Original Article
Food Science and Technology

ISSN 0101-2061 (Print)
ISSN 1678-457X (Online)

DOI: https://doi.org/10.1590/fst.73022

Rapid identification of green tea varieties based on FT-NIR spectroscopy and LDA/QR

Jiabao WANG[1], Xiaohong WU[2,3], Jun ZHENG[4*], Bin WU[5]

Abstract

There are many substances beneficial to human body in tea. In this study, we put forward innovative strategies to quickly and harmlessly identify Chinese green tea varieties. Near-infrared (NIR) spectrometer was used to collect NIR spectral data of tea samples, and the data were preprocessed by Savitzky-Golay (SG) filter to eliminate noise of spectral data. Three feature extraction algorithms: principal component analysis (PCA) combined with linear discriminant analysis (LDA), LDA/QR, generalize singular value decomposition (GSVD) were performed to decrease the dimension and compress the spectral data. Finally, k-nearest neighbor (*k*NN) classifier was utilized to classify the samples according to the NIR spectra of the samples. PCA combined with LDA, GSVD and LDA/QR had the classification accuracy rates 94.19%, 91.86% and 98.84%, respectively. So, LDA/QR showed the highest classification accuracy in classification of NIR spectra of tea samples. We believe that the combination of NIR spectroscopy and feature extraction algorithms can quickly identify the types of tea samples. This method may have the potential to identify other varieties of food.

Keywords: Chinese green tea; near-infrared spectroscopy; Savitzky-Golay filter; discriminant analysis; LDA/QR.

Practical Application: Near infrared spectroscopy is combined with feature extraction algorithm to detect the types of tea. It has many advantages that traditional detection methods do not have, such as fast, convenient and non-destructive.

1 Introduction

Tea is one of the most popular drinks, which has been in great demand all over the world and has occupied a high position in people's hearts (Dekant et al., 2017). The reason why tea has always been in great demand is not only its ability to cultivate the sentiment, but also its beneficial nutrients, such as tea polyphenols, flavonoids, amino acids, alkaloids, proanthocyanidins and so on (Ahammed & Li, 2022). In addition, tea has many beneficial functions. For example, by-products of green tea processing, such as flavonols and polysaccharides, can prevent the absorption and accumulation of lipids in mild cells in the intestine (Yoo et al., 2020). Epigallocatechin gallate (EGCG) is the most active catechin in tea polyphenols and has anticancer properties (Ju et al., 2007). Green tea extract coated with active film has strong antioxidant capacity, so it has broad prospects in the field of meat preservation (Song et al., 2020). White tea has strong antioxidant activity, so it can reduce inflammation. The main components of tea polyphenols and EGCG have the potential free radical scavenging effects and good antibacterial effects (Xia et al., 2021). In addition, carbohydrates extracted from Pu'er tea have hypoglycemic effect by inhibiting glucosidase (Lin et al., 2019). Therefore, in recent years, many scholars began to carry out various studies on tea, such as quantitative analysis of tea components to identify the quality of tea (Pang et al., 2022), the impact of tea nutrients on reducing blood glucose (Uğur et al., 2022). It is a normal phenomenon that there are all kinds of tea with different labels. However, the price gap is huge. People always think that the more expensive the better, but it is not the case. It is common for merchants to mix and sell inferior tea with high-quality tea to gain more profit, but this leads to uneven quality of tea in the market, which greatly damages the rights and interests of consumers and affects the taste of tea. Therefore, it is of great significance to explore a fast and accurate recognition method for discrimination of tea samples.

In order to solve the problem of tea adulteration in the market, researchers have tried some methods. For example, the least absolute shrink and selection operator method (LASSO) and analysis of variance (ANOVA) were used to identify the origin of tea (Pan et al., 2022). Because polyphenols would be oxidized when they are exposed to air during the fermentation stage of tea leaves, which would produce color, aroma, taste and other properties (Zhu et al., 2017), many researchers used chemical methods to classify tea according to this phenomenon. The information collected by electronic nose and electronic tongue can be directly spliced and fused for qualitative and quantitative analysis of tea quality grade (Xu et al., 2021). The volatile components of different grades of tea were analyzed by gas chromatography-mass spectrometry (GC-MS) (Qin et al.,

Received 25 June, 2022
Accepted 05 Aug., 2022
[1] *Institute of Talented Engineering Students, Jiangsu University, Zhenjiang, China*
[2] *School of Electrical and Information Engineering, Jiangsu University, Zhenjiang, China*
[3] *High-Tech Key Laboratory of Agricultural Equipment and Intelligence of Jiangsu Province, Jiangsu University, Zhenjiang, China*
[4] *Department of Electrical and Control Engineering, Research Institute of Zhejiang University-Taizhou, Taizhou, China*
[5] *Department of Information Engineering, Chuzhou Polytechnic, Chuzhou, China*
*Corresponding author: dbzj@netease.com

2013). High performance liquid chromatography diode array detection (HPLC-DAD) method was used for rapid quantitative analysis of ten main components in West Lake Longjing samples (Gu et al., 2020). In addition, some researchers used the changes of polyphenols and coffee concentration in tea to distinguish the tea varieties. Capillary electrophoresis was utilized to measure the content of tea caffeine and catechins (Lee & Ong, 2000). However, these physical and chemical indicators need to be carried out in the laboratory. The disadvantages are cumbersome steps, high cost and long experimental time, which are not suitable for large-scale processing of tea.

Near-infrared spectroscopy (NIRS) is an electromagnetic wave between visible light and mid-infrared light. Near-infrared (NIR) spectroscopy which mainly refers to the absorption of hydrogen-containing groups is the frequency doubling and combined spectral bands of molecular vibration spectroscopy (Liu et al., 2007; Yuan & Dou, 2004; Tao et al., 2016; Kahriman & Egesel, 2016). It contains rich information on the composition and molecular structure of most types of organic substances, and is non-destructive, low-cost, and detectable. Near-infrared spectral information can be used as a technical analysis method for nanoparticles (Brigger et al., 2000). Duo to its high speed and other advantages, it has made possible the wide application as a detection technology in the field of food analysis. In these two literatures (Xu, 2018; Wang et al., 2017), researchers used near-infrared spectroscopy for identification of water pollution and liquid food, respectively. NIRS was used to determine the protein content of sweet potatoes to screen high-protein sweet ones (Laurie et al., 2020). The multivariate classification model of NIRS was applied to the study of detecting adulteration of goat milk (Teixeira et al., 2020). Near infrared spectroscopy was combined with a possibility fuzzy C-means (PFCM) algorithm based on similar particle swarm optimization (SPSO) for apple classification (Xu et al., 2022). Some researchers had proposed an improved discriminant information extraction algorithm, called weighted global fuzzy uncorrelated discriminant transform (WGFUDT), and combined it with near-infrared spectroscopy for grade recognition of two kinds of green tea (Huangshan Maofeng tea and mee tea) (He et al., 2022). Detecting the content of melamine and urea in fat-filled milk powder (FMP) prepared with different vegetable oils through NIRS confirmed and distinguished whether FMP was adulterated (Ejeahalaka & On, 2020). In terms of tea identification, some researchers have completed phased progress by near-infrared spectroscopy. Luypaert et al. (2003) applied infrared spectroscopy and PLS to quickly assess the quality of green tea (Luypaert et al., 2003). The wavelength range of the near-infrared spectrum is 780~2500 nm, so the data collected has the characteristics of high dimension, crest and trough, spectral overlap etc. Spectral analysis techniques will combine some machine learning methods to make the experimental results more accurate and clearer, such as fuzzy C-mean (FCM), principal component analysis (PCA), linear analysis (LDA), decision tree (TD). Furthermore, near-infrared spectroscopy has been applied to tea identification, such as: NIRS combined with support vector machine (SVM) could effectively, quickly and easily identify tea varieties (Zhao et al., 2006). NIRS using multivariate calibration analysis could accurately determine the active ingredients of tea (Chen et al., 2006). Li & He (2008) performed principal component analysis on the data and used artificial neural network to distinguish tea tree varieties after proper spectral preprocessing (Li & He, 2008). Chen et al. made an effort to calculate the first five principal components as the input of the SVM classifier according to the difference in spectral features and PCA. In training, the three types of tea reached 90%, 100%, and 93.33% recognition accuracy (Chen et al., 2007).

LDA, LDA/QR and generalize singular value decomposition (GSVD) can be applied to extract feature and compress data. However, when using LDA, PCA needs to be used to reduce the dimension of the data, otherwise small sample problems will occur (Zhong & Ban, 2022). LDA/QR, a LDA based dimension reduction algorithm (Ye & Li, 2004), achieves the efficiency by introducing QR decomposition on a small-size matrix, while keeping competitive classification accuracy. GSVD is a feature extraction algorithm that can avoid small sample size problem (Howland & Park, 2004). In this study, we tried to make use of NIRS to detect and classify Chinese green tea varieties using PCA, LDA, LDA/QR and GSVD.

In this work, the NIR spectroscopy and three feature extraction algorithms were combined to classify Chinese green tea varieties. The detailed steps are as follows: (1) Obtain the NIR spectra of tea using Antaris II FT-NIR spectrometer; (2) Preprocess the NIR spectra via Savitzky-Golay (SG) filter; (3) Perform PCA + LDA, LDA/QR and GSVD for extracting the features of NIR spectra; (4) Classify the data with k-nearest neighbor (*k*NN) classifier; (5) Compare the classification results of the three pattern recognition algorithms.

2 Materials and methods

2.1 Sample preparation

Four varieties of tea (Yuexi Cuilan, Luan Guapian, Shiji Maofeng, and Huangshan Maofeng) were purchased in the same market. Because each variety of tea samples possessed 65 experimental samples, there were 260 samples in total. The following principles apply to the selection of samples. For the same variety, it must be selected from the same region and processed in the same way. For the different varieties, the dates of manufacture are the same approximately. Firstly, all tea samples were crushed through a 40-mesh sieve with a coffee grinder before carrying on the NIR spectral experiment. Secondly, 5 g sample was weighed as an experimental sample.

2.2 NIR spectra collection

This experiment applied Antaris II FT-NIR spectrometer to collect NIR spectra of tea samples. Because the spectrum is more sensitive to the external environment, laboratory temperature and relative humidity remain constant as far as possible in the process of experiment. The spectrometer was preheated for one hour after being started up. The NIR spectra of tea were collected by reflecting integral ball mode, and each tea sample was scanned 32 times. The wavelength range was 4000-10000 cm^{-1}. The distance of optical wave scanned was 3.587 cm^{-1}. And the data of the collected NIR spectra was 1557 dimensions. Each sample was sampled 3 times to reduce the experimental error.

The mean value of three results was used as the following experimental data.

2.3 Spectral processing method

Through the analysis of tea samples, the information of spectral principal components can be obtained. However, the raw spectral data can be possible to be noisy due to some measurement errors or interference. Due to the presence of these noises, it is possible to have a negative impact on the classification results. SG filter can be used to identify time-varying finite impulse response (FIR) systems (Niedźwiecki et al., 2021). In order to eliminate the influence of noise, SG filter smoothing algorithm was used to process the spectral data.

2.4 Principle component analysis

Principle component analysis (PCA) is an unsupervised method used to reduce the dimension of data and extract major features (Cao et al., 2022). When PCA is used to cut down the dimension of data, some information will be lost, but the most important data will be extracted, and it will also play a role on reducing noise. In addition, using PCA can avoid the problem of long running time of the model due to the high dimensional data. When PCA is performed, the covariance matrix is decomposed to obtain the eigenvectors with the largest eigenvalues. It is generally believed that when the cumulative contribution rate of the selected principal components reaches more than 80%, the number of principal components is considered to be sufficient.

2.5 Feature extraction algorithms

In this study, three feature extraction algorithms (LDA, GSVD and LDA/QR) were used to obtain the discriminant information from the samples' spectral data. Among them, PCA was required to process data before using LDA.

Linear discriminant analysis

LDA is committed to maximizing the distances between classes and minimizing the distances within classes. In this section, make $X=\{x_i\}, i=1,2,...,n$, be a set of samples that is the d-dimensional data. $\omega_1, \omega_2, ..., \omega_c$ are known c data clusters ($x_i \in \omega_j, j \in \{1,2,...,c\}$). The matrix that transforms the high-dimensional data into the low-dimensional data can be obtained by solving the following equation.

The Fisher optimal discriminant function is (Equation 1):

$$\max J_1(G) = \frac{G^T S_b G}{G^T S_w G} \tag{1}$$

Where S_b is the scatter matrix between the data clusters (Equation 2):

$$S_b = \sum_{i=1}^{c} P(\omega_i)(m_i - m_0)(m_i - m_0)^T \tag{2}$$

S_w is the scatter matrix within the data clusters (Equation 3):

$$S_w = \sum_{i=1}^{c} P(\omega_i) E\left\{\frac{(x-m_0)(x-m_0)^T}{\omega_i}\right\} \tag{3}$$

In Equations 2-3, m_0 represents the average value of the whole sample; $m_i, i \in \{1,2,...,c\}$ is the average value of samples belonging to class i; $P(\omega_i)$ is the prior probability of ω_i; generally, $P(\omega_i)$ is equal to $\frac{1}{c}$.

Equation 1 can be solved as an eigenvalue problem (Equation 4):

$$S_w^{-1} S_b G = \lambda G \tag{4}$$

Based on the above calculation process, the discriminant values and the corresponding discriminant vectors can be obtained. Obtain the largest p discriminant values and then sort them in descending order $\lambda_1, \lambda_2, ..., \lambda_p$, and their corresponding discriminant vectors $G_1, G_2, ..., G_p$. The following linear transformation projects the data from R^d to R^p (Equation 5):

$$Y = \left[G_1 G_2 ... G_p\right]^T X \tag{5}$$

LDA/QR

LDA/QR algorithm is a derived algorithm from LDA, and it can solve the small sample size problem of LDA. On the basis of LDA, LDA/QR adds the operation of using QR decomposition on small matrix, which is also the essence of LDA/QR algorithm (Ye & Li, 2004). And the time complexity of the algorithm is linear with the dimension and data size. The calculation of the transformation matrix is carried out on the small dimension matrix, so the stability of the algorithm is quite excellent.

LDA/QR has two stages. The essence of the first stage is to maximize the distances between classes. The first stage aims to solve the following optimization problem (Equation 6):

$$\max J_2(G) = G^T S_b G \; where \; G^T G = I \tag{6}$$

The essence of this optimization problem is to maximize the distances between classes. The solution of this problem can be obtained by calculating the eigenvalue problem of matrix S_b. It can also be obtained equivalently by QR decomposition of matrix H_b. H_b can be calculated from the following equation (Equation 7):

$$H_b = \left[\sqrt{N_1}(m_1 - m), ..., \sqrt{N_k}(m_k - m)\right] \in R^{n \times k} \tag{7}$$

Where N_i represents the size of the ith data cluster; m_i represents the centroid of the ith data cluster; m represents the global centroid of the whole data set; k is the number of data clusters.

In addition, the matrix H_w needs to be calculated from the following equation (Equation 8):

$$H_w = \left[A_1 - m_1 \cdot e_1, ..., A_k - m_k \cdot e_k\right] \in R^{n \times k} \tag{8}$$

Where A_i represents the ith data cluster matrix and $e_i = (1,...,1) \in R^{1 \times N_i}$.

The second stage aims to minimize the distances within the class. The relaxation scheme is used to coordinate the scatter information within clusters. The final optimization problem is exactly the same as the traditional LDA algorithm, but compared with LDA, the matrix that needs to be calculated is very small, so it has good stability in efficiency.

To sum up, LDA/QR can be described in the following steps: (1) Compute the matrix H_b and H_w. (2) Compute QR-decomposition of H_b. (3) Compute the t eigenvectors w_i of $S'^{-1}_b S'_w$, with increasing eigenvalues, where $S'_b = Q^T S_b Q$ and $S'_w = Q^T S_w Q$. (4) Matrix $G = QW$, where $W = [w_1,...,w_t]$ be calculated from the following optimization problem (Equation 9):

$$min\, J_3(W) = \frac{W^T S'_w W}{W^T S'_b W} \quad (9)$$

Based on the above calculations, the following linear transformation projects the data from R^n to R^t (Equation 10):

$$Y = G^T X \quad (10)$$

Generalize singular value decomposition

Before using LDA, if the dimension of the sample is much larger than the number of samples, PCA should be used to reduce the dimension of data, or small sample size problem will happen. LDA can only be used for nonsingular linear transformation. Such a standard requires that the matrix S_w must be a nonsingular matrix. However, matrix S_w is often a singular matrix in multiple application scenarios. The original reason for this phenomenon is that the dimension of the data itself is much larger than the number of samples, that is, the small sample size problem (Fukunaga, 1990). When matrix S_w is a singular matrix, errors will occur in computing S_w^{-1}. For this reason, some researchers used a positive pseudoinverse matrix S_w^{+} to avoid solving S_w^{-1}. In order to solve the small sample size problem, generalize singular value decomposition (GSVD) extends discriminant analysis and theoretically provides the optimal data dimensionality reduction. When S_w is a nonsingular matrix, the classical discriminant analysis extends its solution to the generalized eigenvalue problem. The problem is reformulated by generalized eigenvalue decomposition, and the scope of use is extended to the case when S_w is a singular matrix. The transformation matrix G is obtained by calculating the matrices H_b and H_w from the data set according to Equations 7-8, respectively, and then using the complete orthogonal decomposition and calculating SVD on the basis of this. The detailed description of GSVD can be seen in the article (Howland & Park, 2004).

2.6 Classification algorithm

*k*NN algorithm is an excellent classification algorithm, which is simple and effective. Therefore, in this study, *k*NN algorithm was selected as the classifier. In *k*NN algorithm, the category of a sample data of unknown category depends on the nearest *k* known sample data. In order to avoid ambiguity in the classification results, the value of *k* is generally odd.

2.7 Design of the identification system

Figure 1 shows the structure of this identification system for authentication of tea varieties. The spectral data of tea samples were collected by Antaris II FT-NIR spectrometer and SG filter was used to eliminate the noise in spectral data. Then, it was divided into two cases. The first was to apply PCA to reduce the dimension of the data, and then applied LDA to extract and compressed the features of the data. The second was to use LDA/QR or GSVD directly for feature extraction and compression. Finally, *k*NN classifier was used to determine the category of each test sample, and the accuracy of the recognition system was obtained.

3 Results and discussion

3.1 Spectral data processing

In this study, the wavelength range of the collected spectra was 4000-10000 cm^{-1}, and the entire spectral data were processed. There may be noise in the initial spectral data, so SG filter function sgolayfilt() of MATLIB was applied to eliminate the noise in the spectral data. The spectral data after SGF processing is presented in Figure 2.

3.2 Principle component analysis and linear discriminant analysis

First, the spectral dimension of tea samples was 1557, which was high and contained a lot of redunant information. So PCA was used to reduce the dimensions of the high dimensional spectral data. The near-infrared spectral data of tea samples were projected onto four feature vectors to obtain the 4-dimensional data. The dimensional reduction of PCA was used to obtain a

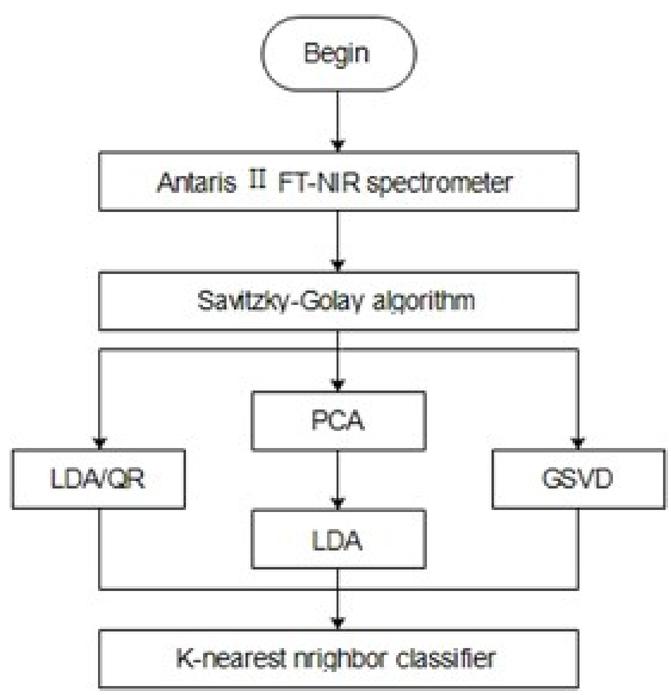

Figure 1. The structure of the identification system.

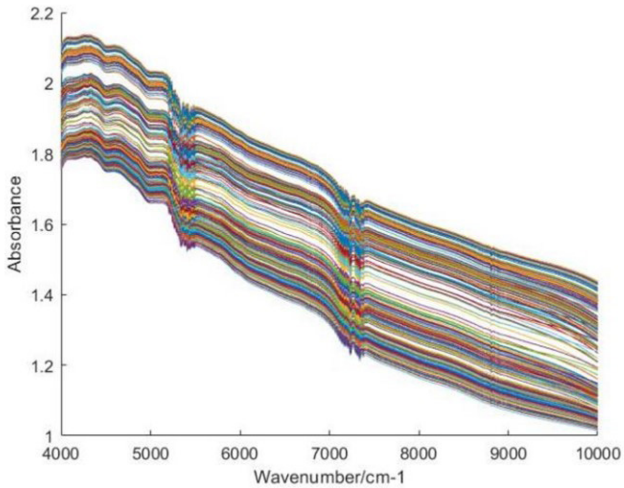

Figure 2. Spectral data by Savitzky-Golay filter.

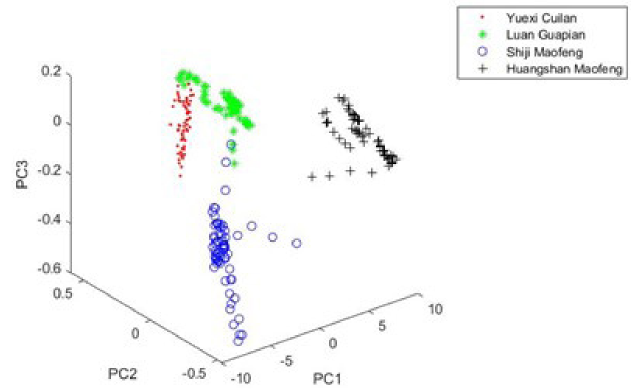

Figure 3. The scores plot of PCA.

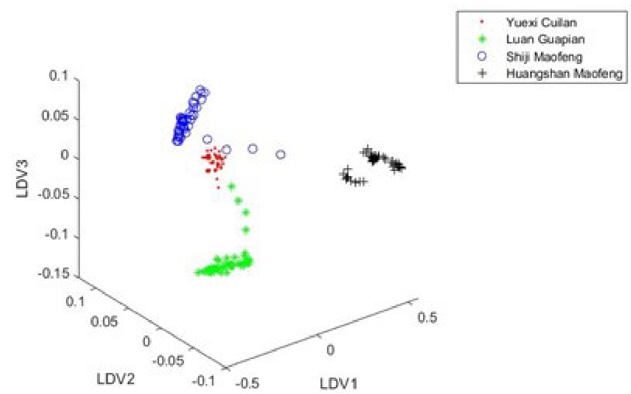

Figure 4. The scores plot of LDA.

260*4 sample set. Then, LDA was applied to continue processing the results of PCA. A new test set changed from 172*1557 to 172*3 was realized by LDA to achieve the purpose of dimensional reduction.

For PCA, eigenvalues of co-variance matrix can be obtained through the training data set and to be arranged in descending order. The cumulative contribution rate of the first four principal components was higher for the training data set, so we selected the first four principal components of PCA. The first four eigenvalues were: $\lambda_1 = 7429.8237$, $\lambda_2 = 30.2302$, $\lambda_3 = 1.2024$, $\lambda_4 = 0.2781$. The first three principal components (PCA1, PCA2, and PCA3) accounted for 99.99% of the total variance. Therefore, the first three principal components were selected to draw the score plots of PCA. The scores plot of PCA are shown in Figure 3.

After the data were processed by PCA, Shiji Maofeng and Luan Guapian had no overlapped sample, and the classification effect was good. However, it was not difficult to see from PCA scores plot, Yuexi Cuilan and Luan Guapian had some overlapped data, and the two kinds of tea samples were difficult to be distinguished. The overlapped data can affect the accuracy of classification. Therefore, LDA was continued to extract feature vectors, and three discriminant vectors were extracted. The calculation results of their eigenvalues were as follows: $\lambda_1 = 45.2504$, $\lambda_2 = 16.3784$, $\lambda_3 = 3.5351$. In order to visualize different categories, we projected the training data into the three-dimensional coordinates through a linear transformation constructed by three discriminant vectors of LDA to obtain the LDA scores plot, as presented in Figure 4.

From the comparison between Figure 3 and Figure 4, it can be seen that after LDA processing, there are less overlapped data. Obviously, Huangshan Maofeng samples were far from the other three kinds of tea samples. In addition, Yuexi Cuilan samples and Shiji Maofeng samples were close to each other, and this heightened the difficulty for the classification of them.

3.3 Generalize singular value decomposition and LDA/QR

The idea of applying GSVD and LDA/QR in feature extraction of spectral data is similar to LDA, which maximizes the distance between classes and minimizes the distance within class. Through some mathematical transformations, the high-dimensional data can be projected into low-dimensional space without causing information loss as much as possible. In the low-dimensional space, the distance between data points belonging to different kinds of tea will become larger, and the distance between data points belonging to the same kind of tea will become smaller, so as to better distinguish different kinds of tea. Figure 5 and Figure 6 are the scores plot of feature extraction using GSVD and LDA/QR, respectively.

Through observation, it can be found that the effect of feature extraction using GSVD and LDA/QR is quite excellent, and the distance or boundary between different species is quite obvious. From the results of GSVD, it can be seen that although there are obvious boundaries between different kinds of tea, there are still some overlaps at the junction. From the results of LDA/QR, we can see that compared with GSVD and LDA, GSVD has the best classification effect, and there are the considerable distances

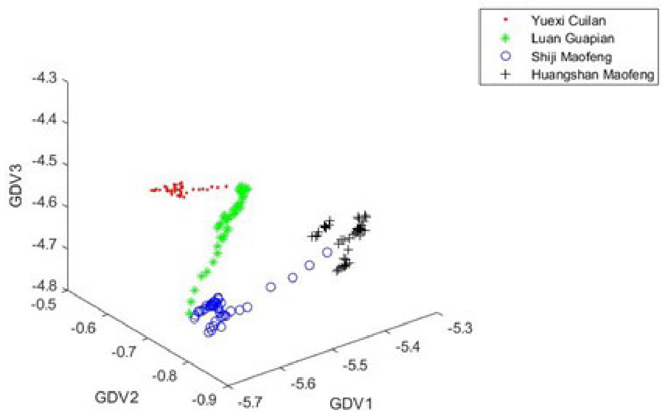

Figure 5. The scores plot of GSVD.

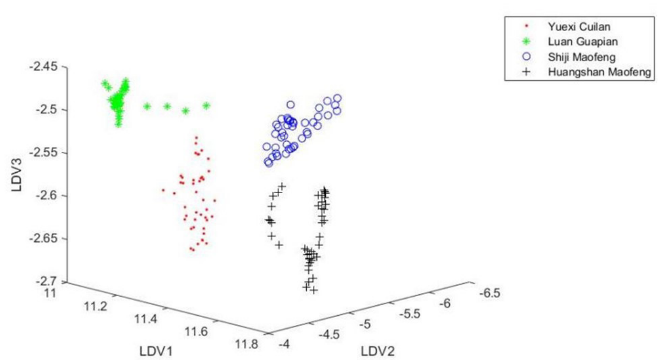

Figure 6. The scores plot of LDA/QR.

between different kinds of data clusters, which clearly reflect that LDA/QR is a very excellent feature extraction algorithm.

3.4 Classification using KNN classifier

Finally, the *k*NN classifier was applied to test the three feature extraction algorithms. Tables 1-3 illustrate the classification accuracy results of three different feature extraction algorithms after using *k*NN classifiers with different *k* values. According to the classification results, it is effective to use these three pattern recognition algorithms for feature extraction and *k*NN algorithm as a classifier when processing the NIR spectral data of tea samples.

When using PCA plus LDA, it is obvious that when $k = 1$ and 3, the classification accuracy can be maintained at more than 90% duo to randomness. However, with the increase of *k*, the classification accuracy rapidly decreased to about 75%. In contrast, when using LDA/QR, the classification accuracy can be maintained at more than 95%. When using GSVD, the classification accuracy can be maintained above 90%. In terms of classification accuracy, LDA/QR has the best performance.

In addition, the average running time of GSVD is the shortest, maintained at about 0.005s. The average running time of LDA and LDA/QR is the same, both of which remain between 0.02 s and 0.025 s, about 4 to 5 times that of GSVD. In terms of running time, GSVD is obviously better than the other two algorithms.

Table 1. The accuracy rate, *k* and running time using PCA plus LDA.

Method	k	Accuracy Rate	Running Time (s)
PCA + LDA	1	94.19%	0.020624
	3	94.77%	0.023061
	5	76.74%	0.024791
	7	76.74%	0.023454
	9	76.16%	0.022810

Table 2. The accuracy rate, *k* and running time using LDA/QR.

Method	k	Accuracy Rate	Running Time (s)
LDA/QR	1	98.84%	0.022036
	3	98.26%	0.023456
	5	97.09%	0.025191
	7	97.09%	0.024521
	9	96.51%	0.024540

Table 3. The accuracy rate, *k* and running time using GSVD.

Method	k	Accuracy Rate	Running Time (s)
GSVD	1	91.86%	0.003845
	3	91.86%	0.005816
	5	91.86%	0.004744
	7	91.86%	0.005334
	9	91.86%	0.005075

Considering the classification accuracy and running time, if this technology is applied to life, for the test of a single sample, although the running time of LDA/QR is longer than that of GSVD, the running time of 0.025 s can be tolerated, and the classification accuracy of LDA/QR is higher than that of GSVD, so LDA/QR is still considered to be the best feature extraction algorithm.

In this work, we explored the use of PCA + LDA plus *k*NN, GSVD plus *k*NN and LDA/QR plus *k*NN combined with near-infrared spectroscopy to classify tea samples. The major idea of GSVD and LDA/QR is the same as that of LDA, and classification is realized by minimizing the distance within classes and maximizing the distance between classes. However, LDA directly solves the eigenvector through the optimization problem of Equation 1. Before using LDA, PCA needs to be used to reduce the dimension of spectral data and the matrix S_w is required to be a nonsingular matrix, which is not met in most application scenarios; GSVD converts the problem into a problem of solving generalized eigenvalues equivalently through a series of mathematical operations, so as to avoid the process of solving the inverse matrix of S_w, and on this basis, complete orthogonal decomposition and calculating SVD are introduced to solve the problem. LDA/QR is divided into two stages: the first stage maximizes the distances between classes and makes some preparations for the second stage; The second stage is to minimize the within-class distances; The second stage aims to minimize the within-class distances and introduce QR-decomposition, so as to avoid the process of calculating the inverse matrix of S_w . The calculation of LDA/QR is carried out on a small matrix, which makes LDA/QR have good numerical stability and good efficiency (the efficiency refers to high classification accuracy).

Some researchers used *k*NN to classify and quantitatively analyze chlorpyrifos residues on tea, and the classification accuracy reached more than 90% (Zhu et al., 2018). In some cases, the effect of using *k*NN as a classifier is quite significant. In these two literatures (Wu et al., 2019; Xin et al., 2018), *k*NN and support vector machine (SVM) were used as classifiers, and the classification accuracy of *k*NN was lower than that of SVM, which also showed that the use of *k*NN had certain limitations. Combined with the results of this study, the classification success rate of *k*NN may be related to the feature extraction algorithm, and this conjecture can be gradually explored in future research. In addition, when the training set is too large, *k*NN algorithm may not be used anymore, because *k*NN algorithm needs to calculate the distance between test samples and all training samples, and the running time will increase with the increase of training set. The method proposed in this paper can be used for determination of tea varieties. Although some traditional testing methods may have higher accuracy, they often consume a lot of time and are unable to carry out nondestructive testing, which is undoubtedly a waste of resources. In addition, traditional detection methods require the detection personnel to have a certain knowledge reserve before they can complete the detection of samples.

4 Conclusion

This paper discussed the application of three different feature extraction algorithms to the classification and detection of tea species. Considering comprehensively, it is considered that LDA/QR algorithm has the highest classification accuracy, which can reach 98.84%. Based on such experimental results, if the feature extraction algorithm is applied to the actual scene, the human and material resources consumed by tea classification and detection can be greatly reduced. Moreover, this method is a fast, accurate, green detection method, which will not produce harmful distances in the process of classification. We believe that in the future, if portable near-infrared devices can be combined with this technology, it will bring good benefits.

Acknowledgements

The authors would like to thank Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), the Undergraduate Innovation and Entrepreneurship Training Program of Jiangsu Province (202110299089Z), the Talent Program of Chuzhou Polytechnic (YG2019026 and YG2019024) and Key Science Research Project of Chuzhou Polytechnic (YJZ-2020-12) for funding this research.

References

Ahammed, G. J., & Li, X. (2022). Hormonal regulation of health-promoting compounds in tea (Camellia sinensis L.). *Plant Physiology and Biochemistry*, 185, 390-400. http://dx.doi.org/10.1016/j.plaphy.2022.06.021. PMid:35785551.

Brigger, I., Chaminade, P., Desmaële, D., Peracchia, M. T., d'Angelo, J., Gurny, R., Renoir, M., & Couvreur, P. (2000). Near infrared with principal component analysis as a novel analytical approach for nanoparticle technology. *Pharmaceutical Research*, 17(9), 1124-1132. http://dx.doi.org/10.1023/A:1026465931525. PMid:11087046.

Cao, H. S., Sun, P. W., & Zhao, L. (2022). PCA-SVM method with sliding window for online fault diagnosis of a small pressurized water reactor. *Annals of Nuclear Energy*, 171, 109036. http://dx.doi.org/10.1016/j.anucene.2022.109036.

Chen, Q. S., Zhao, J. W., Huang, X. Y., Zhang, H. D., & Liu, M. H. (2006). Simultaneous determination of total polyphenols and caffeine contents of green tea by near-infrared reflectance spectroscopy. *Microchemical Journal*, 83(1), 42-47. http://dx.doi.org/10.1016/j.microc.2006.01.023.

Chen, Q., Zhao, J., Fang, C. H., & Wang, D. (2007). Feasibility study on identification of green, black and Oolong teas using near-infrared reflectance spectroscopy based on support vector machine (SVM). *Spectrochimica Acta. Part A: Molecular and Biomolecular Spectroscopy*, 66(3), 568-574. http://dx.doi.org/10.1016/j.saa.2006.03.038. PMid:16859975.

Dekant, W., Fujii, K., Shibata, E., Morita, O., & Shimotoyodome, A. (2017). Safety assessment of green tea based beverages and dried green tea extracts as nutritional supplements. *Toxicology Letters*, 277, 104-108. http://dx.doi.org/10.1016/j.toxlet.2017.06.008. PMid:28655517.

Ejeahalaka, K. K., & On, S. L. W. (2020). Effective detection and quantification of chemical adulterants in model fat-filled milk powders using NIRS and hierarchical modelling strategies. *Food Chemistry*, 309, 125785. http://dx.doi.org/10.1016/j.foodchem.2019.125785. PMid:31732247.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. San Francisco: Academic Press.

Gu, H. W., Yin, X. L., Ma, Y. X., Wang, J., Yang, F., Sun, W. Q., Ding, B. M., Chen, Y., & Liu, Z. (2020). Differentiating grades of Xihu Longjing teas according to the contents of ten major components based on HPLC-DAD in combination with chemometrics. *Lebensmittel-Wissenschaft + Technologie*, 130, 109688. http://dx.doi.org/10.1016/j.lwt.2020.109688.

He, F., Wu, X. H., Wu, B., Zeng, S., & Zhu, X. (2022). Green tea grades identification via Fourier transform near-infrared spectroscopy and weighted global fuzzy uncorrelated discriminant transform. *Journal of Food Process Engineering*, e14109. Ahead of print. http://dx.doi.org/10.1111/jfpe.14109.

Howland, P., & Park, P. (2004). Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8), 995-1006. http://dx.doi.org/10.1109/TPAMI.2004.46. PMid:15641730.

Ju, J., Lu, G., Lambert, J. D., & Yang, C. S. (2007). Inhibition of carcinogenesis by tea constituents. *Seminars in Cancer Biology*, 17(5), 395-402. http://dx.doi.org/10.1016/j.semcancer.2007.06.013. PMid:17686632.

Kahriman, F., & Egesel, C. O. (2016). Comparison of spectral and molecular analyses for classification of long term stored wheat samples. *Guang Pu Xue Yu Guang Pu Fen Xi*, 36(4), 1266-1272. PMid:30052360.

Laurie, S. M., Naidoo, S. I. M., Magwaza, L., Shimelis, H., & Laing, M. (2020). Assessment of the genetic diversity of sweetpotato germplasm collections for protein content. *South African Journal of Botany*, 132, 132-139. http://dx.doi.org/10.1016/j.sajb.2020.03.041.

Lee, B. L., & Ong, C. N. (2000). Comparative analysis of tea catechins and theaflavins by highperformance liquid chromatography and capillary electrophoresis. *Journal of Chromatography A*, 881(1-2), 439-447. http://dx.doi.org/10.1016/S0021-9673(00)00215-6. PMid:10905726.

Li, X. L., & He, Y. (2008). Discriminating varieties of tea plant based on Vis/NIR spectral characteristics and using artificial neural networks. *Biosystems Engineering*, 99(3), 313-321. http://dx.doi.org/10.1016/j.biosystemseng.2007.11.007.

Lin, H. C., Lee, C. T., Yen, Y. Y., Chu, C. L., Hsieh, Y. P., Yang, C. S., & Lan, S. J. (2019). Systematic review and meta-analysis of anti-hyperglycaemic effects of Pu-erh tea. *International Journal of Food Science & Technology*, 54(2), 516-525. http://dx.doi.org/10.1111/ijfs.13966.

Liu, H., Xiang, B.-R., Qu, L.-B., & Xu, J.-P. (2007). Structure analysis of benzoic medicines by near infrared and two dimensional correlation spectroscopy. *Guang Pu Xue Yu Guang Pu Fen Xi*, 27(2), 265-269. PMid:17514952.

Luypaert, J., Zhang, M. H., & Massart, D. L. (2003). Feasibility study for the use of near infrared spectroscopy in the qualitative and quantitative analysis of green tea, Camellia sinensis (L.). *Analytica Chimica Acta*, 478(2), 303-312. http://dx.doi.org/10.1016/S0003-2670(02)01509-X.

Niedźwiecki, M. J., Ciołek, M., Gańcza, A., & Kaczmarek, P. (2021). Application of regularized Savitzky-Golay filters to identification of time-varying systems. *Automatica*, 133, 109865. http://dx.doi.org/10.1016/j.automatica.2021.109865.

Pan, T. H., Yan, R., & Chen, Q. (2022). Geographical origin of green tea identification using LASSO and ANOVA. *Food Science and Technology*, 42, e41922. http://dx.doi.org/10.1590/fst.41922.

Pang, X. M., Chen, F. Y., Liu, G. Y., Zhang, Q., Ye, J. H., Lei, W. X., Jia, X. L., & He, H. B. (2022). Comparative analysis on the quality of Wuyi Rougui (Camellia sinensis) tea with different grades. *Food Science and Technology*, 42, e115321. http://dx.doi.org/10.1590/fst.115321.

Qin, Z. H., Pang, X. L., Chen, D., Cheng, H., Hu, X. S., & Wu, J. H. (2013). Evaluation of Chinese tea by the electronic nose and gas chromatography-mass spectrometry: correlation with sensory properties and classification according to grade level. *Food Research International*, 53(2), 864-874. http://dx.doi.org/10.1016/j.foodres.2013.02.005.

Song, X. C., Canellas, E., Wrona, M., Becerril, R., & Nerin, C. (2020). Comparison of two antioxidant packaging based on rosemary oleoresin and green tea extract coated on polyethylene terephthalate for extending the shelf life of minced pork meat. *Food Packaging and Shelf Life*, 26, 100588. http://dx.doi.org/10.1016/j.fpsl.2020.100588.

Tao, L. L., Huang, W., Yang, X. J., Cao, Z. Y., Deng, J. M., Wang, S. S., Mei, F. Y., Zhang, M. W., & Zhang, X. (2016). Correlations between near infrared spectra and molecular structures of 20 standard amino acids. *Guang Pu Xue Yu Guang Pu Fen Xi*, 36(9), 2766-2773. PMid:30084592.

Teixeira, J. L. D., Carames, E. T. D., Baptista, D. P., Gigante, M. L., & Pallone, J. A. L. (2020). Vibrational spectroscopy and chemometrics tools for authenticity and improvement the safety control in goat milk. *Food Control*, 112, 107105. http://dx.doi.org/10.1016/j.foodcont.2020.107105.

Uğur, H., Çatak, J., Özgür, B., Efe, E., Görünmek, M., Belli, İ., & Yaman, M. (2022). Effects of different polyphenol-rich herbal teas on reducing predicted glycemic index. *Food Science and Technology*, 42, e03022. http://dx.doi.org/10.1590/fst.03022.

Wang, L., Sun, D. W., Pu, H. B., & Cheng, J. H. (2017). Quality analysis, classification, and authentication of liquid foods by near-infrared spectroscopy: a review of recent research developments. *Critical Reviews in Food Science and Nutrition*, 57(7), 1524-1538. http://dx.doi.org/10.1080/10408398.2015.1115954. PMid:26745605.

Wu, M. M., Sun, J., Lu, B., Ge, X., Zhou, X., & Zou, M. L. (2019). Application of deep brief network in transmission spectroscopy detection of pesticide residues in lettuce leaves. *Journal of Food Process Engineering*, 42(3), e13005. http://dx.doi.org/10.1111/jfpe.13005.

Xia, X., Lin, Z., Shao, K., Wang, X., Xu, J., Zhai, H., Wang, H., Xu, W., & Zhao, Y. (2021). Combination of white tea and peppermint demonstrated synergistic antibacterial and anti-inflammatory activities. *Journal of the Science of Food and Agriculture*, 101(6), 2500-2510. http://dx.doi.org/10.1002/jsfa.10876. PMid:33058206.

Xin, Z., Jun, S., Bing, L., Xiaohong, W., Chunxia, D., & Ning, Y. (2018). Study on pesticide residues classification of lettuce leaves based on polarization spectroscopy. *Journal of Food Process Engineering*, 41(8), e12903. http://dx.doi.org/10.1111/jfpe.12903.

Xu, M., Wang, J., & Zhu, L. Y. (2021). Tea quality evaluation by applying E-nose combined with chemometrics methods. *Journal of Food Science and Technology*, 58(4), 1549-1561. http://dx.doi.org/10.1007/s13197-020-04667-0. PMid:33746282.

Xu, P. L. (2018). Research and application of near-infrared spectroscopy in rapid detection of water pollution. *Desalination and Water Treatment*, 122, 1-4. http://dx.doi.org/10.5004/dwt.2018.22559.

Xu, Q. Y., Wu, X. H., Wu, B., & Zhou, H. X. (2022). Detection of apple varieties by near-infrared reflectance spectroscopy coupled with SPSO-PFCM. *Journal of Food Process Engineering*, 45(4), e13993. http://dx.doi.org/10.1111/jfpe.13993.

Ye, J. P., & Li, Q. (2004). LDA/QR: an efficient and effective dimension reduction algorithm and its theoretical foundation. *Pattern Recognition*, 37(4), 851-854. http://dx.doi.org/10.1016/j.patcog.2003.08.006.

Yoo, S. H., Lee, Y. E., Chung, J. O., Rha, C. S., Hong, Y. D., Park, M. Y., & Shim, S. M. (2020). Enhancing the effect of catechins with green tea flavonol and polysaccharides on preventing lipid absorption and accumulation. *Lebensmittel-Wissenschaft + Technologie*, 134, 110032. http://dx.doi.org/10.1016/j.lwt.2020.110032.

Yuan, B., & Dou, X. M. (2004). Near-infrared spectral studies of hydrogen-bond in water-methanol mixtures. *Guang Pu Xue Yu Guang Pu Fen Xi*, 24(11), 1319-1322. PMid:15762465.

Zhao, J., Chen, Q., Huang, X., & Fang, C. H. (2006). Qualitative identification of tea categories by near infrared spectroscopy and support vector machine. *Journal of Pharmaceutical and Biomedical Analysis*, 41(4), 1198-1204. http://dx.doi.org/10.1016/j.jpba.2006.02.053. PMid:16621404.

Zhong, X. P., & Ban, H. (2022). Pre-trained network-based transfer learning: a small-sample machine learning approach to nuclear power plant classification problem. *Annals of Nuclear Energy*, 175, 109201. http://dx.doi.org/10.1016/j.anucene.2022.109201.

Zhu, J. J., Agyekum, A. A., Kutsanedzie, F. Y. H., Li, H. H., Chen, Q. S., Ouyang, Q., & Jiang, H. (2018). Qualitative and quantitative analysis of chlorpyrifos residues in tea by surface-enhanced Raman spectroscopy (SERS) combined with chemometric models. *Lebensmittel-Wissenschaft + Technologie*, 97, 760-769. http://dx.doi.org/10.1016/j.lwt.2018.07.055.

Zhu, M., Li, N., Zhao, M., Yu, W., & Wu, J.-L. (2017). Metabolomic profiling delineate taste qualities of tea leaf pubescence. *Food Research International*, 94, 36-44. http://dx.doi.org/10.1016/j.foodres.2017.01.026. PMid:28290365.