Study on Identification Method of Chinese Herbal Medicine based on Infrared Spectroscopy Characteristics

Shiyi Tang, Jiahao Jue, Yifan Lin

Shenzhen Senior High School, Shenzhen 518000, China

Abstract. The identification of Chinese herbal medicines is a key issue in the field of traditional Chinese medicine. Based on the characteristics of Chinese herbal medicines, the classification of types, producing areas, and quality can be realized. However, traditional identification methods of Chinese herbal medicines mainly rely on manual identification methods, which requires a lot for identification personnel with low efficiency. To solve this problem, we study the intelligent method of identification of Chinese herbal medicines by using data of infrared spectroscopy characteristic. To solve this problem, this paper studies the classification of spectral characteristic data of Chinese herbal medicines from unsupervised and supervised learning. Firstly, an improved K-means clustering algorithm based on Gaussian distribution model is established for unsupervised spectral classification of Chinese herbal medicines. This method "over-classifies" the sample data by K-means clustering algorithm, and further classifies the data by Gaussian mixture model, thus realizing unsupervised classification of Chinese herbal medicines. Secondly, aiming at the supervised classification and recognition of Chinese herbal medicines, an improved discriminant analysis classification method based on Gaussian distribution is established to identify different kinds and producing areas of Chinese herbal medicines. Finally, we test our method on two sets of data with and without tagged information, with Chinese herbal medicines in two data sets identified respectively. The experimental results fully verify the effectiveness of the method, especially in the supervised identification of Chinese herbal medicines. We have proved the effectiveness of our designed model through the comparison of various methods and extensive tests.

Keywords: Infrared Spectroscopy Characteristics; Classification of Chinese Herbal Medicines; K-means Clustering; Linear Discriminant Analysis.

1. Introduction

With the development of several generations, the application of human energy and machine learning algorithms has penetrated into all aspects of people's production and life, making great contributions to promoting the convenience of human production and life in fields of e-commerce, media, transportation, and medicine [1-2]. Through machine learning algorithms, it can solve various data classifications, recognition, detection, prediction, and other problems in people's production and life. Compared with manual processing, using hardware equipment and technology of artificial intelligence data analysis and processing can greatly improve the efficiency and accuracy of work in some fields, thus reducing labor costs and error risks caused by manual operation. Up against social progress, each walk of life puts forward higher requirements for machine algorithms, and the goal is to achieve more accurate and stable results.

Traditional Chinese medicine is an important part of our medicine, thus it is important to identify the types (varieties, producing areas, quality, etc.) of Chinese herbal medicines in the research field of traditional Chinese medicine [1-6]. Infrared spectroscopy features have vital applications in the analysis of material components. We use infrared spectroscopy information to study the identification of Chinese herbal medicines. First of all, from the perspective of purchasing medicinal materials, there may be obvious differences in the quality and medicinal properties of the same medicinal materials in different producing areas, thus leading to great differences in prices. That's why correctly distinguishing the types, producing areas, and quality of Chinese herbal medicines plays an imperative role in Chinese medicine [7, 8]. Accurate identification of the types and producing areas of medicinal materials can effectively eliminate the shoddy goods sold at decent prices, thus protecting the interests of businessmen and buyers. Secondly, from the patient's point of view, there

will be great differences in efficacy between two kinds of medicinal materials with similar appearance and smell. If the wrong medicine is used, the effect of the medicine will be completely different, and even endanger the health and life of patients. Therefore, for patients, the correctness of each Chinese medicine is necessary. Finally, from the perspective of academic significance, the research on data classification algorithms based on spectral features has certain academic value for promoting the theory and application of data classification algorithms [9, 10].

Dimension reduction algorithm can not only reduce the dimension of high-dimensional data through transformation, but also extract more prominent features through transformation to improve the discrimination of data. LDA and PCA are two common dimension-reduction algorithms, which are widely used in problems related to classification and recognition [11, 12]. For example, in face recognition, LDA and PCA algorithms can effectively improve the discrimination of sample features by learning a projection feature, widely used from company punch-in and access control recognition. This kind of algorithm obtains projection features through training, which uses more prominent projection features to calculate the distance between face images. Because the data in the database has tag information, we can search for the sample with the smallest distance from the face to be recognized in the database to determine the identity of the test image. At the same time, this algorithm can also be used in the identification of Chinese herbal medicines. Many kinds of Chinese herbal medicines exist, and there is no obvious difference in appearance characteristics among similar Chinese herbal medicines, which brings great challenges to their identification. In this paper, we will design data classification models based on PCA and LDA algorithms respectively, in addition to studying the classification of Chinese herbal medicines under unsupervised and supervised conditions respectively, so as to realize the classification of Chinese herbal medicines by using hardware equipment, minimize labor costs, and improve classification efficiency and accuracy.

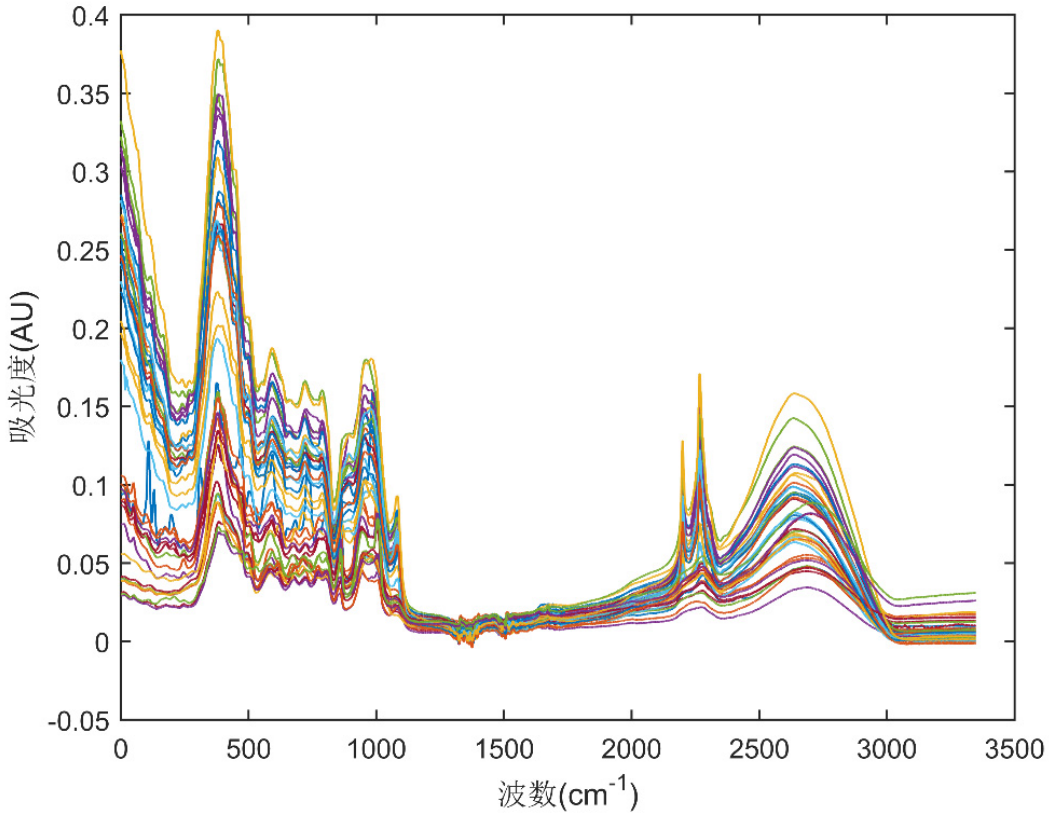


Fig 1. Spectral Image of Chinese Herbal Medicines

To sum up, the identification of Chinese herbal medicines is of great significance in the research of Chinese medicine, so people have put forward various methods for their identification. In the analysis of identifying Chinese herbal medicines based on the infrared spectrum, considering that it

is difficult to get reasonable discrimination when only using near-infrared or mid-infrared data for discriminant analysis, Ding Xueli et al. combined two kinds of spectral data (mid-infrared and near-infrared) of Chinese herbal medicines by standard deviation method and trained samples of Chinese herbal medicines from several definite producing areas by Fisher discriminant analysis (LDA/FDA), finally finding a convenient and quick research method [1]. In view of the problems in the development of quality control of Chinese herbal medicines, Hu Yang and Li Xianzhi, considering that the current mainstream quality control methods of traditional Chinese medicine need to spend a lot of time and energy on sample processing and some other disadvantages, analyzed the quality control technology of traditional Chinese medicine based on infrared characteristics by summarizing the research of near-infrared spectroscopy technology in neutral analysis, quantitative analysis and in-process control [2]. In identifying medicinal materials by spectrum, Zheng Jie et al. collected the spectral information of two-kind medicinal materials by near-infrared hyperspectral technology in their research work. By comparing five different pre-treatment methods, and using partial least square discriminant analysis, support vector machine, and other methods, the research models were established for research, but the number of samples was small [3]. Xie Meiying et al. can quickly and effectively identify the chemical components and producing areas of Achyranthes bidentata through cluster analysis and PCA spectral data analysis, and adopt the technology of combining infrared spectroscopy with second derivative spectroscopy. However, it also has the problem of fewer sample materials [4]. In the research and identification of Chinese herbal medicines based on deep learning, Wu Chong proposed a new identification model of Chinese herbal medicines considering some comprehensive background factors, and modified the model by using the particularity of attention mechanism, thus applying it to actual scenes [5]. Zhou Yongjun et al. determined three different Chinese herbal medicines in a certain range by using terahertz time-domain spectroscopy and obtained corresponding spectral tables with subtle differences, which finally realized the identification of Chinese herbal medicines [6]. In order to quantitatively compare the same medicinal characteristics of different Chinese herbal medicines, Wang Xian-Rui and others quantified the important properties and their medicinal characteristics, and studied the relationship between primary metabolic molecules and “cold-hot” medicinal properties based on the characteristics of Chinese herbs and Fisher discriminant analysis. Meanwhile, the quantitative value “R” of “cold-hot” and “warm-hot” medicinal properties was obtained according to the relevant indexes of medicinal properties, and the fixed value “R” was used to express the strength of medicinal properties. Therefore, it is concluded that golden thread has the coldest property as a medicinal material, which is basically consistent with the cognition of TCM theory. So, this method has certain guiding significance for the quantification of “cold and hot” medicinal properties [7]. On the problem of using spectrum to analyze traditional Chinese medicine, Chen Yan-jiang et al. applied the SVM method to identify samples by THz spectrum. The data of THz spectrum and absorption rate were obtained through experiments, and the characteristic space of traditional Chinese medicine was constructed. At the same time, they also established a BP network model for spectral recognition of traditional Chinese medicine. The final results show that this method has good prediction ability, but the effect of SVM is better than BP artificial neural network for small samples [8]. LiliWanga and CongWangab et al. used pyrolysis gas chromatography to analyze the fingerprint pyrolysis patterns of several samples from different growing sites and different species. At the same time, through the hierarchical cluster analysis of the data matrix which can identify a certain specification (16*40), samples from different growing places and species are distinguished. The combination of Py-GC fingerprint technology and chemometrics is a fast and effective selective classification method, which is more practical for the quality control of herbal raw materials [9].

Therefore, in our research, we hope to classify the spectral data by data classification algorithm and design a newly effective measurement model according to the characteristics of spectral data of Chinese herbal medicines, so as to realize the rapid identification of the characteristics, types, and producing areas of various medicinal materials. In our study, different methods are designed for the

classification of medicinal materials' spectral feature data from two application scenarios, that is, unsupervised and supervised learning. The main innovations in our work are as follows.

(1) A sample classification method based on Gaussian distribution is designed. For unsupervised data classification, we first divide the data into different categories by clustering method. Then, the largest class of data as the first class uses Gaussian mixture distribution to classify the data again, so as to get more reasonable classification results.

(2) An improved LDA data recognition method is proposed. Different from traditional classification methods, for supervised multi-class data recognition, we use sample difference vector to transform multi-class into the two-class problem, and re-train the feature transformation matrix through k similar samples to form potential positive sample data, so that the recognition result is more accurate.

(3) Aiming at the classification problem, we have done a lot of experiments and analyses on the two different data sets, unsupervised and supervised, to verify the effectiveness of our algorithm.

2. Study on Unsupervised Identification Method of Chinese Herbal Medicines based on Cluster Analysis

In the identification of Chinese herbal medicines, one of the application scenarios is to identify and separate many different Chinese herbal medicines. Fig. 1 shows the mid-infrared spectral characteristic curves of medicinal materials. From the data in the figure, it can be seen that the spectral characteristics of medicinal materials are very complex, and there is no obvious characteristic that is easy to distinguish between the characteristic curves of different samples. The problem of unsupervised sample classification and recognition is mainly to design a method through the similarity between samples, so that similar samples can be divided into the same class and samples with great differences can be divided into different classes. However, the characteristics of Chinese herbal medicines used in this paper are high-dimensional and complex. Therefore, the Euclidean distance based on the original mid-infrared spectral features is not ideal to measure the similarity of samples. This paper aims at the unsupervised classification of Chinese herbal medicines. Firstly, principal component analysis (PCA) is used to reduce the dimension of the original sample feature data, in order to get a more discriminative and significant classification feature. Then, based on the processed feature data, an unsupervised identification model of medicinal materials based on K-means algorithm is established.

2.1 Dimension Reduction Processing based on Principal Component Analysis

PCA dimension reduction plays an important role in evaluation and classification, and as an important preprocessing method, it provides a basis for related classification and recognition algorithms. The algorithm seeks a projection feature with more discrimination ability by maximizing the variance of projection data. According to the spectral characteristic curve of Chinese herbal medicines shown in Fig. 1, it can be seen that measuring the similarity of Chinese herbal medicines through original characteristics will be affected by noise. We first centralize the data and then use PCA to pre-process the initial data to obtain a transformation feature. Define the spectral characteristic data of the sample as $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{im})$, where $i = 1, 2, \cdots, N$. Definition symbol $\mathbf{x}_i$ is a feature vector. m is the feature dimension and N is the total number of samples. The feature projection vector in the PCA algorithm is represented by symbol $\mathbf{w} = (w_1, w_2, \cdots, w_m)$. The vector uses each element as a group of coefficients to form a fixed feature transformation mode, and $\mathbf{x}_i$ will be transformed into coordinates $\mathbf{w}^T\mathbf{x}_i$ in the projection direction. The purpose of the principal component analysis is to solve this vector, so that the variance of projected coordinates of all sample individuals is maximized, that is to say, the variance of data in the coordinate set $\{\mathbf{w}^T\mathbf{x}_i\}$ is maximized. Therefore, the principal component analysis model is established as follows:

$$\begin{aligned} &\arg\max_{\mathbf{w}} \quad J(\mathbf{w}) = \mathbf{w}^T \mathbf{X}\mathbf{X}^T \mathbf{w} \\ &s.t. \quad \mathbf{w}^T \mathbf{w} = 1 \end{aligned} \tag{1}$$

Where $\mathbf{X} = \left(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_i, \cdots, \mathbf{x}_N\right)$ represents the sample matrix.

By solving the optimization problem in the model (1) by the Lagrange multiplier method, the eigenvalue problem in formula (2) is obtained.

$$\mathbf{X}\mathbf{X}^T \mathbf{w} = \lambda \mathbf{w} \tag{2}$$

By solving formula (2), we can obtain $\lambda_1, \lambda_2, \cdots, \lambda_m$ as the eigenvalue of $\mathbf{X}\mathbf{X}^T$, and we have a set of eigenvectors $\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m$ corresponding to it. Assuming that we will sort $\lambda_1, \lambda_2, \cdots, \lambda_m$, there is $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$. Then, $\lambda_1 = \max\left\{\lambda_1, \lambda_2, \cdots, \lambda_m\right\}$ is the optimal function value of the objective function in formula (1). λ_2 is the suboptimal value, and the list goes on. Finally, we select only k vectors from the m solutions, and a set of transformation features with greater sample discrimination $\left(\mathbf{w}_1^T \mathbf{x}_i, \mathbf{w}_2^T \mathbf{x}_i, \cdots, \mathbf{w}_k^T \mathbf{x}_i\right); k < m$ is obtained finally. The k value is calculated as follows:

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \geq \alpha \geq \frac{\sum_{i=1}^{k-1} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \tag{3}$$

Where α is the feature selection threshold, which is taken as $\alpha = 0.98$ in this paper. For the original sample feature data measured by the instrument, we substitute the feature vector obtained by solving formula (2) into the transformation equations in formula (4), and then substitute the original feature data of each sample into formula (4) for transformation.

$$\begin{cases} y_1 = w_{11}x_1 + w_{12}x_2 + \cdots + w_{1m}x_m \\ y_2 = w_{21}x_1 + w_{22}x_2 + \cdots + w_{2m}x_m \\ \vdots \\ y_k = w_{k1}x_1 + w_{k2}x_2 + \cdots + w_{km}x_m \end{cases} \tag{4}$$

Each of which $y_1, y_2, \cdots, y_k$ represents a projection feature. It can be seen from Formula (4) that we learn a set of linear transformations through the PCA algorithm, which transforms the original sample features into a feature with a stronger resolution. By mining the common parts of the original data, the PCA algorithm finds a projection feature to separate the data as much as possible, so as to improve the resolution of features. Based on the PCA algorithm, the resulting projection data is shown in fig. 2 below.

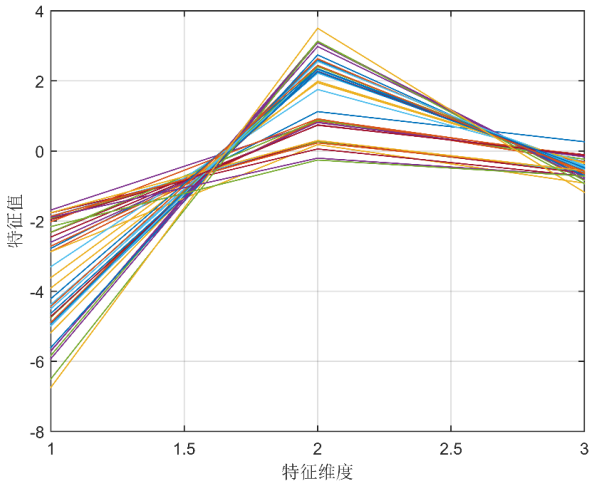


Fig 2. PCA Characteristics of Sample Data

As can be seen from fig. 2, after the above processing process, the feature dimension of the original data can be greatly reduced. From the distribution of samples in various dimensions, it can be seen that the discrimination of different individuals is clearer.

2.2 Unsupervised Classification of Medicinal Materials based on K-means Clustering Algorithm

2.2.1 Clustering Model based on K-means Algorithm

With regard to the unsupervised identification of Chinese herbal medicines, the purpose is to divide the given infrared characteristic data of unlabeled Chinese herbal medicines into different categories according to the trend of spectral characteristics, apart from dividing the medicinal materials of the same kind and the same origin into the same category and marking them with the same symbols, so as to mark the samples of Chinese herbal medicines of different kinds or origins with different labels. It can be seen that the identification of Chinese herbal medicines based on unsupervised infrared spectroscopy is a typical clustering problem. In this paper, based on the K-means algorithm, using the spectral characteristic data of Chinese herbal medicines, the sample data are clustered to achieve the division of different classes of Chinese herbal medicines. The feature data of Chinese herbal medicines have been reduced in the dimension above. For the processed data, we define a symbol $\mathbf{Y}=\left\{\mathbf{y}_i=\left(y_1,y_2,\cdots,y_k\right)\right\}$ to represent the feature data set after sample feature conversion. Among them, the symbolic representation $\mathbf{y}_i$ represents the transformed feature of the i-th individual in the set as a vector. $l_i=1,2,\cdots,n$ is defined and $\mathbf{y}_i$ is the category of tagged individuals as tag information. For the identification of Chinese herbal medicines, it can be regarded as a typical clustering problem. By the clustering method, all samples are divided into different categories according to the feature similarity. $C_r=\left\{\mathbf{y}_i \mid l_i=r\right\}$ is defined as sample collections into different categories.

Using the above-processed data, we use the clustering method to classify unsupervised samples and separate different medicinal materials. In high-dimensional space, we regard each individual in the sample set as a point, and calculate the distance between points to measure the similarity between samples. Assuming that all sample data have r categories, the objective function of clustering analysis is to minimize the sum of squares of total errors, that is to say, the distance from each individual to the center of its category is the smallest, so is the sum of total distances, as shown in Formula (5):

$$D=\sum_{r=1}^{n}\sum_{\mathbf{y}_i\in C_r}\left(d\left(\mathbf{y}_i,center_r\right)\right)^2 \tag{5}$$

Where $d\left(\mathbf{y}_i,center_r\right)$ is a distance function, which calculates the distance of the $center_r$ of the class corresponding to the i-th point. Variable $center_r$ represents the coordinate center of all points belonging to the r-th class,

$$center_r=\frac{1}{N_r}\sum_{\mathbf{y}_i\in C_r}\mathbf{y}_i \tag{6}$$

Where N_r is the total number of points representing the r-th class of data.

It is worth noting that the unsupervised classification of spectral feature data of Chinese herbal medicines in this paper is based on the projection features obtained by PCA algorithm, and measures the similarity of samples. Considering the characteristics of the PCA projection features, we use the distance function in Formula (7):

$$d\left(\mathbf{y}_i,center_r\right)=\sqrt{\sum_{s=1}^{k}\lambda_k\left(y_{is}-center_{rs}\right)^2} \tag{7}$$

Where y_{is} is the s-th projection feature of the sample $\mathbf{y}_i$. $center_{rs}$ represents the s-th feature of the cluster center $center_r$. λ_k is the k-th eigenvalue calculated in 2.1 above. We use this eigenvalue

as the weight of each dimension projection feature to measure the similarity between samples more accurately.

2.2.2 Calculation Method

In the above paper, the classification problem of spectral characteristic data of Chinese herbal medicines is defined, and the data classification method based on the K-means clustering algorithm is described through mathematical language. Below, all calculation methods are explained one by one in this paper in order.

2.2.2.1 Initialization

In the K-means clustering method, the initialization of model parameters is a key step. According to the principle of the algorithm, there are two kinds of parameters that need to be set with one of them to be determined subjectively, including the number of cluster types K and the center of the cluster. Setting the initial values of these two parameters is the key content of the algorithm.

(1) K Value Initialization: The value of K is a value defined subjectively through human experience or according to the needs of practical problems. In the classification of spectral data of Chinese herbal medicines, we roughly determine the number of clustering categories by observing the distribution of spectral characteristic curves.

(2) Category center coordinates: The initialization of this parameter is to define the center coordinates of all categories at the initial time. We use the random initialization method to assign values to the center coordinates of all categories. The method is to randomly select K points in the sample set, and require that the coordinates of these points are different, otherwise the points of the two categories will be divided into the same category. We define variables $center_1$, …, $center_K$ which represent the center coordinates of these categories.

Input: Characteristic data set $\{\mathbf{y}_i = (y_1, y_2, \cdots, y_k)\}$ of medicinal materials.

Output: The category label $\{l_i = 1, 2, \cdots, m\}$ corresponding to the input sample.

2.2.2.2 Iterative Calculation

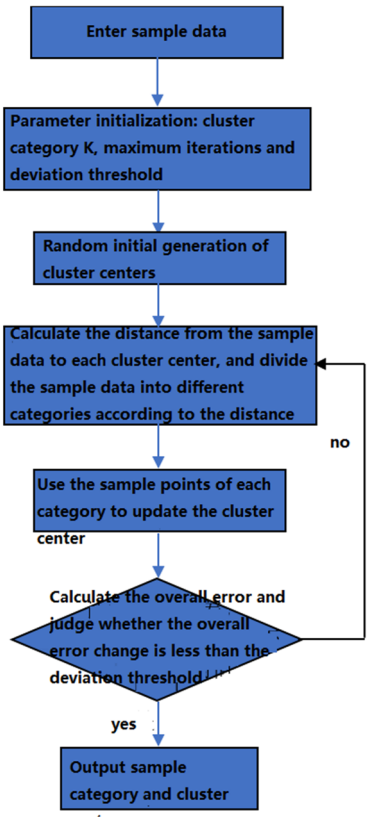


Fig 3. K-means Clustering Flow Chart of Chinese Herbal Medicines

According to the initialization conditions as well as input and output data formats in 2.2.2.1, the data is clustered by iteration, including:

1. According to the initialized class center coordinates, using the distance function in Formula (6), the distance from each sample point to k centers is determined. The smaller the value, the closer it is. Then, we use the distance value to determine which center point each data point is closest to, and mark which category each data point belongs to according to the category of cluster center:

$$l_i = \arg\min_r \left\{ d\left(\mathbf{y}_i, center_r\right), r = 1, 2, \cdots, n \right\} \tag{8}$$

2. According to the classification result $\left\{ y_{ij} \right\}$ obtained in the previous step, the cluster center is updated, that is, the corresponding category center is replaced by calculating the average value of all pixels belonging to each category. The specific calculation method is shown in Formula (3).

3. The above steps 1) and 2) are repeated until the deviation of the total error of the results of the current k-th iteration and the last iteration is less than a given threshold, i.e., $\left| D(k) - D(k-1) \right| \le \alpha$

4. Output $\left\{ l_i \right\}$

The overall flow is shown in fig. 3.

2.2.3 Classification Results and Analysis of Chinese Herbal Medicines

In K-means clustering, the value of the category label does not have any real meaning, but is used to distinguish different categories of data. That is to say, samples with the same label value belong to the same class, while samples with different label values belong to different classes. The initial clustering center is randomly assigned or selected points, so the same data will be labeled differently in each clustering, but which points belong to the same class are the same. Thus, the sample data corresponding to the same kind of Chinese herbal medicines in each clustering result is the same.

According to the above calculation method, the sample data are tested under different clustering parameters (K=3, 4, 5, 6, 7, 8), and the test results are given in Figure 4. The horizontal axis y_1 and vertical axis y_2 are two kinds of conversion features obtained after PCA processing, which correspond to λ_1 and λ_2. In this paper, the principal component analysis parameter $\alpha = 0.98$ is set, and three projection features $\mathbf{y}_i = \left(y_1, y_2, y_3 \right)$ are extracted. In order to show the distribution of clustered samples, we choose features as horizontal and vertical coordinates y_1 and y_2, and use different colors to depict the scatter graph of each type of data. It can be seen from the results in the figure that samples can be well classified by clustering. However, when K is selected with different values, the classification results of samples are quite different. In particular, under different K value settings, the three outliers in the upper left corner of the graph are all divided into the same category. It can be seen from the figure that among the three blue outliers, the two points in the upper left corner are closer, and the remaining one point is farther away from the other two points. However, according to the clustering results, these three special sample points are divided into the same category. It can be seen that there are obvious deficiencies in determining the number of categories and data distribution characteristics based on this method. According to the objective function in formula (5), this classical clustering algorithm mainly considers the classification deviation of the sample population, but does not consider the distribution characteristics of the samples. Especially when the sample distribution is unbalanced, the classification effect of small sample categories is flawed. For three outliers, the K-means algorithm does not fully consider the distribution characteristics of three outliers in order to minimize the overall error (see formula (5)). Therefore, we will optimize the clustering method for the deficiencies, using the characteristics of the data distribution to improve the classification effect.

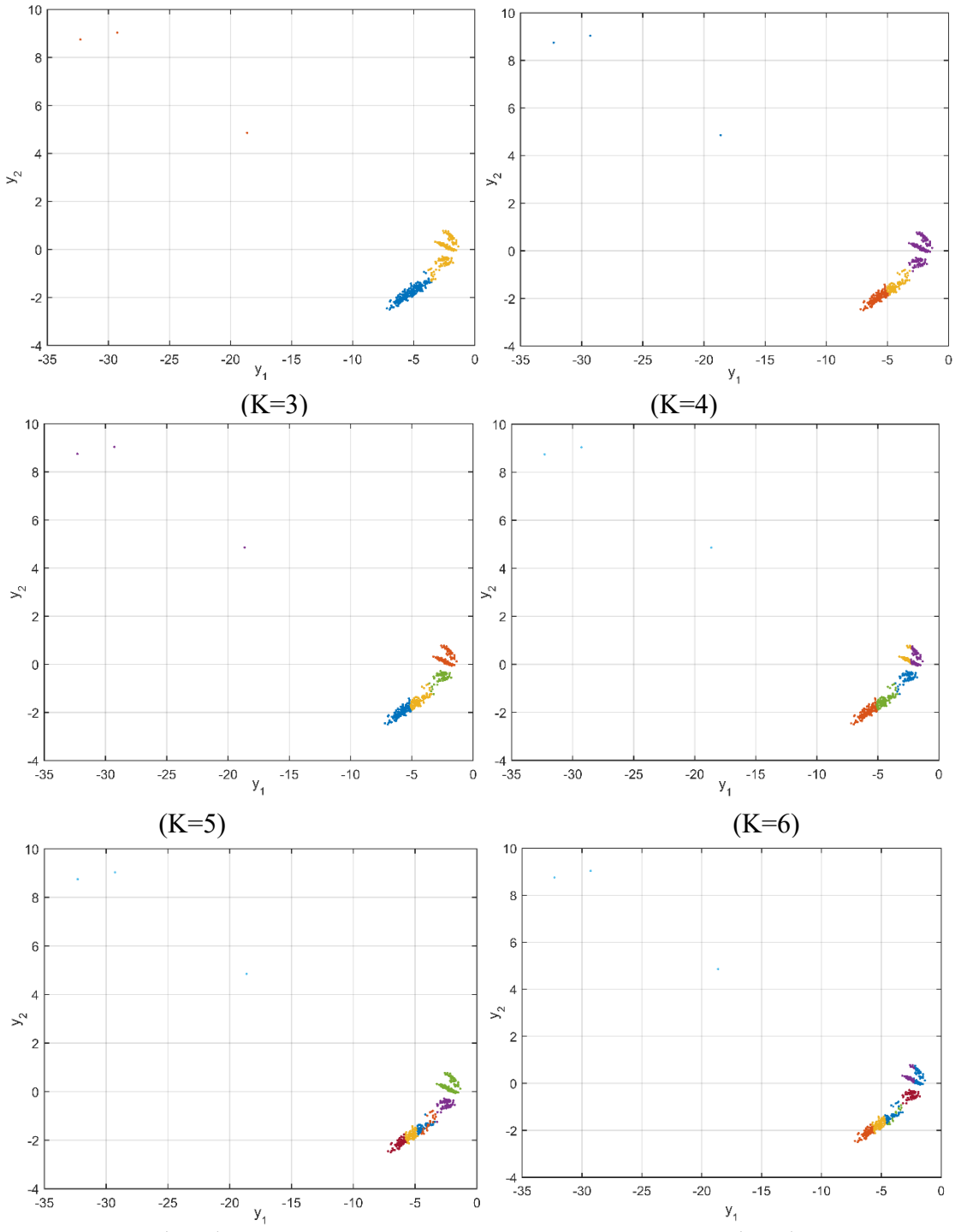


Fig 4. Data Classification Results Under Different K Values

2.3 Improved K-means Algorithm based on Gaussian Model

For unsupervised sample data, we use the clustering method to solve the problem. However, the shortcoming of the unsupervised classification method based on the K-means algorithm is that the number of data categories is unknown, that is to say, the number of clustering categories K is a parameter set manually. There is no reliable basis to accurately calculate the number of categories. Clustering category parameter K is artificially set through data analysis and experience. Therefore, how to accurately judge the number of categories has become a key problem in a K-means clustering algorithm. In this paper, based on the distribution characteristics of spectral data of Chinese herbal medicines, the existing K-means algorithm is improved by using the Gaussian function.

In this paper, it is assumed that the spectral characteristic data of the same class of Chinese herbal medicines obey multivariate Gaussian distribution, as shown in Formula (9):

$$\mathbf{y}_{k,i} \sim N\left(\mu_k, \Sigma_k\right) \tag{9}$$

In the above formula, μ_k is the mean of class k, Σ_k is the corresponding covariance. $\mathbf{y}_{k,i}$ represents a sample in category k. Therefore, we can describe the sample distribution of Chinese herbal medicines by a Gaussian mixture model, which can be expressed by formula (10):

$$p\left(\mathbf{y}\right) = \sum_{k=1}^{m} \alpha_k N\left(\mathbf{y} \middle| \mu_k, \Sigma_k\right) \tag{10}$$

$N\left(\mathbf{y} \middle| \mu_k, \Sigma_k\right)$ represents the component of the k-th Gaussian kernel function.

$$N\left(\mathbf{y} \middle| \boldsymbol{\mu}_k, \Sigma_k\right) = \frac{1}{\left(2\pi\right)^{D/2}} \frac{1}{\left|\Sigma_k\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{y}-\boldsymbol{\mu}_k\right)^T \Sigma_k^{-1}\left(\mathbf{y}-\boldsymbol{\mu}_k\right)\right\} \tag{11}$$

If there are m different classes of samples in the sample data, the overall distribution of the samples can be described by m different multivariate Gaussian distributions, where the mixing coefficient is represented by α_k, related to the proportion of the number of samples of the k-th class and satisfies the condition in formula (12):

$$\begin{gathered} \sum_{k=1}^{m} \alpha_k = 1 \\ 0 \le \pi_k \le 1 \end{gathered} \tag{12}$$

For the parameter estimation method of the Gaussian mixture, we use the EM algorithm and use joint probability distribution to establish the likelihood function as follows:

$$p\left(\mathbf{Y}, \mathbf{Z} \middle| \boldsymbol{\mu}, \Sigma, \boldsymbol{\alpha}\right) = \prod_{n=1}^{N} \prod_{k=1}^{K} \alpha_k^{z_{n,k}} N\left(\mathbf{y}_n \middle| \boldsymbol{\mu}_k, \Sigma_k\right)^{z_{n,k}} \tag{13}$$

$\mathbf{Z} = \left\{z_{n,k}\right\}$ represents the ownership of the n-th data to the k-th category. Logarithmic operation on both sides of the equal sign of formula (13) is as follows:

$$\ln p\left(\mathbf{Y}, \mathbf{Z} \middle| \boldsymbol{\mu}, \Sigma, \boldsymbol{\alpha}\right) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{n,k} \left\{\ln \alpha_k + \ln N\left(\mathbf{y}_n \middle| \boldsymbol{\mu}_k, \Sigma_k\right)\right\} \tag{14}$$

According to formula (14), calculate the posterior probability of variable z, thus obtaining:

$$\gamma\left(z_{n,k}\right) = \frac{\alpha_k N\left(\mathbf{y}_n \middle| \boldsymbol{\mu}_k, \Sigma_k\right)}{\sum_s \alpha_s N\left(\mathbf{y}_n \middle| \boldsymbol{\mu}_s, \Sigma_s\right)} \tag{15}$$

Finally, the mathematical expectation is shown in Formula (16):

$$\mathbf{E}_{\mathbf{Z}}\left[\ln p\left(\mathbf{Y}, \mathbf{Z} \middle| \boldsymbol{\mu}, \Sigma, \boldsymbol{\alpha}\right)\right] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma\left(z_{n,k}\right) \left\{\ln \alpha_k + \ln N\left(\mathbf{y}_n \middle| \boldsymbol{\mu}_k, \Sigma_k\right)\right\} \tag{16}$$

The function (16) derives each variable separately to solve the optimal solution, and the obtained solution is as formula (17):

$$\begin{cases} \boldsymbol{\mu}_k = \dfrac{1}{N_k} \sum_{i=1}^{N_k} \gamma\left(z_{i,k}\right) \mathbf{y}_i \\ N_k = \sum_{i=1}^{N} \gamma\left(z_{i,k}\right) \\ \Sigma_k = \dfrac{1}{N_k} \sum_{i=1}^{N_k} \gamma\left(z_{i,k}\right)\left(\mathbf{y}_i - \boldsymbol{\mu}_k\right)\left(\mathbf{y}_i - \boldsymbol{\mu}_k\right)^T \\ \alpha_k = \dfrac{N_k}{N} \end{cases} \tag{17}$$

Based on the results of K-means clustering, this method is improved by using the distribution information of samples. In the improved method, we first classify the data by K-means. Then, the number of samples in each category is counted, the category with the largest number of samples is selected, and the kernel function of the Gaussian distribution is calculated, so as to obtain the parameters of the first distribution in the Gaussian mixture model. Then, according to the judgment condition of the Gaussian distribution, the samples with less than 3 times standard deviation are classified into the same class, and the data are re-classified by using the Gaussian mixture method.

Based on the above-improved algorithm, the data of Chinese herbal medicines are reclassified, the results are shown in fig. 4 below. It can be seen from the results shown in the figure that according to the improved unsupervised classification method of Chinese herbal medicines, the sample data are divided into four different categories. Different from traditional clustering methods, we introduce Gaussian distribution to judge whether all samples belong to the same class in each class of data, and fully consider the distribution characteristics of samples. Comparing the classification results of K=3 and K=4 in fig. 4, we can clearly see that the classification effect of the improved algorithm is more in line with the actual classification effect of Chinese herbal medicines. Combined with the spectral characteristic curves of Chinese herbal medicines in fig.1, it can be clearly seen that there are obvious differences between the spectral characteristic curves of the three samples and those of other medicinal materials. However, based on the traditional clustering method, these three sample points are divided into the same class, which is inconsistent with the actual situation. In the classification of the improved algorithm, these three samples are divided into two categories, which are consistent with the distribution characteristics of spectral characteristic curves of all samples shown in fig.1. Therefore, according to the experimental results, we can see that our improved method has a more accurate classification effect and accords with the objective fact of data distribution.

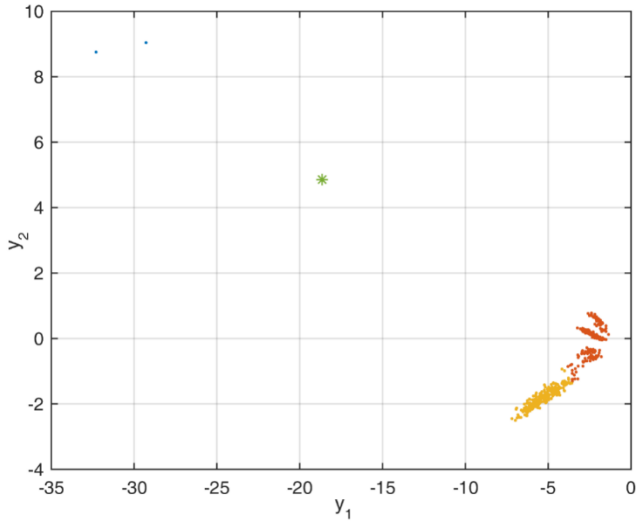


Fig 5. Classification Effect of Improved K-means Clustering Algorithm

3. Identification Method of Chinese Herbal Medicines based on Linear Discriminant Analysis

3.1 Principle of Supervised Identification of Chinese Herbal Medicines

Supervised identification of Chinese herbal medicines requires more detailed classification of spectral characteristics of Chinese herbal medicines. In this paper, it is defined as a supervised classification problem. By classifying the spectral data of Chinese herbal medicines, we learn a spectral feature projection method and project the original data into a subspace with more resolution, so that the distance between a pair of samples belonging to the same category $(\mathbf{x}_i, \mathbf{x}_j)_+$ is significantly

smaller than that of a pair of samples belonging to different categories $(\mathbf{x}_i, \mathbf{x}_k)_-$, that is to say $f(\mathbf{x}_i, \mathbf{x}_j)_+ \ll f(\mathbf{x}_i, \mathbf{x}_k)_-$. Then, for the samples $\mathbf{z}$ to be identified, the distance between them and each known sample is calculated, and the most similar samples are found according to the distance, so as to realize the identification of sample categories $\mathbf{z}$.

According to the above analysis, it can be seen that establishing a strong resolution and robust distance function is the key to the supervised classification of spectral data of Chinese herbal medicines. In this paper, linear discriminant analysis (LDA) is used to maximize the differences of different categories of individuals. At the same time, it is necessary to minimize the differences between individuals of the same class. According to these two goals, the feature transformation pattern is learned. Based on feature projection, we get an effective Mahalanobis distance to measure the similarity of spectral data of Chinese herbal medicines, as shown in Formula (18):

$$d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W}\mathbf{W}^T (\mathbf{x}_i - \mathbf{x}_j) \tag{18}$$

Where the transformation matrix of high-dimensional features is represented by $\mathbf{W}$.

Given the sample data set of spectral characteristics of Chinese herbal medicines is $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_i, \cdots, \mathbf{x}_N\}$, the corresponding sample category labels are $\{y_1, y_2, \cdots, y_i, \cdots, y_N\}$, $i = 1, 2, \cdots, N$.

Based on LDA classification method, a projection space with stronger resolution can be obtained by learning training data. Further, in order to calculate the distance between the two samples more accurately, we construct the Mahalanobis distance on the converted data as shown in Formula (19)

$$d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W}\mathbf{M}\mathbf{W}^T (\mathbf{x}_i - \mathbf{x}_j) \tag{19}$$

Where $\mathbf{M}$ is the metric matrix of Mahalanobis distance.

According to the definitions in the distance models (18), (19), it can be seen that the measure of similarity between samples is based on the sample feature difference vector $\mathbf{x}_i - \mathbf{x}_j$. For the distance measure of samples in projection space, we first assume that the eigenvectors of samples obey multivariate Gaussian distribution. $\mathbf{u}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ represents the feature difference vector of a pair of samples. Therefore, $\mathbf{u}_{ij}$ also obeys multivariate Gaussian distribution. $l_{ij} = 0$ *or* 1 are defined as labels that represent sample pairs $(\mathbf{x}_i, \mathbf{x}_j)$. $l_{ij} = 0$ means that $(\mathbf{x}_i, \mathbf{x}_j)$ belongs to the same class, that is, a positive sample; On the contrary, $l_{ij} = 1$ means that $(\mathbf{x}_i, \mathbf{x}_j)$ belongs to different classes, that is, negative samples.

According to the Gaussian distribution hypothesis, the following probabilities are obtained:

$$\begin{cases} p\left(\mathbf{x}_i, \mathbf{x}_j \mid \boldsymbol{\theta}_0\right) = \dfrac{1}{(2\pi)^{\frac{d}{2}} \left|\Sigma_0\right|^{\frac{1}{2}}} \exp\left(-\dfrac{1}{2}\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \Sigma_0^{-1}\left(\mathbf{x}_i - \mathbf{x}_j\right)\right) \\ p\left(\mathbf{x}_i, \mathbf{x}_j \mid \boldsymbol{\theta}_1\right) = \dfrac{1}{(2\pi)^{\frac{d}{2}} \left|\Sigma_1\right|^{\frac{1}{2}}} \exp\left(-\dfrac{1}{2}\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \Sigma_1^{-1}\left(\mathbf{x}_i - \mathbf{x}_j\right)\right) \end{cases} \tag{20}$$

Where $p\left(\mathbf{x}_i, \mathbf{x}_j \mid \boldsymbol{\theta}_0\right)$ represents the probability of positive sample distribution and $p\left(\mathbf{x}_i, \mathbf{x}_j \mid \boldsymbol{\theta}_1\right)$ the probability of negative sample distribution. $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ represent positive and negative sample distribution parameters respectively. Σ_0 and Σ_1 represent covariance matrices of positive and negative samples respectively. In addition, according to Formula (19), we can see that the distribution probability of positive and negative samples is defined on the basis of characteristic difference vector.

According to the idea of KISSME algorithm, we use likelihood ratio function (such as formula (20)) to define sample distance,

$$y = f\left(\mathbf{u}_{ij}\right) = \log\left(\frac{p\left(\mathbf{x}_i, \mathbf{x}_j \mid \boldsymbol{\theta}_1\right)}{p\left(\mathbf{x}_i, \mathbf{x}_j \mid \boldsymbol{\theta}_0\right)}\right) \tag{21}$$

According to the definition in formula (20), when the probability of a pair of samples belonging to positive samples is greater than the probability of negative samples, the corresponding likelihood ratio function value is less than 0; On the contrary, when the probability of a pair of samples belonging to positive samples is less than that of negative samples, the corresponding likelihood ratio function value is greater than 0. Therefore, it can be seen that as for the distance function defined in formula (20), the higher the sample similarity, the greater the probability of belonging to positive samples and the smaller the corresponding function value. It can be seen from this that a distance function based on sample distribution characteristics can be effectively defined according to formula (20). Substituting formula (19) into formula (20) and taking logarithms for both sides generate the following results.

$$f\left(\mathbf{u}_{ij}\right) = \left(\mathbf{u}_{ij}\right)^T \Sigma_1^{-1}\left(\mathbf{u}_{ij}\right) + \log\left(\left|\Sigma_1^{-1}\right|\right) - \left(\mathbf{u}_{ij}\right)^T \Sigma_0^{-1}\left(\mathbf{u}_{ij}\right) - \log\left(\left|\Sigma_0^{-1}\right|\right) \tag{22}$$

Where $\log\left(\left|\Sigma_1^{-1}\right|\right)$, $\log\left(\left|\Sigma_0^{-1}\right|\right)$ is a constant, and removing these two constants has no effect on the distance measure. Thus, Formula (21) can be rewritten as Formula (22):

$$f\left(\mathbf{u}_{ij}\right) = \left(\mathbf{x}_i - \mathbf{x}_j\right)^T \left(\Sigma_0^{-1} - \Sigma_1^{-1}\right)\left(\mathbf{x}_i - \mathbf{x}_j\right) \tag{23}$$

It is worth noting that we define the distance measure function in the projection space, that is to say, the sample distance function can be rewritten as follows:

$$f\left(\mathbf{u}_{ij}\right) = \left(\mathbf{x}_i - \mathbf{x}_j\right)^T \mathbf{W}\left(\Sigma_0'^{-1} - \Sigma_1'^{-1}\right)\mathbf{W}^T\left(\mathbf{x}_i - \mathbf{x}_j\right)$$

$\Sigma_0'^{-1}, \Sigma_1'^{-1}$ is the covariance of positive and negative samples in projection space.

3.2 Learning Method of Measurement Model based on Discriminant Analysis

3.2.1 LDA Algorithm based on Multi-classification

As pointed out above, we learn a feature transformation mode with resolution through LDA algorithm and transform the original spectral features of Chinese herbal medicines into a classified projection feature, thus improving the discrimination of samples. First, calculate the average value of each class:

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{\mathbf{x}_i \in \mathbf{X}_k} \mathbf{x}_i \tag{24}$$

Where N_k represents the number of samples in the k-th category. Then, calculate the intra-class and inter-class divergence, as shown in formula (25):

$$\begin{cases} \Sigma_I = \sum_{k=1}^{r} \frac{N_k}{N} \sum_{\mathbf{x}_i \in \mathbf{X}_k} \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)^T \\ \Sigma_E = \sum_{k=1}^{r} \left(\boldsymbol{\mu}_k - \boldsymbol{\mu}\right)\left(\mathbf{u}_k - \boldsymbol{\mu}\right)^T \end{cases} \tag{25}$$

Variable Σ_I and Σ_E is intra-class and inter-class divergence, $\mathbf{X}_k$ represents the sample set of the k-th class and $\boldsymbol{\mu}$ represents the mean value of all samples. The classification model of Chinese herbal medicines is established as follows:

$$\arg\max_{\mathbf{w}} J(\mathbf{w}) = \frac{\mathbf{w}^T \Sigma_E \mathbf{w}}{\mathbf{w}^T \Sigma_I \mathbf{w}} \tag{26}$$

Formula (26) is a typical Rayleigh entropy function, with its maximum and minimum are the maximum and minimum eigenvalues of the matrix $\Sigma_I^{-1}\Sigma_E$ respectively. The eigenvalues and

eigenvectors of the matrix $\sum_I^{-1}\sum_E$ are calculated, and the eigenvectors of k eigenvalues with the largest value are found out. Meanwhile, the matrix $\mathbf{W}=(\mathbf{w}_1,\mathbf{w}_2,\cdots,\mathbf{w}_r)$ is constructed as the eigentransformation matrix. Through this matrix, we can transform the original feature into a more prominent feature to identify samples.

3.2.2 LDA Algorithm based on Binary Classification

Traditional data classification methods categorize the data of Chinese herbal medicines of the same kind and origin into the same, and then use multi-classification LDA algorithm to train a group of characteristic linear transformation patterns with the largest inter-class divergence and the smallest intra-class divergence. According to the definition of Mahalanobis distance above, we study the classification of Chinese herbal medicines on the basis of feature search vector, and transform the multi-classification of spectral data of Chinese herbal medicines into a binary classification of positive and negative sample pairs to train a projection space.

Based on a given set of samples $\{\mathbf{x}_i\}$, $\mathbf{u}_{ij}=\mathbf{x}_i-\mathbf{x}_j$. According to the label data of samples, the positive sample set $\mathbf{U}_1=\{\mathbf{u}_{ij}\}_+$ and the negative sample set $\mathbf{U}_0=\{\mathbf{u}_{ij}\}_-$ are obtained. The binary classification LDA algorithm is used to classify the feature query vector data of positive and negative sets. Calculate the covariance matrix of positive and negative classes:

$$\begin{cases}\sum_1=\sum\limits_{\mathbf{u}_{ij}\in\mathbf{U}_1}(\mathbf{u}_{ij}-\boldsymbol{\mu}_1)(\mathbf{u}_{ij}-\boldsymbol{\mu}_1)^T\\ \sum_0=\sum\limits_{\mathbf{u}_{ij}\in\mathbf{U}_0}(\mathbf{u}_{ij}-\boldsymbol{\mu}_0)(\mathbf{u}_{ij}-\boldsymbol{\mu}_0)^T\end{cases}\tag{27}$$

According to the principle of LDA linear discriminant analysis algorithm, we need to learn a projection direction and project the sample data to the same direction, so that the distance between the transformed data and the same kind of data is the smallest and the distance between different kinds of data is the largest. $\mathbf{w}=(w_1,w_2,\cdots,w_m)$ represents characteristic transform coefficients. After transformation, the centers of the two types of data are $\mathbf{w}^T\boldsymbol{\mu}_1$ and $\mathbf{w}^T\boldsymbol{\mu}_0$. According to the idea of LDA algorithm, we should maximize the distance between the two categories $\left\|\mathbf{w}^T\boldsymbol{\mu}_1-\mathbf{w}^T\boldsymbol{\mu}_0\right\|_2^2$. In addition, we want the distance between the sample data of the same class to be as small as possible after projection. By constraining the sum of variances of the sample data $\mathbf{w}^T\sum_1\mathbf{w}$ and $\mathbf{w}^T\sum_0\mathbf{w}$ in the projection direction, we can minimize the intra-class distance.

Based on the above algorithm principle, we establish the model in formula (28):

$$\arg\max_{\mathbf{w}}J(\mathbf{w})=\frac{\left\|\mathbf{w}^T\boldsymbol{\mu}_1-\mathbf{w}^T\boldsymbol{\mu}_0\right\|_2^2}{\mathbf{w}^T(\sum_1+\sum_0)\mathbf{w}}\tag{28}$$

From the optimization model in formula (28), it can be seen that the objective function of LDA linear discriminant analysis is a generalized Rayleigh entropy function, and its equivalent form is:

$$\begin{aligned}&\arg\max_{\mathbf{w}}\ \ J(\mathbf{w})=\left\|\mathbf{w}^T\boldsymbol{\mu}_1-\mathbf{w}^T\boldsymbol{\mu}_0\right\|_2^2\\ &s.t.\quad \mathbf{w}^T(\sum_1+\sum_0)\mathbf{w}=1\end{aligned}\tag{29}$$

For the model in formula (29), the generalized Lagrange multiplier method is used to solve the optimization model, and the solution of the optimization model is transformed into the following eigenvalue problem:

$$(\sum_1+\sum_0)^{-1}\left\|\mathbf{w}^T\overline{\mathbf{q}}_+-\mathbf{w}^T\overline{\mathbf{q}}_-\right\|_2^2\mathbf{w}=\lambda\mathbf{w}\tag{30}$$

The model (30) is solved to obtain a set of features and arrange them from large to small, namely $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$. The corresponding eigenvector is $(\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m)$. We only use r features with the largest feature to form the feature transformation matrix $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_r)$. The method chosen is the same as formula (3).

3.3 Supervised Classification Results and Analysis of Chinese Herbal Medicines

3.3.1 Experimental Data

In this part, we test the supervised classification and recognition methods of Chinese herbal medicines to verify the recognition effect of the model. There are 658 spectral characteristic data of Chinese herbal medicines in the data set used for algorithm verification, and these samples have 11 categories. Moreover, the number of samples in each category is uncertain. Fig. 6 shows the spectral characteristic curves of all these samples. It can be seen from the feature curve that these samples have no obvious discrimination and the features are very complex, so it is necessary to establish a more effective resolution method.

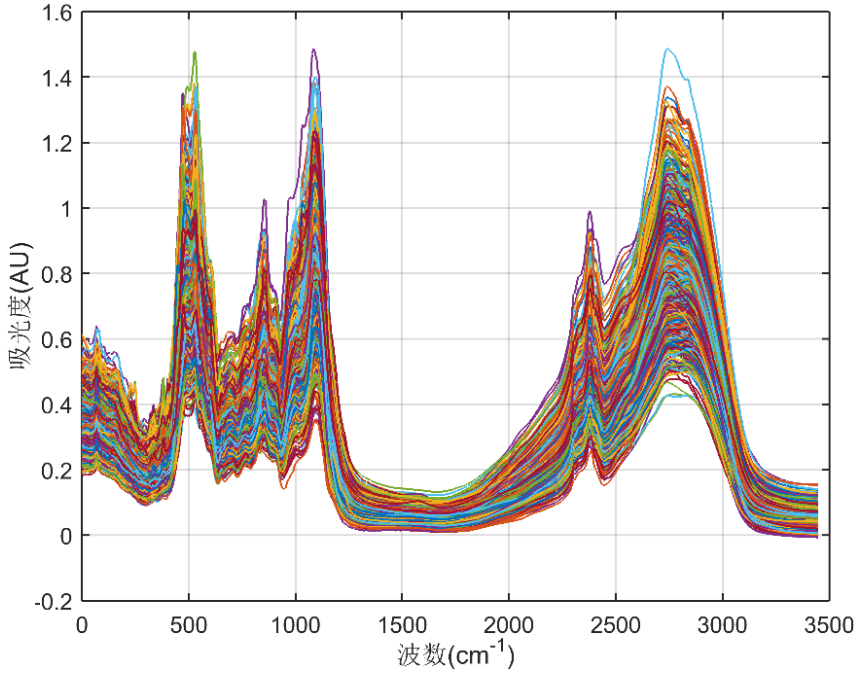


Fig 6. Spectral Characteristic Curve of Sample Set

3.3.2 Test Results and Analysis

Through the data set described in 3.3. 1, we tested the recognition effect of LDA recognition method. When launching the test experiment, all the data are randomly divided into training and test data. Among them, 100 samples are divided into test data. In addition, in order to ensure the fairness of test accuracy and avoid the error caused by random sampling, we carry out 10 independent repeated experiments in the test process, and take the average recognition accuracy as the final recognition result. Through programming calculation, the recognition accuracy of the classification and recognition model of Chinese herbal medicines established by LDA linear discriminant analysis reaches 87.25%. However, the recognition accuracy of the model based on binary LDA method is only 56.73%, which is mainly due to the strong correlation of spectral characteristics of Chinese herbal medicines, seen from the characteristic curve of samples in fig. 6. That's why the two-classification method based on feature difference vector is difficult to find projection features with stronger resolution. Only through multi-classification and more detailed constraint on the discrimination between samples can better results be achieved.

In order to illustrate the classification effect of the algorithm, we take the projection eigenvalues of the first three dimensions as the coordinate values of the samples. Then, we depict the three-

dimensional scatter diagram of the sample distribution, as shown in fig. 7. It can be seen from the effect in the figure that the data of different medicinal materials can be effectively separated based on LDA classification model.

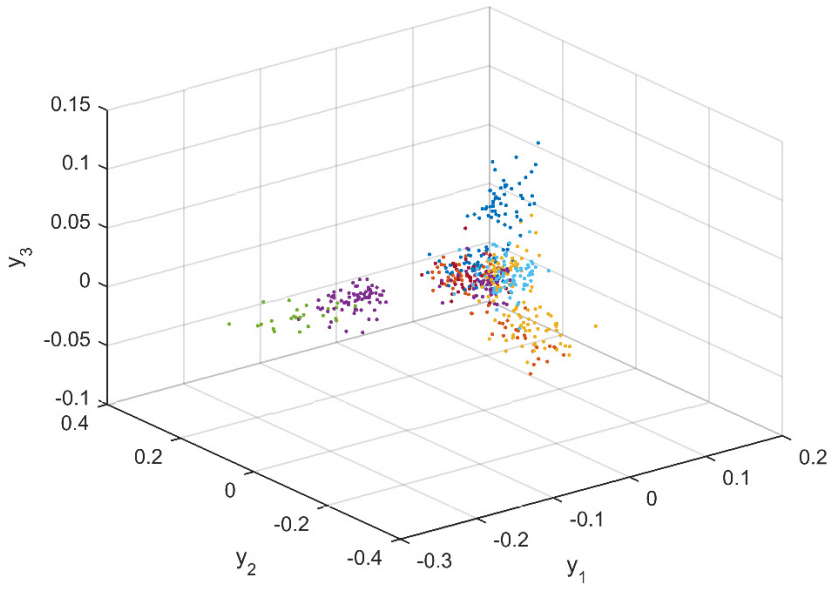


Fig 7. LDA Classification Results

4. Improved LDA Classification Algorithm based on Gaussian Distribution

4.1 Methods of Improvements

In the third part, we studied the supervised classification method of Chinese herbal medicines based on LDA. However, considering the complexity of spectral characteristics of Chinese herbal medicines in practice, due to the cleaning process of surface substances as well as the dry and wet state of Chinese herbal medicines when measuring spectral characteristics of Chinese herbal medicines, the collected spectral characteristics data of Chinese herbal medicines fluctuate violently, which leads to the lack of generalization performance of the data recognition model we learned at last. Therefore, we use the samples with small distance in the recognition results to form "pseudo-positive samples", and then optimize the original method by constructing the distribution constraint of "pseudo-positive samples".

As we all know, training and test data are obtained by random sampling with the same Gaussian distribution. When identifying test individuals, we identify test data by finding the individual closest to the test data in the projection space. Therefore, we use the distance model (19) obtained by model learning in formula (22) to measure the similarity of the test individuals. k most similar samples are determined for each sample tested according to the sample distance. Then the feature difference vector between the test individual and these similar data is regarded as potential positive samples, and the distribution characteristics of these positive samples are constrained. The constraint method is as follows:

$$\Sigma_1' = \sum_{i=1}^{k}\left(\left(\mathbf{z}_j - \mathbf{x}_i\right) - \boldsymbol{\mu}_1\right)\left(\left(\mathbf{z}_j - \mathbf{x}_i\right) - \boldsymbol{\mu}_1\right)^T \tag{31}$$

According to the principle in (31), we use the distance from the potential positive sample to the center of the real positive sample as the objective function, and recalculate the projection vector of the transformed feature by minimizing the divergence.

The specific process of the test is to calculate the distance between the test individual and each training sample by using the initial distance function learned by LDA, and then sort it as shown in Formula (32):

$$rank\left(\mathbf{x}_i,\mathbf{z}_j\right)=\sum_s \operatorname{sign}\left(d\left(\mathbf{x}_i,\mathbf{z}_j\right),d\left(\mathbf{x}_i,\mathbf{z}_s\right)\right)+1 \tag{32}$$

$rank(\bullet)$ is a functional function that sorts the distance measurement results, that is, sorts all similar distances to $\left(\mathbf{z}_j,\mathbf{x}_i\right)$. The mathematical expression of the function $\operatorname{sign}(\bullet)$ is shown in Formula (33):

$$\operatorname{sign}\left(d\left(\mathbf{x}_i,\mathbf{z}_j\right),d\left(\mathbf{x}_i,\mathbf{z}_s\right)\right)=\begin{cases}1, & if \quad d\left(\mathbf{x}_i,\mathbf{z}_j\right)>d\left(\mathbf{x}_i,\mathbf{z}_s\right)\\ 0, & else\end{cases} \tag{33}$$

It can be seen from the calculation principles of (32) and (33) that the corresponding distance $\sum_s \operatorname{sign}\left(d\left(\mathbf{x}_i,\mathbf{z}_j\right),d\left(\mathbf{x}_i,\mathbf{z}_s\right)\right)$ is less than the number of samples $d\left(\mathbf{x}_i,\mathbf{z}_j\right)$. In the improved method, we determine the top k training data samples as similar samples of test individuals, thus forming potential positive samples. Using the projection shadow space training method in 3.2, the potential positive samples are brought into the model for re-training, thereby obtaining an improved model in Formula (34)

$$\arg\max_{\mathbf{w}} J(\mathbf{w})=\frac{\mathbf{w}^T\left(\left\|\mathbf{w}^T\boldsymbol{\mu}_1-\mathbf{w}^T\boldsymbol{\mu}_0\right\|_2^2\right)\mathbf{w}}{\mathbf{w}^T\left(\sum_1+\sum_0+\sum_1'\right)\mathbf{w}} \tag{34}$$

$\sum_1'$ is the covariance of the potentially positive sample, $\mathbf{W}'=(\mathbf{w}_1',\mathbf{w}_2',\cdots,\mathbf{w}_m')$ as an improved feature transformation method is obtained according to the model (34).

4.2 Experimental Results and Comparative Analysis

In order to verify the effect of the improved method, we use the same test method as in 3.3.1 to calculate the recognition accuracy of the improved method. At the same time, we compare with other different methods to more fully illustrate the effect of the improved method, as shown in Table 1. From the results in the table, it can be concluded that the improved method designed by us has the highest recognition accuracy. L2 method uses Euclidean distance to identify on the original data. If only the original high-dimensional features are used, the recognition degree is only 46.6%, and the accuracy is very low. This proves from the side that high-dimensional spectral data is very complex and has insufficient resolution performance in our research task, so it is not advisable to use it directly to identify samples. However, using PCA to extract features and then using Euclidean distance for recognition will reduce the accuracy, which is mainly because this method is unsupervised and the extracted features have no purpose, only considering the differences of data, but not the relationship between similar data and different data.

Table 1. Comparison of Recognition Accuracy of Different Algorithms

Method	Accuracy
L2	46.6%
PCA	38.1%
Multi-classification LDA	94.3%
Binary LDA	95.1%
Methods of Improvements	**96.2%**

5. Summary

In this paper, the identification of Chinese herbal medicines is studied. The identification technology of medicinal materials is of great significance in traditional Chinese medicine. Because there are great differences in medicinal properties and pharmacology among different kinds, different producing areas, and different years of Chinese herbal medicines, it is very crucial to accurately identify the types, producing areas, and quality of Chinese herbal medicines for doctors to accurately use drugs, supervise their quality and patients' safety. However, in real life, the identification of Chinese herbal medicines usually depends on experienced doctors, who evaluate their own experience and identify Chinese herbal medicines according to their color, shape, smell, and other characteristics. This traditional identification method of Chinese herbal medicines is very demanding on personnel, which obviously cannot meet the needs of practical application. Therefore, based on the infrared spectroscopy characteristic data of Chinese herbal medicines, this paper designs a Chinese herbal medicine identification technology by using the artificial intelligence method of machine learning. In this paper, the classification of spectral characteristic data of Chinese herbal medicines is studied from two aspects, that is, unsupervised and supervised learning. Firstly, an improved K-means clustering algorithm based on the Gaussian distribution model is established for the unsupervised spectral classification of Chinese herbal medicines. This method "over-classifies" the sample data by the K-means clustering algorithm and further classifies the data by the Gaussian mixture model, thus realizing unsupervised classification of Chinese herbal medicines. In addition, we focus on the supervised recognition method. Based on the original multi-classification method, we transform the difference vector into two classification problems. Combining the two classification methods, we use the distribution characteristics of recognition samples to carry out semi-supervised learning, thus improving the recognition accuracy and designing a very effective model.

References

[1] Ni, L. J., Luan, S. R. & Zhang, L. G. (2016). Exploration of rapidly determining quality of traditional Chinese medicines by (NIR) spectroscopy based on internet sharing mode. Chinese Journal of Materia Medica, 41(19), 3520-3527.

[2] Tao, O., Lin, Z. Z., Zhang, X. B. et al. (2014). Research on identification model of Chinese herbal medicine by texture feature parameter of transverse section image. Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology, 16(12), 2558-2562.

[3] Wang, H. Y., Luo, H. & Sun, M. (2004). Stability and reliability of RAPD in identification of Chinese herbal medicines. Journal of Chinese Medicinal Materials, 27(1), 4.

[4] Jiang, D. C., Wang, Y. S. & Weng, L. L. (2006). Spectral identification of commonly used traditional Chinese medicines. Beijing: Chemical Industry Press.

[5] Wu, Y. J., Li, W., Xiang, B. R. et al. (2001). Identification of traditional Chinese medicine Baizhi with near-infrared spectrum. Journal of Chinese Medicinal Materials, 24(1), 3.

[6] Liu, S. H., Zhang, X. G., Zhou, Q. et al. (2005). Use of FTIR and pattern recognition to determine geographical origins of Chinese medical herbs. Spectroscopy and Spectral Analysis, 25(6), 4.

[7] Liu, S. H., Zhang, X. G., Zhou, Q. et al. (2006). Determination of geographical origins of Chinese medical herbs by NIR and pattern recognition. Spectroscopy and Spectral Analysis, 26(6), 4.

[8] Ding, N. Y., Li, W., Feng, X. H. et al. (2008). Classification and identification of Chinese traditional medicines with NIR diffuse reflection. Computer and Applied Chemistry, 25(4), 499-502.

[9] Lin, F. B., Pang, Q. C., Ma, J. et al. (2010). Identifications of herbal medicines based on spectral imaging detection. Journal of Applied Optics.

[10] Lin, F. B., Pang, Q. C., Ma, J. et al. (2010). Identifications of herbal medicines based on spectral imaging detection. Journal of Applied Optics, (02), 117-121.

[11] Perlibakas, V. (2004). Distance measures for PCA-based face recognition. Pattern Recognition Letters, 25(6),711-724.

[12] Subasi, A. & Gursoy, M. I. (2010). EEG signal classification using PCA, ICA, LDA and support vector machines. Expert Systems with Applications, 37(12), 8659-8666.

[13] Ding, X. L., Qi, Q. C. & Fang, L. (2021). Classification and recognition of traditional Chinese medicine based on fisher discriminant analysis. Journal of Chifeng University(Natural Science Edition), 37(11), 4.

[14] Hu, Y. et al. (2022). Research progress of near infrared spectroscopy in quality control of traditional Chinese medicine. Journal of Anhui Agricultural Sciences, 50(01), 8-11.

[15] Zheng, J. et al. (2021). Identification of armeniacae semen amarum and persicae semen from different origins based on near infrared hyper-spectral imaging technology. China Journal of Chinese Materia Medica, 46(10), 7.

[16] Xie, M. Y. et al. (2021). Identification of radix achyranthis bidentatae from different producing areas by infrared spectroscopy and infrared derivative spectrophotometry. Journal of Anhui University of Chinese Medicine, 40(06), 86-91.

[17] Wu, C. (2014). Research on identification method of traditional Chinese medicine based on deep learning. Chengdu: Master Degree Thesis of Sichuan University.

[18] Zhou, Y. J. et al. (2014). Identification of Chinese herbal medicines based on Terahertz spectroscopy analysis. Spectroscopy and Spectral Analysis, 34(07), 1840-1843.

[19] Wang, X. R. et al. (2022). Quantification of "Cold-Hot" medicinal properties of Chinese medicines based on primary metabolites and Fisher's analysis. Computational and Mathematical Methods in Medicine. 5790893.

[20] Chen, Y. J. et al. (2009). Chinese traditional medicine recognition by support vector machine (SVM) terahertz spectrum. Spectroscopy and Spectral Analysis, 29(9), 2346-2350.

[21] Wang, L. L., Wang, C., Pan, Z. F., Sun, Y. & Zhu, X. Y. (2011). Application of pyrolysis-gas chromatography and hierarchical cluster analysis to the discrimination of the Chinese traditional medicine Dendrobium candidum Wall. ex Lindl. Journal of Analytical and Applied Pyrolysis, 90(1), 13-17.