PRINCIPLES AND APPLICATIONS OF FUZZY CLUSTERING ALGORITHMS

Zhang Hengrui, Yu.O. German, He Runhai
Department of Information Technologies in Automated Systems,
Belarussian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus
E-mail: 15058556211@163.com, jgerman@bsuir.by, fpm.he@bsu.by

This paper takes the fuzzy C-mean (FCM) clustering algorithm as an example, introduces the basic principles and applications of fuzzy clustering, and compares it with the K-mean algorithm to analyze the similarities and differences between the two. Fuzzy clustering allows data points to belong to multiple clusters at the same time, is suitable for dealing with fuzzy and overlapping data, and has a wide range of practical applications.

Introduction

Clustering algorithms are an important unsupervised learning method in data analysis and are usually categorized into soft and hard clustering. Hard clustering algorithms (e.g., K-mean clustering) require that each data point belongs to only one specific cluster, while soft clustering algorithms allow data points to belong to multiple clusters at the same time, reflecting their uncertainty. Fuzzy clustering is especially suitable for dealing with fuzzy mathematical phenomena, for example, when describing "today's weather is very hot", it is difficult to give a precise temperature range, which is where fuzzy algorithms come into action.[1]

In this paper, we will take the Fuzzy C-means (FCM) clustering algorithm as an example to systematically introduce the basic principles of fuzzy clustering and its applications, and compare it with the K-means algorithm to analyze the similarities and differences between the two in processing data. Through this comparison, it aims to clarify the advantages of fuzzy clustering in practical applications, especially how to reflect the complexity of data more effectively when dealing with data with overlapping and fuzzy boundaries.

I. Fuzzy c-means clustering algorithm

Fuzzy C-means (FCM) clustering algorithm is a soft clustering method widely used in the field of data mining and pattern recognition. Its main goal is to classify a given dataset into c fuzzy clusters which optimize the distance between data points and cluster centers by minimizing an objective function. The algorithm uses an iterative approach to continuously update the cluster centers and the affiliation of the data points until convergence.

The steps of the FCM algorithm are as follows[2]:

Step1. Initialize the clustering center: Random selection c data points as initial clustering centers. This selection has a significant impact on the final clustering results, so different initialization strategies can be used to improve stability.

Step2. Calculate the affiliation matrix U:The affiliation matrix U is an $n * c$ matrix, where n is the number of data points and c is the number of clusters. Randomly assign affiliation values to each data point such that the sum of the affiliations of each data point to all clusters is 1. The affiliation degree is calculated as:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - V_j\|}{\|x_i - V_k\|} \right)^{\frac{2}{m-1}}} \tag{1}$$

Where u_{ij} means the affiliation of the data point x_i in the clustering V_j and m is a parameter to control the degree of fuzzy.

Step3. Updating Cluster Centers: Calculate the new clustering center based on the values of the affiliation matrix and the data points, The new clustering center is calculated as:

$$V_j = \frac{\sum_{i=1}^{n} u_{ij}^m \cdot x_i}{\sum_{i=1}^{n} u_{ij}^m} \tag{2}$$

Step4. Iteration: Repeat steps 2 and 3 until the transformation of the affiliation or cluster center is less than a preset threshold value. The objective function for the convergence of the affiliation matrix is the minimization objective function J, which takes the following form:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^m \cdot d_{ij}^2 \tag{3}$$

Where d_{ij} means the distance between the data point x_i and the clustering center V_j.

II. Comparison with k-means algorithm

In K-means, which is usually applied when the data are clearly separated, the results are simpler and clearer. While Fuzzy C-means is suitable for dealing with fuzzy and overlapping data, and can better reflect the diversity and complexity of the data.[3]

In Figure 1 are shown the results obtained after using K-means and Fuzzy C-means clustering algorithms separately for the same dataset.

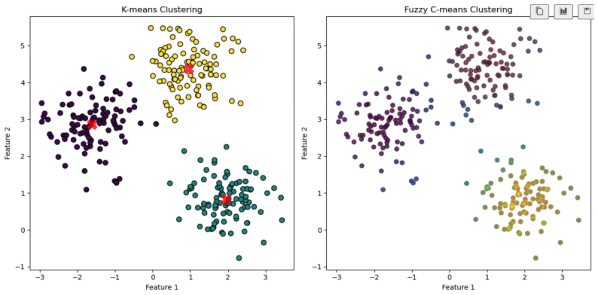

Рис. 1 – Comparison with K-means

This can be clearly seen through the resultant graph: the clustering centers of the K-means, while the Fuzzy C-means shows the affiliation of the data points in different colors.

Fuzzy C-means allows data points to belong to multiple clusters, which is particularly effective for dealing with overlapping or ambiguous data. Visually, it can be seen that the distribution of data points may not be clear and there is a lot of overlap. Its clustering results may be more realistic, especially when the data is not completely separated, and Fuzzy C-means is better at capturing the fuzzy boundaries of the data points.[4]

III. Application of fuzzy clustering algorithm

Fuzzy Clustering has a wide range of real-world applications in several fields. This section will describe some of the major application scenarios of Fuzzy Clustering.

1. Detection of crime hot spots: Using algorithms such as Fuzzy C-means, crime events can be clustered into specific areas to identify crime hot spots. These hot spots are likely to be areas with high crime rates, helping law enforcement allocate resources more efficiently.[5]

2. Medical Image Processing: In medical image processing, it can be used to segment tumor tissue from healthy tissue.[6]

3. Risk assessment in the financial sector: In the field of finance, fuzzy clustering can help predict and assess the risk of financial markets such as stocks.[7]

4. Text Topic Recognition: In Natural Language Processing (NLP), fuzzy clustering is able to deal with the ambiguity of textual data and identify the topic of an article or document.[8]

In summary, the versatility of fuzzy clustering makes it a valuable tool across various domains, enabling more effective decision-making and resource management in complex and uncertain environments.

Summary

Fuzzy clustering is an advanced cluster analysis method designed to effectively deal with ambiguity and uncertainty in data. Unlike traditional hard clustering methods, fuzzy clustering allows data points to belong to multiple clusters simultaneously with different degrees of affiliation. This property allows fuzzy clustering to more accurately reflect the complexity of data in the real world, especially when dealing with fuzzy or overlapping boundaries, providing more detailed and rich clustering information. Therefore, fuzzy clustering has been widely used in several fields, such as image processing, medical diagnosis and market analysis.

In addition, fuzzy clustering not only plays an important role in data analysis, but also provides a powerful tool for decision support. By analyzing historical data with fuzzy clustering, decision makers are able to identify potential risks and key issues, and thus develop more effective strategies. For example, in the financial sector, fuzzy clustering can be used for customer risk assessment to help financial institutions optimize resource allocation and risk control. In terms of visualization, fuzzy clustering results can be presented through various graphical tools, which enhances the intuition and comprehensibility of data analysis, and at the same time provides an important basis for the prediction of future trends. Therefore, fuzzy clustering shows its unique value and advantages in processing complex data and supporting decision-making.

1. Kruse, R., Döring, C., and Lesot, M. J. (2007). Fundamentals of fuzzy clustering. Advances in fuzzy clustering and its applications, 3-30.
2. Li, J., and Lewis, H. W. (2016, November). Fuzzy clustering algorithms—review of the applications. In 2016 IEEE International Conference on Smart Cloud (SmartCloud) (pp. 282-288). IEEE.
3. Dubey, A. K., Gupta, U., and Jain, S. (2018). Comparative study of K-means and fuzzy C-means algorithms on the breast cancer data. International Journal on Advanced Science, Engineering and Information Technology, 8(1), 18-29.
4. Velmurugan, T., and Santhanam, T. (2010). Performance evaluation of k-means and fuzzy c-means clustering algorithms for statistical distributions of input data points. European Journal of Scientific Research, 46(3), 320-330.
5. Grubesic, T. H. (2006). On the application of fuzzy clustering for crime hot spot detection. Journal of Quantitative Criminology, 22, 77-105.
6. Ren, T., Wang, H., Feng, H., Xu, C., Liu, G., and Ding, P. (2019). Study on the improved fuzzy clustering algorithm and its application in brain image segmentation. Applied Soft Computing, 81, 105503.
7. Mo, H., Niu, Y., and Zhang, Y. (2015). Application of parallel clustering algorithms for big data in the division of stock. Big Data Research, 1(4), 9-17.
8. Anam, S. A., Rahman, A. M., Saleheen, N. N., and Arif, H. (2018, June). Automatic text summarization using fuzzy c-means clustering. In 2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR) (pp. 180-184). IEEE.