Identification of Traditional Chinese Medicine Based on KNN Algorithm and Random Forest

Yanning Li*

Tianjin Electronic Information College 300350, Tianjin, Chine

*Corresponding author: jueyuanti291@163.com

Abstract. Medicinal materials are a system of components and complex mixtures, and spectroscopic principles can provide in-depth analysis of the composition mechanism of traditional Chinese medicinal materials through the determination of their material structures. Chinese medicinal materials of different origins and varieties exhibit different spectral characteristics due to differences in chemical composition and organic matter. However, spectral features have high-dimensional attributes, so PCA principal component dimensionality reduction is considered to condense spectral features into representative feature variables. Then, a hierarchical clustering method with a clear hierarchy can be used to classify multiple Chinese medicinal materials into three categories based on their spectral characteristics; At the same time, in order to solve the problem of origin identification, based on the situation that there are many classifications of origins, the KNN algorithm is combined to achieve the requirements of origin identification; Using both mid infrared and near infrared spectral data and using KNN algorithm to verify the origin of Chinese medicinal materials is more accurate; Due to the small number of categories, the prediction of the types of Chinese medicinal materials is implemented using commonly used random forests. The realization of the above methods demonstrates the feasibility of identifying Chinese medicinal materials through infrared spectroscopy, and it is also worth further exploration and research.

Keywords: Principal component analysis, Hierarchical clustering, Random forest.

1. Introduction

As a unique treatment method in China, traditional Chinese medicine has shown excellent efficacy in multiple clinical fields. However, there are many types of traditional Chinese medicine, and there may be certain similarities between different types of traditional Chinese medicine in terms of odor, morphology, and taste, which poses a great challenge to the identification of traditional Chinese medicine types. At the same time, the authenticity of Chinese medicinal materials is mainly determined by the origin, and the identification of the origin is particularly important for the quality identification of medicinal materials. Therefore, the identification of traditional Chinese medicine, which is a crucial link in the scientific control and management of the quality of traditional Chinese medicine, has always been a hot topic in the field of traditional Chinese medicine, regardless of its type or origin. With the gradual improvement of modern technology, spectroscopy has gradually become a common method for the identification of traditional Chinese medicine with its high efficiency and accurate recognition efficiency. The key factor that enables the identification of traditional Chinese medicine using spectroscopy is that, on the one hand, different types of traditional Chinese medicine have significant differences in spectral characteristics, which can be further analyzed through spectral data to achieve the identification of the type of traditional Chinese medicine. On the other hand, due to the different growth environments of the same medicinal material from different origins, there may be differences in the chemical composition and organic matter of inorganic elements, which makes the same medicinal material exhibit different spectral characteristics during spectral irradiation. However, there is a problem that the spectra of the same medicinal material from different origins are relatively close within the same wavelength band, which undoubtedly increases the difficulty of spectral origin identification and increases the error of correct identification. In this case, the comprehensive verification of multi type spectral data plays a unique role. Some Chinese medicinal materials have obvious near-infrared differences, while others have obvious mid-infrared differences. Therefore, it is possible to comprehensively verify and identify the

origin of Chinese medicinal materials from both near-infrared and mid-infrared spectral data. Based on the different characteristics of spectral methods in the identification of traditional Chinese medicine, this study aims to explore the feasibility and rationality of their use in the identification of traditional Chinese medicine through example data.

2. Requirement analysis

2.1. Identification of the origin of the same medicinal material

Based on the provided mid infrared spectral data of a certain medicinal material, analyze the characteristics and differences of medicinal materials from different origins, identify the origin of the medicinal material, and fill in the table with the identification results of the number of medicinal materials given in the following table.

No	3	14	38	48	58	71	79	86	89	110	134	152	227	331	618
OP															

2.2. Identification of the same medicinal material origin based on different spectral data

Identify the origin of a certain medicinal material based on the provided near-infrared and mid-infrared data, and fill in the identification results of the medicinal material origin with the numbers given in the table below.

No	4	15	22	30	34	45	74	114	170	209
OP										

2.3. Double identification of different medicinal materials and origin

Based on the near infrared spectral data of several medicinal materials, try to identify the category and origin of the medicinal materials, and fill in the identification results of the medicinal materials with the numbers given in the table below.

No	94	109	140	278	308	330	347
Class							
OP							

3. Theoretical model

Based on the differences in spectral characteristics of Chinese medicinal materials, the clustering idea in unsupervised learning can be used to automatically classify the types of medicinal materials according to the differences reflected in the data itself. At the same time, through preliminary exploration of the data, it is found that according to the high dimensional and low sample situation of the example data, we adopt corresponding dimensionality reduction processing for the data. As a preprocessing method for high-dimensional data, dimensionality reduction preserves the most important features of high-dimensional data, removes noise and unimportant features, and achieves the goal of improving data processing speed. In this study, principal component analysis (PCA), which is widely used in dimensionality reduction methods, was used. As a linear dimensionality reduction method, PCA not only can perform data conversion between original data and dimensionality reduction data through a certain functional relationship, but also can ensure the comprehensive utilization of dimensionality reduction information based on the degree of extraction of dimensionality reduction information. Clustering algorithm is proposed to identify the types and origins of medicinal materials. At the same time, targeted methods are selected based on the different needs of the problem. The main methods used are hierarchical clustering, nearest neighbor algorithm, and random forest.

3.1. Principal component analysis

As a statistical method for data dimensionality reduction, principal component analysis can achieve dimensionality reduction while retaining the features that contribute the most to the data. It is a process of combining many previously relevant indicators into a new set of unrelated comprehensive indicators to replace the original indicators. That is, to reveal the internal structure of multiple variables through a few principal components, while retaining as much information as possible about the original variables, and not related to each other [1]. It mainly relies on orthogonal transformation to convert the original random vector whose components are related to each other into a new random vector whose components are not related, and then performs dimensionality reduction processing on a multidimensional variable system to enable it to be converted into a low-dimensional variable system with a higher accuracy. Then, by constructing appropriate functions, the low-dimensional system is further converted into a one-dimensional system [2]. The algorithm flow is described as follows:

Assume that there are n samples, each with p variables, forming the data matrix X of $n \times p$.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \tag{1}$$

Centralize all samples first

$$x^{(i)} = x^{(i)} - \frac{1}{m}\sum_{j=1}^{m} x^{(j)} \tag{2}$$

Then calculate the covariance matrix XX^T of sample X

$$XX^T = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{np} \end{pmatrix} \tag{3}$$

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})(x_{ij}-\overline{x_j})}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})^2 \sum_{k=1}^{n}(x_{ij}-\overline{x_j})^2}} \tag{4}$$

The eigenvalues $(\lambda_1, \lambda_2, \cdots \lambda_p)$ and unit eigenvectors $(a_1, a_2, \cdots a_p)$ of the covariance matrix XX^T will be calculated to construct the principal component F_i

$$a_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{bmatrix} \quad a_2 = \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{p1} \end{bmatrix} \quad a_p = \begin{bmatrix} a_{p1} \\ a_{p1} \\ \vdots \\ a_{p1} \end{bmatrix} \tag{5}$$

$$F_i = a_{1i}\lambda_1 + a_{2i}\lambda_2 + \cdots a_{pi}\lambda_p \; i = 1,2, \ldots p \tag{6}$$

Rank the feature values in descending order, select the top k in the ranking, and calculate their principal component contribution rate and cumulative contribution rate. Generally, the first, second,..., and m principal components F_m (m<p) corresponding to the feature value $\lambda_1, \lambda_2, \cdots \lambda_m$ with a cumulative contribution rate of 80% to 95% are taken.

$$\text{Where, Contribution rate} = \frac{\lambda_i}{\sum_{i=1}^{p}\lambda_k} (i=1,2,\ldots p) \tag{7}$$

$$\text{Cumulative contribution rate} = \frac{\sum_{k=1}^{i}\lambda_k}{\sum_{i=1}^{p}\lambda_k} (i=1,2,\ldots p) \tag{8}$$

The selection of principal components here is based on the maximum variance theory. The larger the variance, the greater the amount of information. Each feature vector of the covariance matrix is a projection surface, and the corresponding eigenvalue of each feature vector is the variance of the

original feature projected onto this projection surface. The original features are projected onto these feature vectors, and the projected values are the new principal components.

3.2. Hierarchical clustering method

Hierarchical clustering refers to clustering one layer at a time. Its basic idea is to calculate the similarity between nodes through a certain similarity measure, and sort the two data points that are most similar among all data points according to the similarity degree from high to low. The process of combining different categories of clusters is to determine the similarity between them by calculating the distance between the data points of each category and all data points, The smaller the distance, the higher the similarity, and combine the two closest category clusters to form the final cluster [3]. Hierarchical clustering mainly includes two major categories of algorithms: split method and aggregation method. This article uses the aggregation method, which is a bottom-up strategy. The algorithm flow is described as follows:

Firstly, all sample points in the sample set are treated as a mutually independent cluster; Then calculate the distance between any two types of clusters, find the two types of clusters C1 and C2 with the smallest distance, and merge C1 and C2 into a single type of cluster. The Euclidean distance is used to calculate the distance between different types of data points as an indicator of similarity. The Euclidean distance formula is:

$$D_{European} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \tag{9}$$

Repeat the second step above until all sample points[4] are included in a cluster as shown in Fig 1 .

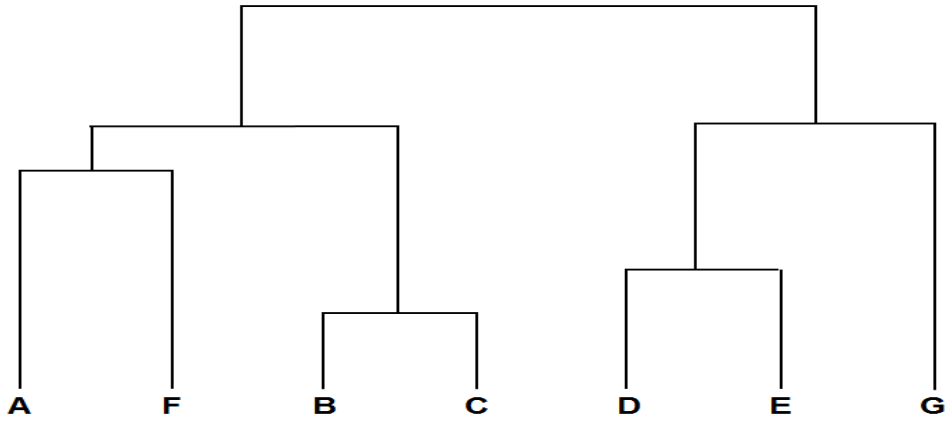

Fig. 1 Schematic diagram of hierarchical clustering

3.3. KNN algorithm

The KNN algorithm is known as the K-nearest neighbor algorithm, which measures the distance between different feature values for classification. Its algorithm idea is that if most of the k most similar samples in the feature space belong to a certain category, the sample also belongs to this category (where K is usually an integer not greater than 20) [5]. In the actual algorithm operation, the KNN algorithm can input test data when the training set data and tags are known, compare the characteristics of the test data with the corresponding characteristics in the training set, and find the first K data in the training set that are most similar to them. The corresponding category of the test data is the category that appears the most frequently among the K data [6], Specifically, first, calculate the distance between the test data and each training data; Then, the K points with the smallest distance are sorted according to the increasing relationship of distance; Then determine the frequency of occurrence of the category of the first K points; Finally, the category with the highest frequency among the first K points is returned as the prediction classification of the test data as shown in Fig 2.

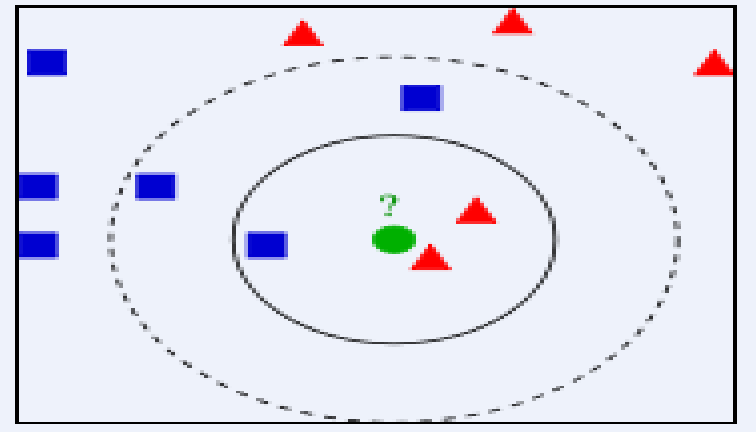

Fig. 2 Example of KNN algorithm

3.4. Random forest

Random forest is an algorithm that integrates multiple decision trees through the idea of ensemble learning. Among them, integrated learning refers to the construction and combination of multiple learner training models, which can achieve significantly superior generalization performance than a single learner [7]. Random forests are mainly classifiers that integrate multiple decision trees, and the final output category is determined based on the mode of each individual decision tree output category, that is, the principle that the minority follows the majority. The principle process can be summarized as two stages: sample randomization and feature randomization. Specifically:

Random samples: Assume that the training dataset contains a total of M object data, and randomly select N samples from the sample data by using a Bootstrap method. The samples taken each time are not identical. These samples form the training dataset for the decision tree.

Feature Randomness: Assuming that each sample data has K features, randomly select k features (k<K) from all features, and select the best segmentation feature as a node based on node segmentation indicators such as information gain or Gini index to establish a CART decision tree. Then repeatedly repeat the previous steps to establish m CART trees to form the final forest (these trees are required to be fully grown and not pruned). The prediction category required for the problem is determined by voting based on the prediction results of m trees as shown in Fig 3.

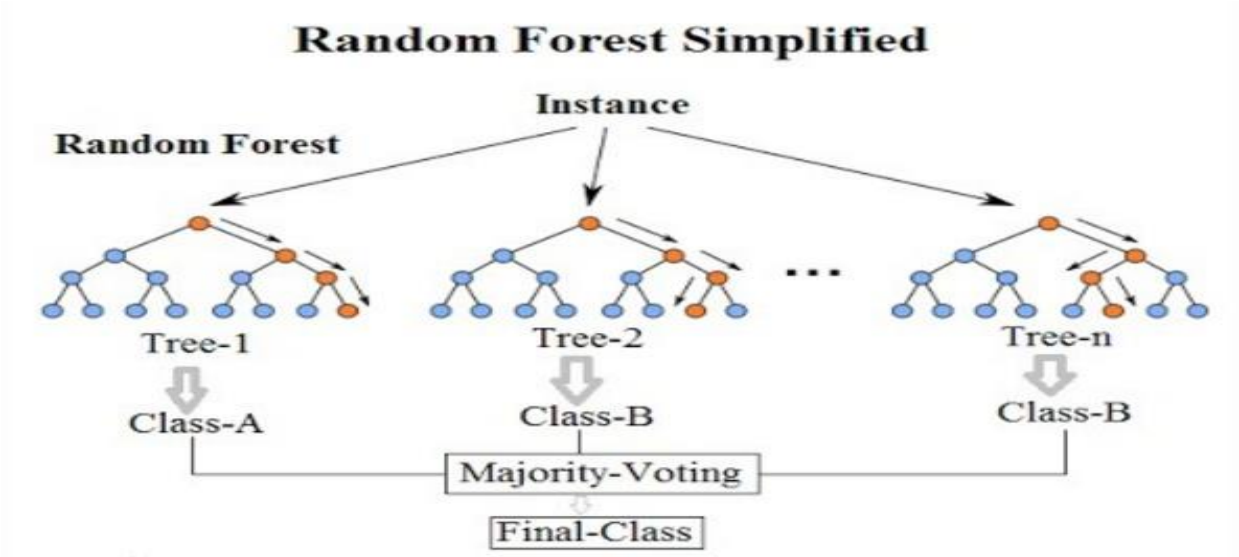

Fig. 3 Example of a random forest

4. Model establishment

4.1. Analyze the characteristics and differences of medicinal materials from different habitats

4.1.1 PCA dimensionality reduction

According to the task requirements, this research plan uses a combination of PCA dimensionality reduction and KNN nearest neighbor algorithm to meet the above requirements. The main reasons for consideration are: ① excessively high dimensions are not conducive to subsequent feature analysis; ② There are 11 classifications of origin, which are not suitable for prediction algorithms and have low accuracy. However, the KNN algorithm assigns a label to the unlabeled Chinese medicinal material node by finding the closest k labeled "neighbor" points in space[8]. In this way, the origin label information of the above medicinal materials can be obtained more accurately.

Repeat the process in 1 and first find the optimal number of dimensionality reducing principal components through the gravel map:

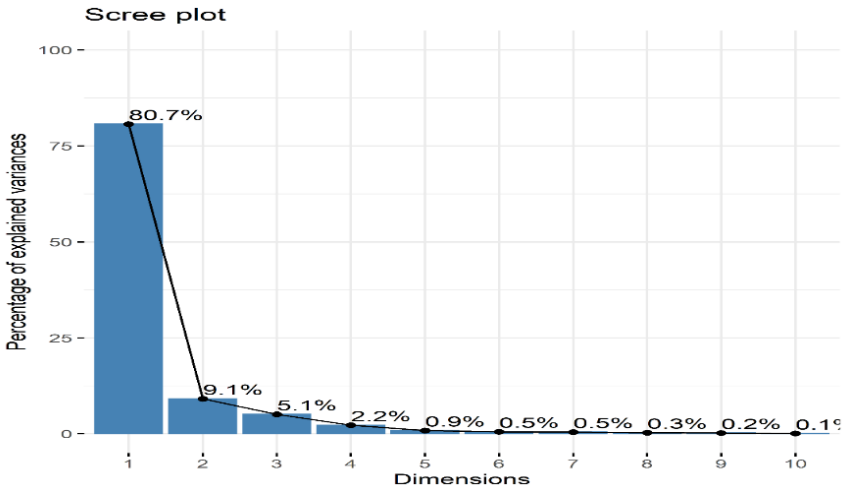

Fig. 4 Infrared spectrum PCA gravel map

As can be seen from Fig 4, there are four components with a contribution degree greater than 1%, and the cumulative contribution degree reaches 97.1%. In the study of this question, the analytic hierarchy process was used to analyze whether there are still outliers of traditional Chinese medicine nodes

4.1.2 KNN nearest neighbor analysis

Use the kknn function in the kknn package of the R language to label the nearest neighbor found in the unlabeled dataset (multiple neighbors can also be specified, and the specified one will be studied here) [9].

```
str(data2_clust)
train_set1<-data2_clust[!is.na(data2_clust$OP),-c(9,10)]
test_set1<-data2_clust[is.na(data2_clust$OP),-c(9,10)]
install.packages("kknn")
library(kknn)
bmt2_kknn<-kknn(OP~.,train=train_set1,test=test_set1,
      distance=1,kernel="triangular")
summary(bmt2_kknn)
fit<-fitted(bmt2_kknn)
fit
bmt2_test_predict<-cbind(test_set1,fit)
bmt2_test_predict
```

The final prediction obtained the label data of 15 unlabeled Chinese herbal medicine nodes, and the results are shown in Table .1:

Table 1. Prediction Results of 15 Producing Areas of Chinese Medicinal Materials

No	**3**	**14**	**38**	**48**	**58**	**71**	**79**	**86**	**89**	**110**	**134**	**152**	**227**	**331**	**618**
OP	6	8	10	6	6	1	1	3	3	6	10	2	2	4	1

4.1.3 Try to identify the origin of this medicinal material based on its near-infrared and mid-infrared data

No	4	15	22	30	34	45	74	114	170	209
OP										

The research in this part draws on the idea of PCA+KNN in the above section. PCA dimensionality reduction is performed on near-infrared and mid-infrared data, respectively. Then, the dimensionality reduction data of the two parts are combined and the KNN algorithm is used to obtain the origin identification information of unmarked medicinal materials[10].

4.1.4 Near-infrared PCA dimensionality reduction results

PCA dimensionality reduction is performed on near-infrared data to obtain a gravel map as shown in Fig 5:

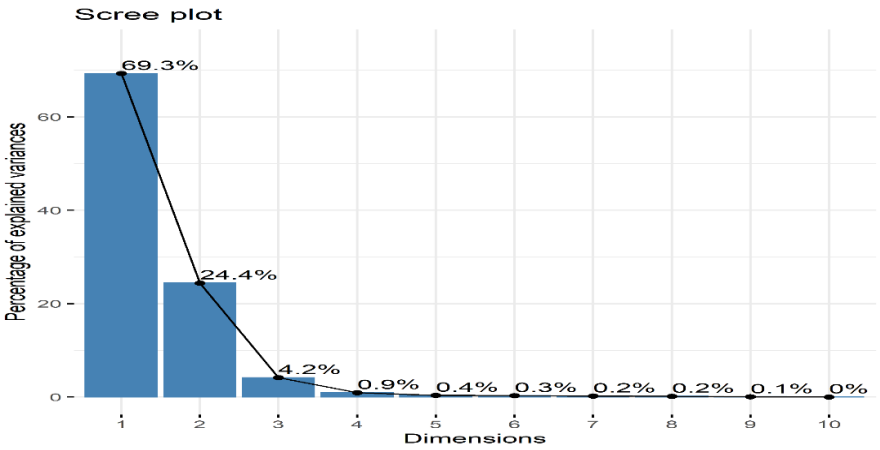

Fig. 5 Near infrared spectrum PCA gravel map

4.1.5 Middle infrared PCA dimensionality reduction results

Perform PCA dimensionality reduction on near-infrared data to obtain a gravel map as shown in Fig 6:

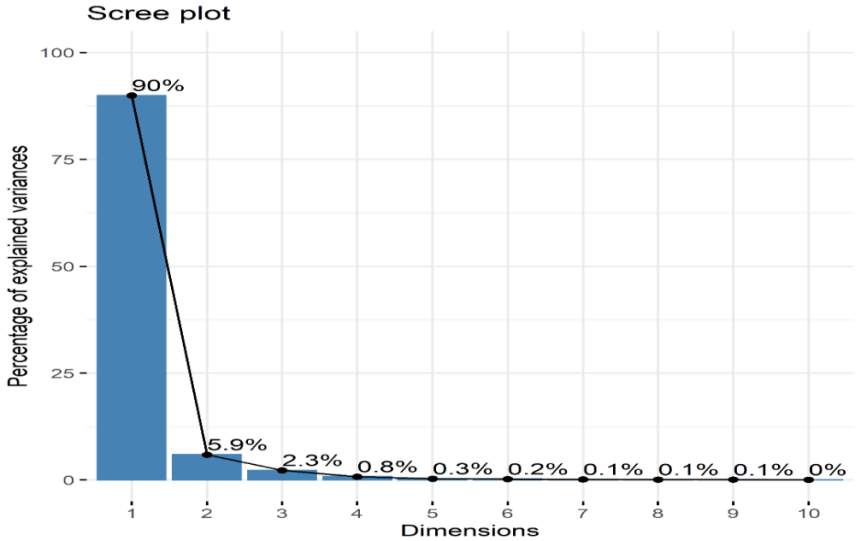

Fig. 6 Middle infrared spectrum PCA gravel map

4.1.6 KNN nearest neighbor analysis

Among the 255 nodes of traditional Chinese medicine, 10 nodes do not have labels. Finding the most adjacent node to assign labels to these 10 nodes is a quick and effective method. The results obtained are shown in Table 2:

Table 2. KNN Algorithm for Predicting Place of Origin Results

No	**4**	**15**	**22**	**30**	**34**	**45**	**74**	**114**	**170**	**209**
OP	15	1	1	2	16	3	4	10	9	10

5. Conclusion

The main research method for identifying traditional Chinese medicine currently being studied is cluster analysis of spectral curves of medicinal materials. Cluster analysis methods can better "understand" the characteristics of the constituent features of medicinal materials. However, future research should provide statistical information on which spectral features of medicinal materials have high commonalities or reflect the same regularity, in order to identify these medicinal materials as the same type of medicinal materials. In current research, existing clustering algorithms are widely used due to their advantages such as simple principles, easy programming, and fast running speed. In recent years, intelligent colony algorithm has been introduced into pattern recognition research, and has achieved good application results. However, both classic K-means algorithms and intelligent optimization algorithms such as Particle Swarm Optimization (PSO) have the disadvantages of being sensitive to the initial clustering center and being prone to falling into local optimums, resulting in low clustering quality and poor stability of clustering results. In the future, it should focus on solving these problems in models.

Reference

[1] Dimension reduction and visualization in principal component analysis. Anal Chemistry 2008:80 (13): 4933-4944.

[2] Beattie JR. Esmonde White FWL Principal Component Analysis: Using spectroscopy to intuitively derive principal component analysis pectrosc2021:75(4):361-375. doi: 101177/0003702820987847

[3] Bu Jun, Liu Wei, Pan Zhen, Ling Kun. Comparative study on hydrochemical classification based on Dif é rent clustering analysis method. International Journal of Environment and Public Health. 2020:17(24):9515. Issued on December 18, 2020.

[4] Research on Liu Xingbo's Cohesive Hierarchical Clustering Algorithm [J. Science and Technology Information (Science Teaching and Research), 2008 (11): 202

[5] Geng Lijuan, Li Xingyi Research on KNN Algorithm for Big Data Classification [J]. Computer Application Use Research, 2014,31 (05): 1342-1344+1373.

[6] Dou Xiaofan. Overview of KNN Algorithm [J]. Communication World, 2018 (10): 273-274.

[7] Comprehensive genetic and epigenetic prediction of coronary artery type Y diabetes in Framingham radiotherapy study, PLoS One.2018; 13(1): e0190549

[8] Le, S., Josse, J.Husson, F.(2008) FactoMineR: The R package for multivariate analysis. Journal of Statistical Software. 25(1)

[9] Hechenbichler K, Schliep K. Weighted K-nearest neighbor technique and ordered classification [J]. Discussion Paper Sfb, 2004.

[10] Breiman, L. (2001), Random Forest, Machine Learning 45 (1), 5-32