Using fuzzy C-means Clustering to Identify Heavy Metal Polluted Soil in a Certain Area of Shanghai

Yuhan Xu, Luolei Zhang, Peng Yu, Chongjin Zhao
State Key Laboratory of Marine Geology, Tongji University, 1239 Siping Road, Shanghai 200092, China

Abstract. Fuzzy clustering is an important branch of fuzzy pattern recognition, which has been widely used in many fields such as data mining, image processing, big data analysis and so on. Among many fuzzy clustering algorithms, fuzzy c-means algorithm is the most widely used. In this paper, the fuzzy c-means clustering is applied to the identification of polluted soil. To solve the problem of determining the optimal number of clusters for this method, the method we choose is obtaining an initial clustering result firstly and then merging. Based on the important characteristic of fuzzy c-means method that the objective function of fuzzy c-means clustering decreases rapidly as the number of clusters increases, and the rate become slow after exceeding the optimal number of clusters, we choose the clustering result whose objective function is reduced to a certain degree as the initial clustering result. Then use the parameter estimation method in statistics to estimate the distance between classes, determine the reasonable range of distance between classes, combine the initial classification results, and finally perform the final clustering results according to the clustering validity function. Evaluation and experiments prove the feasibility and effectiveness of the scheme.

1. Introduction
With the continuous development of China's economy, environmental pollution is increasing. In order to meet the people's needs for a better living environment, the construction of ecological civilization is speeding up. Therefore, it is urgent for the market to identify the abnormal characteristics of polluted soil[1].

Fuzzy clustering analysis is an analysis method that cluster objective things by establishing fuzzy similarity relationship according to the characteristics, degree of intimacy and similarity between them[2,3]. There are many algorithms to realize clustering analysis, among which fuzzy C-means (FCM) algorithm is the most widely used and successful one. The core idea of this algorithm is to get the membership degree of each sample point to all class centers by optimizing the objective function, and then determine the ownership of sample points by membership degree, so as to achieve the purpose of automatic classification of sample data.

2. Algorithm

2.1. Related concept
Fuzzy c-means algorithm is widely used in many fuzzy clustering algorithms. It is developed from hard c-classification[4]. For the data set \( X = \{x_1, x_2, \ldots, x_n\} \) composed of \( n \) samples, if the predetermined number of classificaitons is \( c \), the set \( V = \{v_1, v_2, \ldots, v_c\} \) represents the clustering center of each classification, then the objective function of hard c-classification can be expressed as:
\[ J = \sum_{i=1}^{c} \sum_{y \in \Gamma_i} \| y - v_i \|^2 \]  

(1)

Based on the hard c-classification, Dunn introduced the concept of membership degree to represent the degree of a data belonging to a certain category. If \( u_j(x_i) \) is the membership function of the i-th sample to the j-th class, and \( u_j(x_i) \) satisfies:

\[ \sum_{j=1}^{c} u_j(x_i) = 1, i = 1, 2, \ldots, n \]  

(2)

Then the objective function of fuzzy c-means defined by membership degree is:

\[ J_e = \sum_{i=1}^{c} \sum_{j=1}^{n} [u_j(x_i)]^b \| x_i - v_j \|^2 \]  

(3)

Among them, \( b \) is a constant greater than 1 that controls the degree of ambiguity of the clustering result. When the objective function gets the minimum value, the clustering result is the best. According to the Lagrange multiplier method, the necessary conditions for obtaining the minimum value of \( J_e \) from equations (2) and (3) are as follows:

\[ m_j = \frac{\sum_{i=1}^{n} [u_j(x_i)]^b x_i}{\sum_{i=1}^{n} [u_j(x_i)]^b}, j = 1, 2, \ldots, c \]  

(4)

\[ u_j(x_i) = \left( \frac{1}{\| x_i - v_j \|^2} \right)^{1/(b-1)}, i = 1, 2, \ldots, n \]  

(5)

2.2. Principle of algorithm

The process of fuzzy c-means clustering is given below.

Step 1: Set parameters

The parameters need to be set include the parameter \( b \), which controls the fuzzy degree of clustering results, the current number of clusters \( c=1 \), the maximum number of clusters \( c_{\text{max}} \).

Step 2: Initialize cluster center

When the number of cluster is 1, "density method" is used to select the first cluster center. The so-called "density" refers to the total number of samples falling into a sphere with each sample as the center of the sphere, making a sphere neighborhood according to a given radius. After calculating the "density" of each sample, select the point with the largest "density" as the first clustering center. In order to avoid clustering centers together, calculate the distance between each point in the class and the clustering center, and take the point with the largest distance as the next cluster center.

After that, each new clustering center is the sample point with the largest distance from the clustering center under the number of previous clusters and the largest sum of square errors within the cluster.

Step 3: Calculate membership function and update cluster center according to membership function

The initial clustering center in step 2 is used to calculate the membership function according to formula (5), and the clustering center is updated according to the membership function until the membership value of each sample is stable or reaches the upper limit of iteration times.

Step 4: Calculate \( J_{ec} \)

According to formula (3), calculate the \( J_{ec} \) under the current number of clusters and judge: if the value of \( J_{ec} \) drops below 99% of the value of \( J_{e1} \), take the current clustering result as the initial clustering result; otherwise, repeat step 2 – step 4.

After getting the initial clustering results, calculate the Xie-Beni effectiveness index \( V_{\text{Xie}} \) according to formula (6) to evaluate the effectiveness of clustering.
\[ V_{Xie}(U, V, C) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} u_{ij} \|v_i - x_j\|^2 }{ \min_{i\neq j} \|v_i - v_j\|^2 } \] (6)

Step 5: Combine and evaluate the initial clustering results

Calculate the distance \( X \) between all classes. Since \( X \) satisfies the normal distribution \( N(\mu, \sigma^2) \), choose the maximum likelihood method to estimate the parameters of \( \mu, \sigma^2 \). Then compare the relationship between \( X_{\text{min}} \) and \( \mu - \sigma \): if \( X_{\text{min}} < \mu - \sigma \), merge the corresponding two classes, take the average value of two classes of elements as the new clustering center, and calculate the distance between all the elements in the corresponding two classes and all the clustering centers, classify the elements into the nearest class, and then repeat step 5 for judgment: if \( X_{\text{min}} > \mu - \sigma \), consider the final clustering result as the most reasonable one.

After class merging, calculate \( V_{Xie} \) again. Since the indicator molecule represents the degree of tightness within the class, the molecule represents the degree of separation between classes. If the value drops, the combination is valid.

2.3. Algorithm instance

Based on the clustering process above, we have identified the inversion results of the measured data of four high-density lines. The results are as follows:

(1) Line 1

![Figure 1. Inversion result of Line 1](image1)

![Figure 2. Initial clustering result of Line 1](image2)

The number of clusters in the initial clustering results is 10, and the value of validity index \( V_{Xie} \) is 26.3905.

![Figure 3. Final clustering result](image3)

After merging, the final clustering result are 4 categories, and the value of validity index \( V_{Xie} \) is 0.0722.

(2) Line 2
Figure 4. inversion result of Line 2

The number of clusters in the initial clustering result is 8, and the value of the validity index $V_{\text{Xie}}$ is 18.5565.

Figure 5. initial clustering result of Line 2

The final clustering result after merging is 5 categories, and the value of the validity index $V_{\text{Xie}}$ is 0.0754.

(3) Line 3

Figure 7. inversion result of Line 3

The number of clusters in the initial clustering result is 11, besides, the value of validity index $V_{\text{Xie}}$ is 22.3459.
After merging, the final clustering result are 4 categories, and the value of validity index $V_{Xie}$ is 0.0720.

(4) Line 4

The number of clusters in the initial clustering result is 10. The value of the validity index $V_{Xie}$ is 23.8386.

Compared figure 1 and 2, figure 4 and 5, figure 7 and 8, Figure 10 and 11, it can be seen that there are many unreasonable classifications due to the excessive number of initial clusters. The high value of $V_{Xie}$ also shows that the preliminary classification is not reasonable. After merging, the decline of $V_{Xie}$ proves the rationality of merging. Meanwhile, the final clustering results such as figure 3, figure 6, Figure 9 and figure 12 clearly show that the spatial distribution of pollutants is consistent with the inversion results.

3. Conclusions

For the intelligent discrimination of pollutants and abnormal bodies by applying fuzzy C-means clustering, the method we use is to first determine the initial clustering result according to the degree of reduction of the objective function, and then to estimate the parameters based on the distance between all classes of the initial clustering result. Then merge item by item and use Xie-Beni validity
index $V_{sir}$ to evaluate the clustering results. The results of clustering the inversion results of high-density measured data with this method have proved the correctness of the method.

4. Acknowledgments
The authors express their sincere thanks for support from the Shanghai Science and Technology Commission(18DZ1204200), and the Fundamental Research Funds for the Central Universities (22120180529).

References
[1] Bian Z, Zhang X Pattern recognition (Second Edition) [M] Tsinghua University Press 2000
[2] Tang M, Yang Y Research progress on the validity of fuzzy clustering [J] Computer engineering and science 2009 (09) 126-128
[3] Zheng J, Wang W and Yang W Text clustering evaluation method based on the estimation of distance parameter between classes [J] Computer Engineering 2009 35 (9) 37-39
[4] Xie X, Beni G A validity measure for fuzzy clustering[J] IEEE Transactions on Pattern Analysis and Machine Intelligence 1991 13(8) 841-847