Identification of lambda-cyhalothrin residues on Chinese cabbage using fuzzy uncorrelated discriminant vector analysis and MIR spectroscopy

Xiaohong Wu[1,2*], Tingfei Zhang[1], Bin Wu[3], Haoxiang Zhou[4]

(1. *School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, Jiangsu, China*;
2. *High-tech Key Laboratory of Agricultural Equipment and Intelligence of Jiangsu Province, Jiangsu University, Zhenjiang 212013, Jiangsu, China*;
3. *Department of Information Engineering, Chuzhou Polytechnic, Chuzhou, 239000, Anhui, China*;
4. *Research Institute of Zhejiang University-Taizhou, Taizhou 317700, Zhejiang, China*)

Abstract: Excessive pesticide residues on Chinese cabbage will be harmful to people's health. Therefore, an identification system was designed for qualitative analysis of lambda-cyhalothrin residues on Chinese cabbage leaves. In order to extract discriminant information from mid-infrared (MIR) spectra of Chinese cabbage effectively, fuzzy uncorrelated discriminant vector (FUDV) analysis was proposed by introducing the fuzzy set theory into uncorrelated discriminant vector (UDV) analysis. In this system, the Cary 630 FTIR spectrometer was used to scan four samples of Chinese cabbage with different concentrations of lambda-cyhalothrin. The MIR spectra were preprocessed by standard normal variable (SNV) and Savitzky-Golay smoothing (SG). Next, the high-dimensional MIR spectra were processed for dimension reduction by principal component analysis (PCA). Furthermore, UDV, FUDV, and some other discriminant analysis algorithms were used for feature extraction, respectively. Finally, the *K*-nearest neighbor (KNN) classifier was employed to classify the data. The experimental results showed that when FUDV was used as the feature extraction algorithm, the identification system reached the maximum classification accuracy of 100%. The results indicated that FUDV combined with MIR spectroscopy was an effective method to identify lambda-cyhalothrin residues on Chinese cabbage.

Keywords: Chinese cabbage, mid-infrared spectroscopy, fuzzy uncorrelated discriminant vector, uncorrelated discriminant vector, lambda-cyhalothrin residues

DOI: 10.25165/j.ijabe.20221503.6486

Citation: Wu X H, Zhang T F, Wu B, Zhou H X. Identification of lambda-cyhalothrin residues on Chinese cabbage using fuzzy uncorrelated discriminant vector analysis and MIR spectroscopy. Int J Agric & Biol Eng, 2022; 15(3): 217–224.

1 Introduction

As a green and healthy vegetable, Chinese cabbage is very popular among consumers in China. Due to the widespread use of pesticides in agricultural production, there are more and more cases of food poisoning caused by pesticide residues[1] in recent years. In the cultivation of Chinese cabbage, one of the most commonly used insecticides is lambda-cyhalothrin, which is used to control various pests. As a pyrethroid insecticide, lambda-cyhalothrin is very harmful to the human body. Research results showed that regular consumption of foods containing pyrethroid pesticide residues increases the risk of developmental and neurological diseases in children aged 3-11[2]. Pyrethroids can disrupt reproductive hormones, thereby destroying the male reproductive system[2]. Therefore, it is necessary to design an efficient, fast and accurate identification system to analyze lambda-cyhalothrin residues for the health of consumers[3-5].

Several detection methods, such as quantum dots (QDs)[6], laser light[7], and spectroscopy, have been widely used in the detection of pesticide residues. Spectroscopy has the advantages of being fast and non-destructive and does not need complicated sample pretreatment, so it is widely used in agricultural production and food detection[8]. Spectroscopy technology commonly includes near-infrared (NIR) spectroscopy[9,10], mid-infrared (MIR) spectroscopy[11] and hyperspectral imaging (HSI) technology[12,13]. The wavelength ranges of NIR and MIR are 780-2526 nm and 2526-25 000 nm, respectively. Spectral information of MIR arises from fundamental molecular vibrations in functional groups, and NIR spectra contain spectral information on overtones and combinations of fundamental vibrations. In comparison, MIR is more sensitive than NIR in detecting complex molecular structures. HSI is a combination of imaging technology and spectral detection technology. HSI can obtain the external image features of the sample while collecting spectral information. However, the detection process of HSI is so complicated that it is not conducive to large-scale use. In order to study a fast and accurate pesticide residues identification system, this study used MIR spectroscopy to collect spectral information. Some research results showed that MIR can effectively obtain the characteristic information of agricultural products and food. For example, Yang et al.[14] studied a model composed of MIR and partial least squares regression (PLSR) to detect pesticide residues in Chinese herbal

Received date: 2021-01-27 **Accepted date:** 2021-09-07

Biographies: Tingfei Zhang, MS candidate, research interest: NIR detection of food and agricultural products and pattern recognition. Email: 2221907104@stmail.ujs.edu.cn; **Bin Wu**, MS, Associate Professor, research interest: machine learning and NIR detection, Email: wubin2003@163.com; **Haoxiang Zhou**, MS, research interest: fuzzy clustering and data analysis, Email: 2221807065@stmail.ujs.edu.cn.

***Corresponding author: Xiaohong Wu**, PhD, Professor, research interest: nondestructive detection of food and agricultural products and pattern recognition. School of Electrical and Information Engineering, Jiangsu University, Xuefu Road No.301, Zhenjiang 212013, Jiangsu, China. Tel: +86 51188791245, Email: wxh419@ujs.edu.cn.

medicines. Etzion et al.[15] took raw milk as the research object, and utilized MIR spectroscopy to detect protein concentration in milk. Yang et al.[16] used MIR spectroscopy to determine the nitrate content in Chinese cabbage. With the development of instrument and software technology, some emerging technologies, such as attenuated total reflectance (ATR) technology, are widely used in the field of agricultural products and food detection. ATR has the advantages of simple operation and high detection sensitivity. Su et al.[17] used ATR-MIR spectroscopy to identify potato varieties and detected potato doneness degree. The above researches proved that the models established using MIR were effective in identifying the variety and quality of agricultural and food products. These researches provided the theoretical reference for this study. Based on the above researches, the purpose of this study was to explore the potential of ATR-MIR for detecting lambda-cyhalothrin residues on Chinese cabbage leaves.

Fisher's linear discriminant analysis (LDA)[18] is an important tool in statistical pattern recognition. LDA aims to find a discriminant vector set that makes training data the biggest ratio of between-class distance to within-class distance. At the same time, the discriminant vectors are mutually orthogonal. But the projections of the training samples on the feature space of the discriminant vector set are statistically correlated[19]. In order to solve this problem, uncorrelated discriminant vector (UDV) analysis[20] added uncorrelated constraints when computing the discriminant vector set. So the discriminant vector set obtained by UDV is very effective when it is applied in classification.

MIR spectral data sets more and less contain the overlapped data points that traditional classification methods are difficult to distinguish. So this article introduces fuzzy theory on the basis of UDV to solve this problem. Fuzzy set theory was established by Zadeh[21] is widely used in many fields such as pattern recognition, image processing, and data mining. This theory was designed to deal with poorly defined concepts. It allows the fuzzy membership of samples to change between one (complete belonging) and zero (complete exclusion), instead of taking a value of one and zero as in ordinary set theory. A number of examples can prove fuzzy classification algorithms are always better than traditional methods when dealing with some problems. For example, Wu et al.[22] studied the model of fuzzy Foley-Sammom transformation (FFST) to classify Chinese vinegar varieties, and researched fuzzy discriminant principal component analysis (FDPCA) to deal with overlapped data points[23]. Chen et al.[24] studied the method of fuzzy linear discriminant analysis (FLDA) to deal with overlapped data points. Lin et al.[25] created fuzzy support vector machines (FSVM) to reduce noises and outliers. Ning et al.[26] studied the application of fuzzy C-means clustering in image analysis of critical medicine. Cadenas et al.[27] used fuzzy *K*-nearest neighbor classifier to deal with imperfect data. In order to qualitatively analysis of pesticide residues on Chinese cabbage effectively and quickly, this study proposed fuzzy uncorrelated discriminant vector (FUDV) analysis to extract discriminant information from the collected spectral data of Chinese cabbage. FUDV was designed by introducing fuzzy set theory into UDV.

In this study, the identification system of pesticide residue levels on Chinese cabbage contains two parts: spectral data collection and machine-learning algorithms. The part of machine-learning algorithms consists of five algorithms: standard normal variable (SNV), Savitzky-Golay smoothing (SG), principal component analysis (PCA), FUDV, and *K*-nearest neighbor (KNN) algorithm. SNV and SG are commonly used for preprocessing spectral data. PCA[28] and FUDV are feature extraction algorithms, and they are applied for reducing the data dimensionality and extracting features. KNN is a common classifier and is utilized to acquire the classification accuracy of this model.

2 Materials and methods

2.1 Uncorrelated discriminant vector analysis

The objective of UDV is to find a transformation matrix that consists of uncorrelated discriminant vectors linearly transforming the data into the uncorrelated feature space, and UDV tends to find a linear transformation matrix that maximizes the ratio of the between-classes scatter S_b to the within-classes scatter matrix S_w in the uncorrelated space.

Suppose the following data set of training samples is given as $X=\{x_1, x_2, x_3, \ldots, x_n\}$. The sample x_i belongs to one of c classes ($\omega_1, \omega_2, \omega_3, \ldots, \omega_c$). Then some definitions and equations of UDV are given as follows[20]:

$$a_i = \frac{\sum_{k=1}^{N_i} x_k}{N_i} \tag{1}$$

where, a_i is the mean of samples in class ω_i, $1 \le i \le c$; N_i is the number of samples belonging to class ω_i; x_k is the kth sample of class ω_i, $1 \le k \le N_i$.

$$S_b = \sum_{i=1}^{c} N_i (a_i - a)(a_i - a)^{\mathrm{T}} \tag{2}$$

where, S_b is the between-class scatter matrix; a is the mean of total samples.

$$S_w = \sum_{i=1}^{c} \sum_{k=1}^{N_i} (a_i - x_k)(a_i - x_k)^{\mathrm{T}} \tag{3}$$

where, S_w is the within-class scatter matrix.

$$S_t = \sum_{i=1}^{N} (x_i - a)(x_i - a)^{\mathrm{T}} = S_b + S_w \tag{4}$$

where, S_t is the total scatter matrix of the whole data set; N is the number of all samples. To obtain the find discriminant vector set φ, UDV defines the discriminant criterion function $J_F(\varphi)$ as[20]:

$$J_F(\varphi) = \frac{\varphi^{\mathrm{T}} S_b^u \varphi}{\varphi^{\mathrm{T}} S_t^u \varphi} \tag{5}$$

where, S_b^u is the between-class scatter matrix S_b in uncorrelated space and $S_b^u = V^{\mathrm{T}} S_b V$. S_t^u is the total scatter matrix S_t in uncorrelated space and $S_t^u = V^{\mathrm{T}} S_t V$. Because of $\varphi^{\mathrm{T}} S_t \varphi = 1$, Equation (5) can be simplified as follows[20]:

$$J_F(\varphi) = \varphi^{\mathrm{T}} S_b^u \varphi \tag{6}$$

In the uncorrelated space, and the first uncorrelated discriminant vector φ_1 can be acquired by the following equation:

$$\varphi_1 = \arg\max(J_F(\varphi)) \tag{7}$$

The rth Fisher discriminant vectors φ_r can be computed by the following equation:

$$\varphi_r = \underset{\varphi^{\mathrm{T}} \cdot \varphi_i = 0, i=1,2,\ldots,r-1}{\arg\max} (J_F(\varphi)) \tag{8}$$

The discriminant vector set φ is based on Fisher discriminant criterion. However, to make φ satisfy the uncorrelated constraint, φ is transformed by the uncorrelated transforming matrix V. Let $M=V\varphi$, and M is the uncorrelated discriminant vector set.

UDV algorithm is described in the following steps:

Step 1 Compute S_b, S_w, and S_t using Equations (2)-(4), respectively;

Step 2 Compute Λ and U of S_t, $U^{\mathrm{T}}S_tU=\Lambda$, and then $V=U\Lambda^{-1/2}$;

Step 3 Compute the rth ($r=c-1$) discriminant vector φ_r using Equations (7) and (8);

Step 4 Let $M=V\varphi$.

Although UDV can effectively classify the data points, UDV cannot get a satisfactory result when there are a few overlapped data points in the data set. For a simple reason, each sample in the same class has the same weight when calculating the scatter matrixes and mean value. In order to solve this problem, fuzzy set theory is introduced into UDV to produce FUDV.

2.2 Fuzzy uncorrelated discriminant vector analysis

In UDV, all data points are considered to have the same weight. When there are overlapped data points in the data set, these data will greatly affect classification accuracies. Therefore, FUDV introduces the fuzzy membership values to determine the extent that which the data belong to one class, and then different data points have different contributions. When calculating the scatter matrixes and mean value, the contribution of overlapped data points is much smaller than normal data points. Such a mechanism can effectively reduce the influence of overlapped data points on classification accuracy. A detailed description of FUDV is given in the following.

Suppose the following data set of training samples is given as $X=\{x_1, x_2, x_3, \ldots, x_n\}$. The sample x_i belongs to one of c classes ($\omega_1, \omega_2, \omega_3, \ldots, \omega_c$). The equation of fuzzy membership value μ_{ij} can be written as

$$\mu_{ij}=\left[\sum_{k=1}^{c}\left(\frac{\|x_j-a_i\|^2}{\|x_j-a_k\|^2}\right)\right]^{-1} \tag{9}$$

Fuzzy membership value μ_{ij} changes between zero and one, and it indicates the extent that which the jth sample belongs to class ω_i.

After μ_{ij} is calculated, the fuzzy between-class scatter matrix S_{fB}, fuzzy within-class scatter matrix S_{fW} and fuzzy total scatter matrix S_{fT} are described as follows:

$$S_{\mathrm{fB}}=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}(a_i-a)(a_i-a)^{\mathrm{T}} \tag{10}$$

$$S_{\mathrm{fW}}=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}(a_i-x_j)(a_i-x_j)^{\mathrm{T}} \tag{11}$$

$$S_{\mathrm{fT}}=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}(x_j-a)(x_j-a)^{\mathrm{T}} \tag{12}$$

where, m is the weight index of fuzzy membership value μ_{ij}.

Like FLDA, FUDV tries to find the discriminant vector ψ by solving the following equation[29]:

$$J_F(\psi)=\frac{\psi^{\mathrm{T}}S_{\mathrm{fB}}\psi}{\psi^{\mathrm{T}}S_{\mathrm{fT}}\psi} \tag{13}$$

Based on the theory of uncorrelated discriminant transform[19] and UDV[20], the projections of the training samples on the discriminant vectors should be statistically uncorrelated, and for any vector γ in uncorrelated space, it satisfies $\gamma^{\mathrm{T}}C_{\mathrm{fT}}\gamma=1$ where $C_{\mathrm{fT}}=P^{\mathrm{T}}S_{\mathrm{fT}}P=I$ and I is the identity matrix. S_{fT} is a real symmetric matrix, so there is a real orthogonal matrix U satisfying $U^{\mathrm{T}}S_{\mathrm{fT}}U=\Lambda$, where Λ is the diagonal matrix. Because of $U^{\mathrm{T}}S_{\mathrm{fT}}U=\Lambda$ and $P^{\mathrm{T}}S_{\mathrm{fT}}P=I$, there is $P=U\Lambda^{-1/2}$. Let $\psi=P\gamma$, Equation (13) can be written as follows:

$$\frac{\psi^{\mathrm{T}}S_{fB}\psi}{\psi^{\mathrm{T}}S_{fT}\psi}=\frac{\gamma^{\mathrm{T}}P^{\mathrm{T}}S_{fB}P\gamma}{\gamma^{\mathrm{T}}P^{\mathrm{T}}S_{fT}P\gamma}=\gamma^{\mathrm{T}}P^{\mathrm{T}}S_{fB}P\gamma=\gamma^{\mathrm{T}}C_{fB}\gamma \tag{14}$$

where, $C_{\mathrm{fB}}=P^{\mathrm{T}}S_{\mathrm{fB}}P$, and Equation (13) can be simplified as $J_F(\gamma)=\gamma^{\mathrm{T}}C_{\mathrm{fB}}\gamma$. γ has r ($r=c-1$) uncorrelated discriminant vectors and $\gamma=[\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_r]$. The rth uncorrelated discriminant vector γ_r can be acquired by utilizing the following equation.

$$\gamma_r=\underset{\omega^{\mathrm{T}}\cdot\omega_i=0,i=1,2,\ldots,r-1}{\arg\max}\ (J_F(\gamma)) \tag{15}$$

Let $Q=P\gamma$ and Q be the fuzzy uncorrelated discriminant matrix.

FUDV algorithm can be described in the following steps:

Step 1 Compute μ_{ij} using Equation (9);

Step 2 Compute S_{fB}, S_{fW}, and S_{fT} using Equations (10), (11), and (12), respectively;

Step 3 Compute the diagonal matrix Λ and eigenvector matrix U of S_{fT}, $U^{\mathrm{T}}S_{\mathrm{fT}}U=\Lambda$, and then $P=U\Lambda^{-1/2}$;

Step 4 Compute the rth discriminant vector γ_r using Equation (15);

Step 5 Let $Q=P\gamma$.

2.3 Experiment materials

Fresh Chinese cabbages (*Brassica rapa*, Chinese group) were selected as the experimental samples. A total of 160 pieces of Chinese cabbage leaf samples were collected that had similar growth conditions. The Chinese cabbage leaves were washed adequately using water (45°C) to clean dust and stored in sealed bags.

The selected pesticide was lambda-cyhalothrin (5% EC, Shandong Shenda Crop Science Co., Ltd., Shouguang, China) which is widely used to kill cabbage bugs during the growth of cabbage. The concentration ratio of 1:500-1:600 is the recommended concentration of lambda-cyhalothrin pesticide by relevant pesticide manufacturers.

In order to produce experimental samples of Chinese cabbage leaves with different pesticide residue concentrations, 160 cabbage leaves were randomly divided into 4 groups with 40 leaves in each group. The leaves of Group A were sprayed with water as the control group. The leaves of Group B were sprayed with a solution of pesticide to water ratio of 1:500, and Group B was called the mild residue group. Similarly, the leaves of Group C and Group D were sprayed with a solution of 1:100 and 1:20, respectively, called the moderate residue group and the severe residue group, respectively. All the prepared samples were placed in a cool and ventilated place for 24 h to reduce the influence of water. Before MIR spectral collection, each sample of Chinese cabbage leaf was made into a 2 mm×2 mm small sample.

2.4 MIR detection and data analysis software

The MIR spectral data of Chinese cabbage were acquired using the Cary 630 FTIR spectrometer (Agilent, USA), ranging from 590-4289 cm^{-1} with the resolution of 8 cm^{-1} and 64 scans for the background and samples, which resulted in the 971-dimensional spectra. Two kinds of spectral analysis software, Micro lab PC and Resolutions pro, were used to record the MIR spectra. All algorithms and spectral data were processed by Matlab 2016a (Mathworks Co., Natick, MA, USA) based on Windows 10 system in this study.

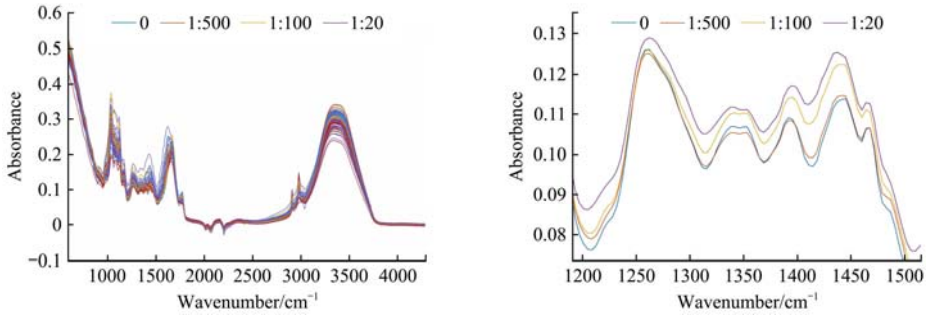
2.5 MIR spectra collection

The MIR spectra of all samples were collected in a constant temperature and humidity laboratory (temperature was about 25°C and humidity was about 50%). The steps of collecting spectral data are described as follows: At first, the lens of the instrument was cleaned with anhydrous alcohol to reduce errors caused by the device before collecting the spectral data of samples. Secondly, the spectrometer was used to detect the background spectrum for decreasing deviation caused by environmental factors. Finally,

the samples were placed on the spectrometer to acquire the spectral data, and each sample was detected three times, and the average value of three experimental data was the final spectral datum for subsequent experiments.

2.6 MIR spectral analysis

The MIR spectral wavenumber range of Chinese cabbage leaf was 590-4289 cm^{-1}. The original MIR spectral curves of four different pesticide residue levels are shown in Figure 1a. Since fresh Chinese cabbage leaf contains more than 90% moisture, the spectral data will be greatly affected by moisture. The two main absorption peaks in the 3000-3800 cm^{-1} and 1500-1800 cm^{-1} regions are the specific absorption of moisture[30]. In the wavenumber range of 1200-1500 cm^{-1}, moisture has little effect on the spectral data. In this range, the spectral data are related to some chemical bonds and functional groups. For example, chemical bonds such as C−O, P−O, and C−O stretching vibrations range from 1100 cm^{-1} to 1200 cm^{-1}. The region of 1200-1500 cm^{-1} mainly contains the C−H, N−H distortion vibrations and the N−O, N=O stretching vibrations. In addition, the lambda-cyhalothrin pesticide contains a unique group C−F bond, and the absorption range of the bond is between 1000-1400 cm^{-1}. Because Chinese cabbage leaves with four different levels of pesticide residues have different functional group information, MIR spectra are able to accurately express all samples. In order to show these differences more clearly, the average MIR spectral curves between 1200-1500 cm^{-1} were enlarged as shown in Figure 1b. It could be seen from Figure 1b that there were certain differences in the average MIR spectral curves of four different pesticide residue levels. These differences indicated that MIR spectroscopy had the potential to identify different levels of pesticide residue.



a. Original spectra of each group of Chinese cabbage leaves b. Partially enlarged area of the average spectra

Figure 1 Original spectra of each group of Chinese cabbage leaves and partially enlarged area of the average spectra

3 Results and discussion

3.1 Data preprocessing

When the spectrometer was used to collect MIR spectra, it would inevitably be interfered with by noise. The original spectra of each group of Chinese cabbage leaves were plotted in Figure 1a. As shown in Figure 1a, the original spectra of different pesticide residue concentrations had obvious overlapped parts. There are two reasons for this phenomenon. First, it was caused by the samples themselves. Second, the collected spectral data contained a certain deviation due to the influence of noise interference and instrument detection. In order to eliminate the influence of these factors on original spectral data, it was necessary to preprocess the data.

Six different algorithms, multiplicative scattering correction (MSC), SNV, SG, mean centering (MC), first derivative (FD), and second derivative (SD), were used to preprocess the spectral data in this experiment. These preprocessing algorithms were divided into four categories. MC can amplify weak signals. FD and SD can reduce the influence of instrument errors in the spectral data. SG can eliminate noise in the spectral data. MSC and SNV can eliminate the scattering influence in the spectral data. Table 1 lists the impact of different preprocessing algorithms and the combined application of preprocessing algorithms on classification accuracy. It could be seen in Table 1 that the classification accuracy of FUDV reached 100% when the spectral data were processed by SG. And the accuracies of UDV and FLDA were significantly improved when SNV and SG were used in combination. Finally, SNV and SG were selected as preprocessing algorithms in this identification system.

Table 1 Classification accuracies under different preprocessing algorithms

Feature extraction	Classification accuracy									
	MSC	SNV	SG	MC	FD	SD	SNV-SG	FD-SG	SNV-FD	MSC-SG
UDV	92.5%	87.5%	92.5%	90.0%	92.5%	90.0%	97.5%	92.5%	90.0%	95.0%
FLDA	87.5%	75.0%	87.5%	90.0%	92.5%	82.5%	95.0%	92.5%	92.5%	85.0%
FUDV	97.5%	92.5%	100%	92.5%	92.5%	92.5%	100%	95.0%	95.0%	100%

Note: UDV: Uncorrelated discriminant vector; FLDA: Fuzzy linear discriminant analysis; FUDV: Fuzzy uncorrelated discriminant vector; MSC: Multiplicative scattering correction; SNV: Standard normal variable; SG: Savitzky-Golay smoothing; MC: Mean centering; FD: First derivative; SD: Second derivative. The same as below.

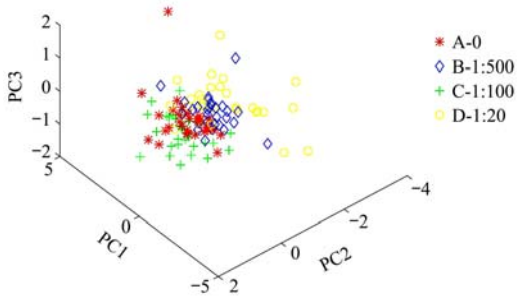
3.2 Dimension reduction

The dimensionality of the spectral data was 971 dimensions and the wavenumbers ranged from 590 cm^{-1} to 4289 cm^{-1}. If feature extraction algorithms, such as LDA, were directly used to deal with the original data, they would encounter computational difficulty. On the one hand, it is computationally challenging by using huge data matrices to calculate eigenvalues. On the other hand, those huge matrices will always encounter a small sample size problem[30] when the dimensionality of samples vastly exceeds the number of samples. The scatter matrix is close to the singular matrix if the spectra with 971 dimensions are directly used for feature extraction because of the small sample size problem. To solve this problem and improve system efficiency and accuracy, PCA was utilized for spectral dimensionality reduction. PCA method can map multiple features into some comprehensive features by finding a set of orthogonal features that can express all

the original features as much as possible, thereby achieving the purpose of dimensionality reduction.

To find the most important features which can beat represent the original spectral data, the concept of accumulative contribution rate was used to decide the number of principal components. The accumulative contribution rate illustrates the proportion of the first L principal components in total data (accumulative contribution rate is equal to the sum of eigenvalues of the first L principal components divided by the sum of the total eigenvalues). The accumulative contribution rate achieved 99.16% when the number of principal components is 14. But in actual application, it was found that the accuracy of classification was not ideal. In order to reduce the data redundancy while keeping the maximum amount of information as much as possible, the number of principal components was set to 24, and accordingly, the accumulative contribution rate was 99.61%.

To visualize the spectral data information processed by PCA, the three-dimensional scatter points processed by PCA are shown in Figure 2. In Figure 2, there are a lot of overlapped data points in the data set, and such distribution would cause great difficulty in classification. Therefore, feature extraction algorithms were used to solve this problem.



Note: PC: Principal component; A: water; B: 1:500 pesticide solution; C: 1:100 pesticide solution; D: 1:20 pesticide solution.

Figure 2 Three-dimensional distribution of data by PCA

3.3 Feature extraction

In this section, the spectral data would be divided into training sets and test sets according to the ratio of 3:1. The training set had 120 samples and the test set had 40 samples. FUDV, UDV, FLDA, uncorrelated discriminant transform (UDT)[19], foley-sammon transform (FST)[31], and FFST[22] were used to extract features from the spectral data of Chinese cabbage, respectively. Finally, KNN classifier was used to compute the classification accuracies.

3.3.1 Classification with UDV

The objective of feature extraction is to find a discriminant vector set. Here, UDV was used as a feature extraction algorithm to obtain the discriminant vector set that included three uncorrelated discriminant vectors. These three uncorrelated discriminant vectors formed a hyperplane. After the spectral data of Chinese cabbage were projected onto the hyperplane, it was more convenient to classify these spectral data. Figure 3 shows the three-dimensional spectral data points processed by UDV. It could be seen that there was not only a significant reduction in the extent of data confusion but also an obvious boundary between Group A and Group D. However, there were some groups, such as Group A and Group C, whose spectral data were somewhat similar. Therefore, even after feature extraction by UDV, there was still some overlapped data between these groups. The impact of these overlapped data points from the classification results is listed in Table 2, and the average classification accuracy of UDV was 93.9%.

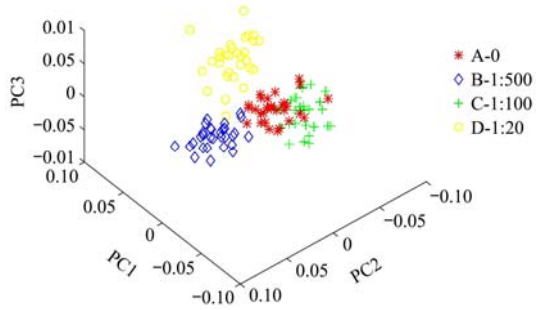


Figure 3 Three-dimensional distribution of data points by UDV

3.3.2 Classification with FUDV

To improve the classification accuracy, FUDV was proposed by introducing the fuzzy set theory into UDV. Spectral data can be classified effectively with FUDV, even if there were overlapped data points in the data set. Before FUDV was utilized to extract the features from spectral data, the fuzzy membership values u_{ik} were calculated. Fuzzy membership values are shown in Figure 4, where the ordinate represented the fuzzy membership values that changed between 0 and 1, and the abscissa represented the number of samples (each group of Chinese cabbage had 30 samples).

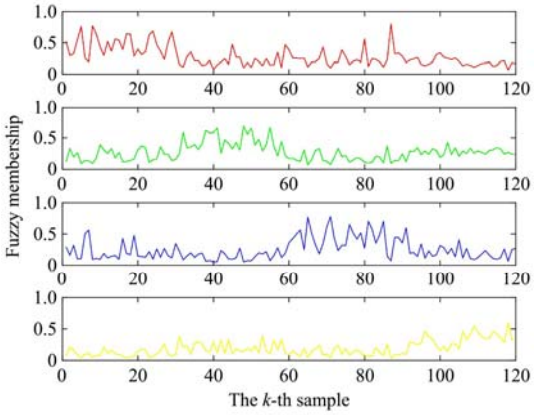


Figure 4 Fuzzy membership values of FUDV

As shown in Figure 4, the fuzzy membership values indicated the extent that a spectral data belonged to one class. For the kth sample x_k, x_k belongs to the ith class if the fuzzy membership value u_{ik} is the biggest than those values belonging to other classes. On the contrary, if the fuzzy membership value of x_k is not the biggest value, then the sample x_k does not belong to the ith class. In the four subgraphs of Figure 4, for cabbage samples belonging to one kind of Chinese cabbage group, their fuzzy membership values were greater than those of other kinds. Based on fuzzy set theory, fuzzy membership values in FUDV were beneficial for dealing with overlapped samples. But in Figure 4, the fuzzy membership values of a few samples that did not belong to one class were greater than those of other classes. This problem would have a certain influence on the classification accuracies.

When FUDV was used for extracting features, it reduced the dimensions of spectral data from 24 to 3. Figure 5 shows the distribution of spectral data reduced by FUDV when $m = 2$ on the three-dimensional view. Compared with the spectral data processed by UDV in Figure 3, Figure 5 shows that the spectral data clusters in the same group were more compactly aggregated, and there were obvious boundaries between different groups. This distribution of the spectral data points processed by FUDV

could be explained that FUDV could effectively reduce the influence of overlapped points in feature extraction. It could be found from Table 2 that after the KNN classier was utilized to compute the accuracy, the average classification accuracy of FUDV was 97.5% and the maximum classification accuracy was 100%.

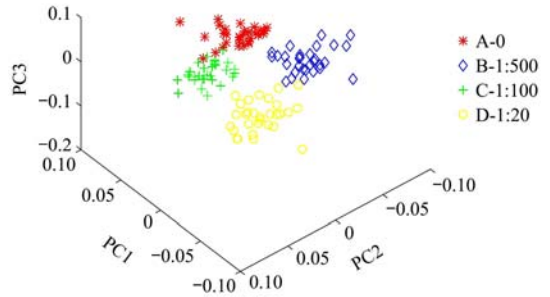


Figure 5 Three-dimensional distribution of data by FUDV

3.3.3 Classification with FLDA

FLDA was developed by introducing fuzzy set theory into LDA. The difference between FUDV and FLDA is that the data points projected by FUDV are statistically uncorrelated, but the data points projected by FLDA are statistically correlated. To illustrate the advantages of uncorrelated discriminant vectors, FLDA was performed for feature extraction as a comparison in this experiment. The process of running FLDA was similar to FUDV. Firstly, FLDA computed the fuzzy membership values of each sample. Then FLDA was run to obtain the discriminant vector set and projected the spectral data onto the feature space which consisted of discriminant vectors. Finally, the KNN algorithm was used to classify these spectral data and get the accuracy of FLDA. The spectral data points after feature extraction by FLDA are shown in Figure 6. In Figure 6, FLDA could also complete the task of feature extraction. But compared with FUDV, the spectral data clusters processed by FLDA had some confusing data points between group B and group D. By contrast, there were less confusing data points in the spectral data clusters obtained by FUDV. It could be found from Figure 5 and Figure 6 that the spectral data clusters processed by FUDV had obvious boundaries between different groups, and the same group of spectral data clusters became more compact than the spectral data clusters processed by FLDA. This result showed that the data processed by FUDV could be classified more accurately than the data by FLDA. The reason for this result was that FLDA just satisfied the biggest ratio of fuzzy between-class distance to fuzzy total-class distance and FUDV satisfied not only the condition of FLDA but also the uncorrelated constraints.

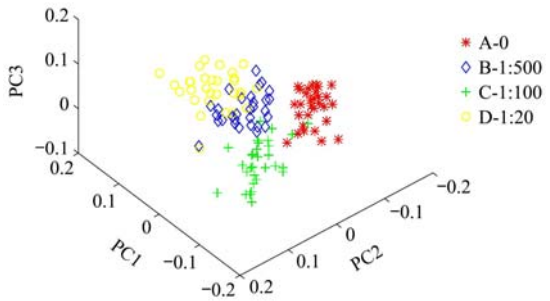


Figure 6 Three-dimensional distribution of data by FLDA

As could be seen from the accuracies of FUDV and FLDA in Table 2, the average classification accuracy computed by the KNN algorithm after feature extraction with FLDA was 88.9%, and the average classification accuracy of FUDV was 97.5%. This experiment proved the advantage of uncorrelated discriminant vectors when FUDV was used to process overlapped data points.

3.4 Different values of K with KNN

Due to the value of K having a significant impact on the performance of KNN, several values of K were selected (K=1, 3, 5, 7, 9, 11, 13, 15, 17) when KNN was utilized to classify the spectral data. In order to illustrate the advantages of FUDV, in addition to UDV, FUDV, and FLDA, some other feature extraction algorithms, including UDT[19], FST[31], and FFST[22], were also applied in this system. The classification results of all feature extraction algorithms are listed in Table 2. It could be seen from Table 2 that when the values of K were 3, 5, and 11, the highest classification accuracy of FUDV was 100%. With the change of K, the accuracy of FUDV was very stable, and the average classification accuracy of FUDV was significantly better than other feature extraction algorithms. The results showed that, compared with other feature extraction algorithms, FUDV was a more efficient feature extraction algorithm with stronger anti-interference ability.

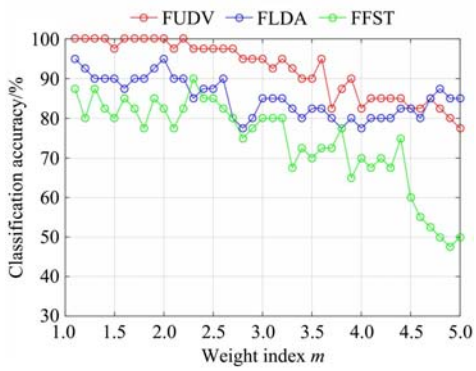
Table 2 Identification accuracies with the different values of *K*

Feature extraction	Identification accuracy									
	1/%	3/%	5/%	7/%	9/%	11/%	13/%	15/%	17/%	Average /%
UDV	95.0	92.5	97.5	95.0	95.0	95.0	92.5	87.5	95.0	93.9
FLDA	90.0	87.5	95.0	87.5	87.5	95.0	95.0	85.0	77.5	88.9
FUDV	97.5	100	100	97.5	95.0	100	95.0	97.5	95.0	97.5
UDT	92.5	87.5	77.5	82.5	87.5	80.0	70.0	62.5	77.5	79.7
FST	85.0	82.5	77.5	80.0	87.5	80.0	70.0	62.5	77.5	78.1
FFST	92.5	87.5	82.5	82.5	87.5	87.5	77.5	77.5	87.5	84.7

Note: UDT: Uncorrelated discriminant transform; FST: Foley-sammon transform; FFST: Fuzzy Foley-Sammom transformation.

3.5 Selection of optimal weight index

Weight index m is a significant parameter in a fuzzy discriminant algorithm, which has a profound influence on the extraction of fuzzy information. Bezdek[32] considered that this parameter controlled the degree of fuzzy membership sharing between classes. The value of weight index m is specified in the interval $(1, +\infty)$. As m approaches infinity, FUDV will lose the characteristics of feature extraction, and the data from different groups will also become vaguer[33]. To improve the accuracy of the pesticide residue identification system, it is necessary to select an appropriate weight index. However, there is not a general method to acquire the optimal weight index, and it depends greatly on the distribution characteristics of the data. The trend of classification accuracy using FUDV, FLDA, and FFST in different values of weight index is presented in Figure 7. With the increase in weight value, the classification accuracy of FUDV, FLDA, and FFST all showed a general downward trend. And it was noteworthy that the classification accuracy of FFST suffered a cliff-like drop when weight index m was greater than 4.5. In contrast, the classification accuracies of FUDV and FLDA were less affected. The results showed that FUDV and FLDA have strong stability. From the figure, it was found that the classification accuracies of FUDV were all above 95% when the range of weight value $m \in [1.1, 3]$. Finally, the optimal weight value m was selected as 2 and the optimal classification accuracy of FUDV was 100%. The results could provide a technical reference for the selection of the weight index of FUDV.



Note: FUDV: Fuzzy uncorrelated discriminant vector; FLDA: Fuzzy linear discriminant analysis; FFST: Fuzzy Foley-Sammom transformation.

Figure 7 Classification results of FUDV and FLDA in different weight index

4 Conclusions

In this study, a lambda-cyhalothrin residue identification system was designed to distinguish four different lambda-cyhalothrin levels of Chinese cabbage. Due to overlapped data points, the accuracy of the identification system was often not satisfactory. To improve the accuracy of this system in detecting lambda-cyhalothrin concentration, FUDV was developed by introducing fuzzy set theory into UDV for feature extraction from the MIR spectral data.

In the above experiments, the UDV, FLDA, FUDV, UDT, FST, and FFST were run for feature extraction, respectively. Through comparing the classification accuracies, it could be found that the classification accuracies of FUDV were better than other feature extraction algorithms in extracting the features from the spectra. From the experimental results, the average classification accuracies of FUDV, UDV, FLDA, UDT, FST, and FFST were 97.5%, 93.9%, 88.9%, 79.7%, 78.1%, and 84.7%, respectively. The maximum classification accuracy of FUDV reached 100%. So this result proved that FUDV had an advantage in feature extraction when there were some overlapped spectral data points in the data clusters. The experimental results indicated that this identification model combined with PCA, FUDV, and KNN algorithm was a very effective method of classifying lambda-cyhalothrin concentration of Chinese cabbage.

Acknowledgement

The authors sincerely acknowledge that this work was financially supported by the National Natural Science Foundation of China (Grant No. 31471413), the Undergraduate Scientific Research Project of Jiangsu University (Grant No. 17A274), and the University Natural Science Research Project of Anhui Province (Grant No. KJ2019A1129).

[References]

[1] Liu P J, Guo Y Z. Current situation of pesticide residues and their impact on exports in China. Journal of Agricultural Science and Technology, 2017; 19(11): 8–14. (in Chinese)

[2] Zhu Q Y, Yang Y, Zhong Y Y, Lao Z T, O'Neill P, Hong D, et al. Synthesis, insecticidal activity, resistance, photodegradation and toxicity of pyrethroids (A review). Chemosphere, 2020; 254: 126779. doi: 10.1016/j.chemosphere.202.126779.

[3] Hassan M M, Li H H, Ahmad W, Zareef M, Wang J J, Xie S C, et al. Au@Ag nanostructure based SERS substrate for simultaneous determination of pesticides residue in tea via solid phase extraction coupled multivariate calibration. LWT, 2019; 105: 290–297.

[4] Yang N, Wang P, Xue C Y, Sun J, Mao H P, Oppong P K. A portable detection method for organophosphorus and carbamates pesticide residues based on multilayer paper chip. Journal of Food Process Engineering, 2018; 41(8): e12867. doi: 10.1111/jfpe.12867.

[5] Ma P, Wang L Y, Xu L, Li J Y, Zhang X D, Chen H. Rapid quantitative determination of chlorpyrifos pesticide residues in tomatoes by surface-enhanced Raman spectroscopy. European Food Research and Technology, 2020; 246(1): 239–251.

[6] Zhou J W, Zou X, Song S H, Chen G H. Quantum dots applied to methodology on detection of pesticide and veterinary drug residues. Journal of Agricultural & Food Chemistry, 2018; 66(6): 1307–1319.

[7] Boydas M G, Ozbek I Y, Kara M. An efficient laser sensor system for apple impact bruise volume estimation. Postharvest Biology & Technology, 2014; 89: 49–55.

[8] Yao Y, Zhang P, Chen Q J, Liu W F, Zeng J, Xie J J, et al. Characterization of pesticide residual dynamics by in situ attenuated total reflection FTIR. Spectroscopy and Spectral Analysis, 2012; 32(12): 3217–3219. (in Chinese)

[9] Sun J, Ge X, Wu X H, Dai C X, Yang N. Identification of pesticide residues in lettuce leaves based on near infrared transmission spectroscopy. Journal of Food Process Engineering, 2018; 41(6): e12816. doi: 10.1111/jfpe.12816.

[10] Chen Q S, Cai J R, Wan X M, Zhao J W. Application of linear/non-linear classification algorithms in discrimination of pork storage time using Fourier transform near infrared (FT-NIR) spectroscopy. LWT - Food Science and Technology, 2011; 44(10): 2053–2058.

[11] Armenta S, Quintas G, Garrigues S, Guardia M. Mid-infrared and Raman spectrometry for quality control of pesticide formulations. Trends in Analytical Chemistry, 2005; 24(8): 772–781.

[12] Jiang S Y, Sun J, Xin Z, Mao H P, Wu X H, Li Q L. Visualizing distribution of pesticide residues in mulberry leaves using NIR hyperspectral imaging. Journal of Food Process Engineering, 2017; 40(4): e12510. doi: 10.1111/jfpe.12510.

[13] Sun J, Jin X M, Mao H P, Wu X H, Tang K, Zhang X D. Identification of lettuce leaf nitrogen level based on adaboost and hyperspectrum. Spectroscopy and Spectral Analysis, 2013; 33(12): 3372-3376.

[14] Yang T M, Zhou R, Jiang D, Fu H Y, Su R, Liu Y X, et al. Rapid detection of pesticide residues in Chinese herbal medicines by Fourier transform infrared spectroscopy coupled with partial least squares regression. Journal of Spectroscopy, 2016; 2016: 9492030. doi: 10.1155/2016/9492030.

[15] Etzion Y, Linker R, Cogan U, Shmulevich I. Determination of protein concentration in raw milk by mid-infrared Fourier transform infrared/attenuated total reflectance spectroscopy. Journal of Dairy Science, 2004; 87(9): 2779–2788.

[16] Yang J B, Du C W, Shen Y Z, Zhou J M. Rapid determination of nitrate in Chinese cabbage using Fourier transforms mid-infrared spectroscopy. Chinese Journal of Analytical Chemistry, 2013; 41(8): 1264–1268.

[17] Su W H, Bakalis S, Sun D W. Potato hierarchical clustering and doneness degree determination by near-infrared (NIR) and attenuated total reflectance mid-infrared (ATR-MIR) spectroscopy. Journal of Food Measurement and Characterization, 2019; 13(2): 1218–1231.

[18] Fisher R A. The use of multiple measurements in taxonomic problems. Annals of Human Genetics, 2012; 7(7): 179–188.

[19] Jin Z, Yang J Y, Hu Z S, Lou Z. Face recognition based on the uncorrelated discriminant transformation. Pattern Recognition, 2001; 34(7): 1405–1416.

[20] Chen M S, Chen H X, Liu W. A new method for resolving the uncorrelated set of discriminant vector. Chinese Journal of Computers, 2004; 27: 913–917. (in Chinese)

[21] Zadeh L A. Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems, 1978; 1(1): 3–28.

[22] Wu X H, Zhu J, Wu B, Huang D P, Sun J, Dai C X. Classification of Chinese vinegar varieties using electronic nose and fuzzy Foley - Sammon transformation. Journal of Food Science and Technology-Mysore, 2020; 57(5): 1310–1319.

[23] Wu X H, Zhu J, Wu B, Zhao C, Sun J, Dai C X. Discrimination of Chinese liquors based on electronic nose and fuzzy discriminant principal component analysis. Foods, 2019; 8(1): 38. doi: 10.3390/foods8010038.

[24] Chen Z P, Jiang J H, Li Y, Liang Y Z, Yu R Q. Fuzzy linear discriminant analysis for chemical data sets. Chemometrics & Intelligent Laboratory Systems, 1999; 45(1): 295–302.

[25] Lin C F, Wang S D. Fuzzy support vector machines. IEEE Transactions

on Neural Networks, 2002; 13(2): 464–471.
[26] Ning Y W, Shi X Y, Yin J G, Xie D W. Application of fuzzy C-means clustering method in the analysis of severe medical images. Journal of Intelligent and Fuzzy Systems, 2020; 38: 1–11.
[27] Cadenas J M, Garrido M C, Martinez R, Munoz E, Bonissone P P. A fuzzy K-nearest neighbor classifier to deal with imperfect data. Soft Computing, 2018; 22: 3313–3330.
[28] Dong C W, Yang Y E, Zhang J Q, Zhu H K, Liu F. Detection of thrips defect on green-peel citrus using hyperspectral imaging technology combining PCA and B-spline lighting correction method. Journal of Integrative Agriculture, 2014; 13(10): 2229–2235.
[29] Wu X H, Wu B, Sun J, Yang N. Classification of apple varieties using near infrared reflectance spectroscopy and fuzzy discriminant C-means clustering model. Journal of Food Process Engineering, 2017; 40(2): e12355. doi: 10.1111/jfpe.12355.
[30] Rozza A, Lombardi G, Casiraghi E, Campadelli P. Novel Fisher discriminant classifiers. Pattern Recognition, 2012; 45(10): 3725–3737.
[31] Foley D H, Sammon J W. An optimal set of discriminant vectors. IEEE Transactions on Computers, 1975; 24(3): 281–289.
[32] Bezdek J C. Pattern recognition with fuzzy objective function algorithms. New York: Plenum, 1981, 1–256.
[33] Barra V, Boire J Y. Tissue segmentation on MR images of the brain by possibilistic clustering on a 3D wavelet representation. Journal of Magnetic Resonance Imaging Jmri, 2015; 11(3): 267–278.