tr>
<th>S. N.</th>
<th>Number of Cluster</th>
<th>FCM Time Complexity</th>
<th>MFPCM Time Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2000</td>
<td>2000</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>10000</td>
<td>5000</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>26000</td>
<td>8500</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>45000</td>
<td>11000</td>
</tr>
</tbody>
</table>

1. Calculate \( T = \{ t_{ij} \} \) represents a typical partition matrix, is defined as:

\[
t_{ij} = \left( \sum_{k=1}^{n} \left( \frac{d_j x_j v_i}{d_j x_j v_k} \right)^{2/(m-1)} \right)^{-1}
\]

2. Calculate \( V = \{ v_{ij} \} \) represents \( c \) centers of the clusters, is defined as:

\[
v_{ij} = \frac{\sum_{j=1}^{n} \left( \mu_{ij}^{m} \omega_{ij}^{m} + t_{ij} \omega_{ij}^{m} \right) x_j}{\sum_{j=1}^{n} \left( \mu_{ij}^{m} \omega_{ij}^{m} + t_{ij} \omega_{ij}^{m} \right)}
\]

**RESULTS**

**Time complexity of FCM and MFPCM by varying no. of Clusters on Iris Data**

The implementation of FCM & MFPCM is done on iris Data in MATLAB. The data t contains 3 classes of 150 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other two; the latter are NOT linearly separable from each other [8, 9]. The dataset contain four attribute which are given below:

The time complexity of FCM (Almedia and Sousa, 2006) is \( O(ndc^2t) \) and time complexity of MFPCM is \( O(ncd) \). Now keeping no. of data points constant, let’s assume \( n=100, d=3, c=3 \) and varying no. of clusters, we obtain the following table and graph. Where \( n=\)number of data point, \( c=\)number of cluster,

\( d=\)dimension, \( l=\)number of iteration.

**Time Complexity when Number of cluster varying**

Now keeping no. of cluster constant, lets assume \( n=140, d=3, c=3 \) and varying no. of iteration, we obtain the following table and graph.

**Comparison of space complexity of FCM and MFPCM**

The space complexity of FCM is \( O(nd + nc) \) and
Table 2. Time Complexity when Number of Iterations varying.

<table>
<thead>
<tr>
<th>S. N.</th>
<th>Number of Iteration</th>
<th>FCM Time Complexity</th>
<th>MFPCM Time Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>5000</td>
<td>4000</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>11000</td>
<td>5000</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>15000</td>
<td>7000</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>25000</td>
<td>11000</td>
</tr>
</tbody>
</table>

Figure 2: Time complexity of FCM and MFPCM by varying no. of Iterations.

MFPCM is \( O(nd) \). Now keeping no. of data points constant, lets assume \( n=140, d=3 \) and varying no. of clusters we obtain the following graph.

Complexity Analysis of FCM Algorithm:

The asymptotic efficiency of the algorithm has following notations:

i. \( I \) number FCM over entire dataset.
ii. \( n \) number of data points.
iii. \( c \) number of clusters
iv. \( d \) number of dimensions

The time complexity of the fuzzy c mean algorithm is \( O(n^d c I) \), where empirically \( I \) grows very slowly with \( n \), \( c \) and \( d \). The memory complexity of FCM is \( O(n^d + n c) \), where \( n \) is the size of data set and \( nc \) the size of \( U \) matrix. For data sets, which cannot be loaded into memory, FCM will have disk accesses every iteration. Thus the disk input output complexity will be \( O(n d^2) \). It is likely that for those data sets the \( U \) matrix cannot be kept in memory too. Thus, it will increase the disk input/output complexity further.

Complexity Analysis of MFPCM Algorithm

The asymptotic efficiency of the algorithm has following notations:

i. \( I \) number of \( k \) means passes over entire data set
ii. \( n \) number of data points.
iii. \( c \) number of clusters
iv. \( d \) number of dimensions.

The time complexity of the hard c mean algorithm is \( O(n^d c I) \), where empirically \( I \) grows very slowly with \( n \), \( c \) and \( d \).

The memory complexity of MFPCM is \( O(n cd) \). I/O complexity of MFPCM is \( O(n^d) \).
Table 3. Space Complexity when Number of Clusters varying.

<table>
<thead>
<tr>
<th>S.No</th>
<th>Number of Cluster</th>
<th>FCM Space Complexity</th>
<th>MFPCM Space Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>400</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>600</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>800</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>1000</td>
<td>12</td>
</tr>
</tbody>
</table>

CONCLUSION

In partitioning based clustering algorithms, the number of final cluster (k) needs to be defined beforehand. Also, algorithms have problems like susceptible to local optima, sensitive to outliers, memory space and unknown number of iteration steps required to cluster. The time complexity of the MFPCM is $O(nd^2)$. The memory complexity of MFPCM is $O(nd + nc)$, and the disk input output complexity will be $O(ndl)$. Fuzzy clustering, which constitute the oldest component of soft computing, are suitable for handling the issues related to understandability of patterns, incomplete/noisy data, mixed media information and human interaction, and can provide approximate solutions faster. They have been mainly used in discovering association rules and functional dependencies and image retrieval. The time complexity of the Fuzzy C-Mean algorithm is $O(ndc^2)$. The memory complexity of FCM is $O(nd + nc)$, and the disk input output complexity will be $O(ndl)$.

REFERENCES

learning from examples”, IEEE Transactions on Medical imaging, 19: 973-985.