MATEC Web of Conferences **119**, 01007 (2017) DOI: 10.1051/matecconf/201711901007
IMETI 2016

Clustering with fuzzy supervised algorithm

Fong-Jhu Yih[1], Yuan-Horng Lin[2] and Jeng-Ming Yih[3,a]

[1]*Department of Information Engineering, National Chung Cheng University, Chiayi 621, Taiwan*
[2]*Department of Mathematics Education National Tai-Chung University of Education, Taichung City 403, Taiwan*
[3]*Department of General Education, Min-Hwei Junior College of Health Care Management, Tainan 736, Taiwan*

Abstract. 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. A Fuzzy C-Means algorithm based on Mahalanobis distance (FCM-M) was proposed to improve those limitations of GG and GK algorithms, but it is not stable enough when some of its covariance matrices are not equal. In this paper, an improved Supervised 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 Supervised Clustering Algorithm.

1 Introduction

To overcome the drawback due to Euclidean distance, we could try to extend the distance measure to Mahalanobis distance (MD). However, Krishnapuram and Kim (1999)pointed out that the Mahalanobis distance can not be used directly in clustering algorithm. Gustafson-Kessel (GK) clustering algorithm and Gath-Geva (GG) clustering algorithm were developed to detect non-spherical structural clusters. In GK-algorithm, a 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. In GG algorithm, the Gaussian distance can only be used for the data with multivariate normal distribution. 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 Mahalanobis distance (FCM-M).Fuzzy partition clustering is a branch in cluster analysis and it is widely used in pattern recognition. Among many well- known fuzzy partition clustering algorithms, Bezdek's Fuzzy C-Means (FCM) (1981), Pal, Pal and Bezdek's Possibility C-Means (PCM) (1993), and Pal, Pal and Bezdek's Fuzzy Possibility C-Means (FPCM) (1997) are all based on Euclidean distance measure for clustering. Hence, those fuzzy partition clustering algorithms can only be used for the data set with the same super spherical shape for each class. Instead of using Euclidean distance measure, Gustafson and Kessel (1979) proposed the G-K algorithm which employs the Mahalanobis distance. It is a fuzzy partition clustering algorithm which can be used for the classes with different geometrical shapes in the data set. However, without the prior information of the shape volume for each class, the G-K

[a] Corresponding author : yih@mail.ntcu.edu.tw

algorithm can only be utilized for the classes with the same volume. In other words, if any dimension of a class is greater than the number of samples in the class, the estimated covariance matrix of the class may not be fully ranked. Hence, the algorithm will induce the singular problem for the inverse covariance matrix. This is an important issue need be addressed when we use the G-K algorithm for clustering. To overcome the issues, a new solution is proposed. A regulating factor of the covariance matrix is added to each class in the objective function, and the constraint of the determinant of the covariance matrices defined in the G-K algorithm is removed. Furthermore, the FCM-AM algorithms included two algorithms, FCM-M and FCM-CM, proposed by our previous works (Hsiang-Chuan Liu, Jeng-Ming Yih, Shin-Wu Liu, 2007, 2008).

2 Fuzzy Partitions Clustering Algorithms Based on Euclidean Distance

The popular fuzzy c-means algorithm 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. The objective of a fuzzy clustering algorithm is to partition the data into clusters so that the similarity of data objects within each cluster is maximized and the similarity of data objects among clusters is minimized. In the objective function based methods, the objective function is a function of data matrix, membership matrix and prototypes of clusters. It measures the overall dissimilarity of data objects within each cluster. Hence, by minimizing the objective function, we can obtain the best partition of the data set.

2.1 Fuzzy C-Means Algorithm

To overcome the drawback due to Euclidean distance, we could try to extend the distance measure to Mahalanobis distance (MD). However, Krishnapuram and Kim (1999) pointed out that the Mahalanobis distance can not be used directly in clustering algorithm. Fuzzy C-Means Algorithm (FCM) which objective function of FCM is given.

2.2 Fuzzy Clustering

Clustering technique plays an important role in data analysis and interpretation. Fuzzy clustering is a branch in clustering analysis and it is widely used in the pattern recognition field. Fuzzy clustering algorithms can only be used to detect the data classes with the same super spherical shapes. To overcome the drawback due to Euclidean distance, we could try to extend the distance measure to Mahalanobis distance (MD). However, Krishnapuram and Kim (1999) pointed out that the Mahalanobis distance can not be used directly in clustering algorithm. Gustafson-Kessel (GK) clustering algorithm and Gath-Geva (GG) clustering algorithm were developed to detect non-spherical structural clusters. In GK-algorithm, a 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. In GG algorithm, the Gaussian distance can only be used for the data with multivariate normal distribution. 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 Mahalanobis distance (FCM-M) was proposed, and then For improving the stability of the FCM-M clustering results, Replace all of the covariance matrices with the same common covariance matrix in the objective function in the FCM-M algorithm.

2.3 The FCM-AM Algorithm and Its Special Cases

Using the Liu-algorithm, we can obtain the objective function of the Fuzzy C-Means algorithm based on Alternative Mahalanobis distances (FCM-AM) as following.

2.4 FCM-NM Algorithm

In this study, a fuzzy clustering algorithm called Fuzzy C-Means algorithm based on normalized Mahalanobis distance (FCM-NM) is used, which was improved by normalizing for each feature in the objective function and also replacing the threshold in the FCM-CM algorithm. We can obtain the objective function of FCM-NM as following:

3 Experiment Real Data

In this study, Linear algebra test for university students is designed by author. The instrument consists of 19 dichotomous items which measure 6 concepts. The data set used in the experimental study is an educational data from university students in Taiwan. There are 231 university students from Taiwan in this test. The tool consist of six concepts, its contents are shown in Table 1.

Table 1. The content of Concepts

Classes	**Concepts**
1	Operation of matrix
2	System of linear equations
3	Determinants
4	Vector space and the property of R^n
5	Eigen-value and eigenvector
6	Geometry of linear algebra

Applying fuzzy clustering algorithms as we mentioned above, the clustering performances of each algorithm are calculated with same fuzzier m = 2 and the clustering accuracies are compared and shown in Table 2.

Table 2. The content of Concepts

Algorithm	Accuracies (%)
FCM	0.738
FCM-AM	0.823
FCM-NM	0.894

The Mean clustering Accuracies of 100 different initial value sets of FCM, FCM-AM and FCM-NM for the Dataset was shown in TABLE 2. From this table, we can find that the performance of FCM algorithm always worse than FCM-AM for above dataset. Although the performance of FCM-NM algorithm is better than which of FCM-AM algorithm in the dataset. In other words, our proposed two algorithms, FCM-NM and FCM-AM are better than FCM algorithm. Hence, the new algorithm, FCM-NM, has the best performance.

4 Conclusion

Clustering technique plays an important role in data analysis and interpretation. It 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. 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-NM 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

1. J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Springer Science & Business Media, 65-70 (1981)
2. H.C. Liu, J.M. Yih, D.B. Wu, and S.W. Liu, Fuzzy possibility c-mean clustering algorithms based on complete mahalanobis distances, International Conference on Wavelet Analysis and Pattern Recognition, **1**, 50-55 (2008)
3. R. Krishnapuram and J. Kim, A note on the Gustafson-Kessel and adaptive fuzzy clustering algorithm, IEEE Transactions on Fuzzy Systems, **7** (4), 453-461 (1999)
4. D.E. Gustafson and W.C. Kessel, Fuzzy clustering with a fuzzy covariance matrix, *Proc. IEEE Conf. Decision Contr. San Diego*, CA, 761-766 (1979)
5. I. Gath and A.B. Geva, Unsupervised optimal fuzzy clustering, IEEE Trans. Pattern Anal. Machine Intell, **11** (7), 773-781 (1989)
6. J.C. Dunn, A fuzzy relative of the isolated data process and its use in detecting compact, well-separated clusters, J. Cybern, **3** (3), 32-57 (1973)
7. F. Höppner, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis*, John Wiley and Sons (1999)
8. H.C. Liu, J.M. Yih, and S.W. Liu, Fuzzy C-means algorithm based on mahalanobis distances and better initial values, *Proceedings of the 10th Joint Conference& 12th International Conference on Fuzzy Theory & Technology*, **1**, 1398-1404 (2007)
9. H.C. Liu, J.M. Yih, D.B. Wu, and S.W. Liu, Fuzzy C-means algorithm based on "complete" Mahalanobis cistances, *Proceedings of International conference on Machine Learning and Cybernetics*, **7** (6), 3569-3574 (2008)
10. H.C. Liu, J.M. Yih, W.C. Lin, and T.S. Liu, Fuzzy C-means algorithm based on PSO and Mahalanobis distances, International Journal of Innovative Computing, Information and Control , **5** (12B), 5033-500 (2009)
11. H.C. Liu, J.M. Yih, W.C. Lin, and D.B. Wu, Journal of Multiple Valued Logic & Soft Computing, **15**, 581-595 (2009)
12. B. Balasko, J. Abonyi, and B. Feil, Fuzzy clustering and data analysis Toolbox for use with Matlab from http://www.mathworks.com/ matlabcentral/ fileexchange/7473 (2013)
13. H.C. Liu, B.C. Jeng, J.M. Yih, and Y.K. Yu, Fuzzy C-means algorithm based on standard mahalanobis distances, *Proceedings of the 2009 International Symposium on Information Processing* (*ISIP'09*), 422-427 (2009)
14. C. Ding, T. Li, and W. Ping, On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing, Computational Statistics and Data Analysis, **52** (8), 3913-3927 (2008)