Clustering Accuracies on Concepts of Mathematics

Fong-Jhu Yih

Abstract. Fuzzy c-means algorithm (FCM) based on Euclidean distance function converges to a local minimum of the objective function, which can only be used to detect spherical structural clusters. Gustafson-Kessel clustering algorithm and Gath-Geva clustering algorithm were developed to detect non-spherical structural clusters. However, Gustafson-Kessel clustering algorithm needs added constraint of fuzzy covariance matrix, Gath-Geva clustering algorithm can only be used for the data with multivariate Gaussian distribution. In GK-algorithm, modified Mahalanobis distance with preserved volume was used. However, the added fuzzy covariance matrices in their distance measure were not directly derived from the objective function. Improved Normalized Mahalanobis Clustering Algorithm Based on FCM by taking a new threshold value and a new convergent process is proposed. The experimental results of real data sets show that our proposed new algorithm has the best performance. Not only replacing the common covariance matrix with the correlation matrix in the objective function in the Normalized Mahalanobis Clustering Algorithm.

Keywords: Mahalanobis distance, Fuzzy c-means algorithm (FCM), Clustering Algorithm.

Introduction

Fuzzy clustering is a branch in clustering analysis and it is widely used in the pattern recognition field. The well-known ones, such as Bezdek’s Fuzzy C-Means (FCM) and Li et al.’s Fuzzy Weighted C-Means (FWCM) [1,2], are based on Euclidean distance. These fuzzy clustering algorithms can only be used to detect the data classes with the same super spherical shapes.

The Fuzzy C-Means algorithm based on adaptive Mahalanobis distances, common Mahalanobis distance and standardized Mahalanobis distance, respectively (FCM-M and FCM-CM), [8-12,13] were proposed, and then, the fuzzy covariance matrices in the Mahalanobis distance can be directly derived by minimizing the objective function. In our three previous works, to add a regulating factor of Each covariance matrix to each class in the objective function, and deleted the constraint of the determinants of
covariance matrices in the GK algorithm, the Fuzzy C-Means algorithm based on adaptive Mahalanobis distances, common Mahalanobis distance and standardized Mahalanobis distance, respectively (FCM-M and FCM-CM), [8-12,13] were proposed, and then, the fuzzy covariance matrices in the Mahalanobis distance can be directly derived by minimizing the objective function.

Literature References

Groups data into clusters so that the data objects within a cluster have high similarity in comparison to one another, but are very dissimilar to those data objects in other clusters. **GK Algorithm** FCM can only work well for spherical shaped clusters. In the objective function the distances between data points to the centers of the clusters are calculated by Euclidian distances. To overcome the above drawback, we could try to extend the distance measure to Mahalanobis distance (MD). However, Krishnapuram and Kim (1999) [2] pointed out that the Mahalanobis distance cannot be used directly in clustering algorithm. Gustafson and Kessel (1979) extended the Euclidian distances of the standard FCM by employing an adaptive norm, in order to detect clusters of different geometrical shape without changing the clusters’ sizes in one data set.

**GG Algorithm** Gath-Geva (GG) fuzzy clustering algorithm is an extension of Gustafson-Kessel (GK) fuzzy clustering algorithm, and also takes the size and density of clusters for classification. [7] Hence, it has better behaviors for irregular features. Probabilistic interpretation of GG clustering is shown by Equation (1):

\[
P(X | \eta) = \sum_{\eta} p(X, \eta) = \sum_{\eta} p(\eta) P(X | \eta) \tag{1}
\]

Gath and Geva (1989) [9] assumed that the normal distribution with expected value and covariance matrix is chosen for generating a datum with prior probability, satisfying

\[
p(\xi | \eta) = \frac{p}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp \left[ -\frac{1}{2}(\xi - \mu)^T \Sigma^{-1} (\xi - \mu) \right] \tag{2}
\]

Where \( X = [\xi_1, \xi_2, \ldots, \xi_r], \xi_r \in R^p, j = 1,2,\ldots, n \) is the data matrix, the covariance matrix of cluster \( i \) is \( \Sigma_r \in R^{p \times p} \). \( p \) is the number of dimension of data, \( c \) is the number of clusters, \( d^2(\xi_r, \bar{\mu}) \) for GG algorithm is chosen to be indirectly proportional to Equation (2) which is the posterior probability (likelihood) function. A small distance means a high probability, and a large distance means a low probability for membership. GG algorithm is based on minimization of the sum of weighted square distances between the data and the cluster centers of the objective function in Equation (3)

\[
J^*_{c,n}(U,A,\Sigma,X) = \sum_{i=1}^{c} \sum_{\eta \in \Sigma} \mu_i^a d^2(\xi_r, \bar{\mu}) \tag{3}
\]

Conditions for probability clusters partition are

\[
m \in [1, \infty); U = [\mu_i], \mu_i \in [0,1], i = 1,2,\ldots, c, j = 1,2,\ldots, n
\]

\[
\sum_{i=1}^{c} \mu_i = 1, j = 1,2,\ldots, n, 0 < \sum_{i=1}^{c} \mu_i < n, i = 1,2,\ldots, c
\]

\[
d^2(\xi_r, \bar{\mu}) = \frac{1}{p(\xi | \eta)} \left( \frac{2\pi)^{p/2}|\Sigma|^{1/2}}{p} \exp \left[ -\frac{1}{2}(\xi - \mu)^T \Sigma^{-1} (\xi - \mu) \right] \right] \tag{5}
\]
Minimizing the objective function respect to all parameters in Equation (4), with the constraint (5), we can obtain the following GG algorithm.

**FCM-M Algorithm** For improving the limitation of GK algorithm and GG algorithm, we added a regulating factor of covariance matrix, 

\[
D = \sum_{j=1}^{c} \sum_{i=1}^{n} \left( x_i - \mu_0^{(i)} \right) \left( x_i - \mu_0^{(i)} \right)^T > 0
\]

and each class in the objective function, and deleted the constraint of the determinant of covariance matrices. We can obtain the objective function of Fuzzy C-Means based on adaptive Mahalanobis distance (FCM-M) as following [8-12]. The steps of the FCM-M are listed as follows [8].

**Step 1**: Determining the number of cluster; c and m-value (let m=2), given converge error, (such as). Randomly choose the initial membership \( u^{(0)}_{ij}, i = 1, 2,..., c, j = 1, 2,..., n \), such that

\[
\sum_{j=1}^{c} u_{ij}^{(0)} = 1, \quad j = 1, 2,..., n
\]

\[
u_{(0)}^{(i)} = \left[ \sum_{k=1}^{n} \mu_{ik}^{(0)} \right]^{-1} \sum_{k=1}^{n} \mu_{ik}^{(0)}, \quad i = 1, 2,..., c
\]

\[
D = \sum_{j=1}^{c} \sum_{i=1}^{n} \left( x_i - \mu_0^{(i)} \right) \left( x_i - \mu_0^{(i)} \right)^T > 0
\]

\[
\Sigma_{(0)}^{(i)} = \left[ \sum_{k=1}^{n} \mu_{ik}^{(0)} \right] \left( x_i - \mu_0^{(i)} \right) \left( x_i - \mu_0^{(i)} \right)^T \sum_{k=1}^{n} \mu_{ik}^{(0)}
\]

\[
\text{if } \left| \xi_{(0)}^{(i)} \right| > D \quad \text{or } \quad \left| \xi_{(0)}^{(i)} \right| < \frac{1}{D} \quad \text{then } \Sigma_{(0)}^{(i)} = I
\]

**Step 2**: Find

\[
\xi_{(0)}^{(i)} = \left[ \sum_{k=1}^{n} \left( x_i - \mu_0^{(i)} \right) \left( x_i - \mu_0^{(i)} \right)^T \right]^{-1} \sum_{k=1}^{n} \left( x_i - \mu_0^{(i)} \right)
\]

\[
\mu_{(0)}^{(i)} = \left[ \sum_{k=1}^{n} \left( x_i - \mu_0^{(i)} \right) \left( x_i - \mu_0^{(i)} \right)^T \right]^{-1} \sum_{k=1}^{n} \left( x_i - \mu_0^{(i)} \right)
\]

\[
\text{if } \left| \xi_{(0)}^{(i)} \right| > D \quad \text{or } \quad \left| \xi_{(0)}^{(i)} \right| < \frac{1}{D} \quad \text{then } \Sigma_{(0)}^{(i)} = I
\]

**Step 3**: Increment k; \( \frac{1}{c} \sum_{i=1}^{c} \left| \xi_{(0)}^{(i)} - \mu_0^{(i)} \right| < \varepsilon \)

Note that FCM is a special case of FCM-M, when covariance matrices equal to identity matrices [8].

**FCM-CM Algorithm** For improving the stability of the clustering results, we replace all of the covariance matrices with the same common covariance matrix in the objective function in the FCM-M algorithm. An improved fuzzy clustering method called FCM-CM is proposed. We can obtain the objective function of FCM-CM as following:

\[
J_{\text{FCM-CM}}(U, A, \Sigma, X) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^{(i)} d^2 \left( x_j, \Sigma \right)
\]
\[ m \in [1, \infty); U = \left[ \mu_i \right]_{i=1}^{c} \ ; \mu_i \in [0,1], i = 1, 2, \ldots, c, j = 1, 2, \ldots, n \]
\[ \sum_{j=1}^{n} \mu_{ij} = 1, j = 1, 2, \ldots, n, 0 \leq \sum_{i=1}^{c} \mu_{ij} \leq n, i = 1, 2, \ldots, c \]
\[ d'(z-x)^{\sum (z-x)^{-1} (x-x)^{-1}} \]
\[ d'(z-x)^{\sum (z-x)^{-1} (x-x)^{-1}} \]

Minimizing the objective function respect to all parameters in Equation (17) with the constraint (16), we can obtain the following FCM-CM algorithm [12].

New Clustering Algorithm

FCM-NM Algorithm Not only z-score normalizing for each feature in the objective function in the FCM-CM algorithm, but also replacing the threshold \( D \), where
\[ D = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu^{(c)}_{ij} \left[ (z - \mu^{(c)}) (z - \mu^{(c)}) \right] > 0 \]

With the determinant value of the crisp correlation matrix, the new fuzzy clustering method, called the Fuzzy C-Means algorithm based on normalized Mahalanobis distance (FCM-NM) is proposed.

Experiment of Real Data

The features of the Iris data contain Length of Sepal, Width of Sepal, Length of Petal, and Width of Petal. The samples were assigned the original 3 clusters based on the clustering analysis. The results were shown in Table 1.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Samples size</th>
<th>Concepts</th>
<th>Average distance of center</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>Setosa</td>
<td>0.48170523644</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>Versicolor</td>
<td>0.70687020404</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>Virginica</td>
<td>0.81933940766</td>
</tr>
</tbody>
</table>

The performances of three clustering methods, FCM, GG, GK, FCM-M, FCM-CM and FCM-NM, are compared in the experiments. The Iris Data [14] with sample size 150 is used as first example. The classification accuracies testing samples were shown in Table 2.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracies</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCM</td>
<td>0. 8933</td>
</tr>
<tr>
<td>GG</td>
<td>0. 7649</td>
</tr>
<tr>
<td>GK</td>
<td>0.9000</td>
</tr>
<tr>
<td>FCM-M</td>
<td>0.9000</td>
</tr>
<tr>
<td>FCM-CM</td>
<td>0.9279</td>
</tr>
<tr>
<td>FCM-NM</td>
<td>0. 9467</td>
</tr>
</tbody>
</table>

Summary

The well-known FCM is based on Euclidean distance function, which can only be used to detect spherical structural clusters. GK algorithm and GG algorithm were developed to detect non-spherical structural clusters. However, the former needs added constraint of fuzzy covariance matrix, the later can only be used for the data with multivariate...
Gaussian distribution. Three improved Fuzzy C-Means algorithm based on different Mahalanobis distance, called FCM-M, FCM-CM, and FCM-SM were proposed by our previous works. In this paper, a further improved Fuzzy C-Means algorithm based on a normalized Mahalanobis distance (FCM-NM) by taking a new convergent process is proposed. The experimental result of the real data set shows that our proposed new algorithm has the best performance.

References