Sensors and Materials, Vol. 33, No. 6 (2021) 2095–2108
MYU Tokyo

S & M 2596

Feature Reduction Method Coupled with Electronic Nose for Quality Control of Tea

Chao Wang,[1,2*] Jizheng Yang,[1,2] and Junhui Wu[3]

[1]Faculty of Electronic Information Engineering, Henan Polytechnic Institute, Nanyang 473000, China
[2]Henan Material Forming Equipment Intelligent Technology Engineering Research Centre, Nanyang 473000, China
[3]Department of Mechanical and Electrical Engineering, Jiangxi Water Resources Institute, Nanchang 330000, China

(Received March 31, 2021; accepted May 26, 2021)

Keywords: electronic nose, sensor detection, feature reduction method, quality control, tea

An effective feature reduction method is a key issue to improve the detection performance of the electronic nose (e-nose). In this study, a feature reduction method coupled with a support vector machine (SVM) was proposed to enhance the detection performance of the e-nose for the quality detection of tea. Firstly, the time-domain features were extracted, which can represent the original gas information of different grades of tea. Secondly, to consider the importance of the relationship between each feature and output category, a subset of multiple features with the best variable importance of projection (VIP) score was generated to obtain the optimized feature set. Finally, kernel principal component analysis (KPCA) and kernel entropy component analysis (KECA) were performed to further reduce the correlation between features to obtain the best feature set. The results indicated that VIP-KECA can obtain the best feature set effectively, and a good classification accuracy of 98% was obtained. This study shows that the feature reduction method is effective for enhancing the detection performance of the e-nose. It also provides an effective technique to monitor the quality of tea.

1. Introduction

The electronic nose (e-nose) is an artificial olfactory system that imitates the human sense of smell.(1) By means of cross-sensitivity, it can obtain the overall smell information of a sample quickly and accurately. The e-nose system is mainly composed of three parts: a sensor array, signal processing, and pattern recognition.(2) The sensor array obtains the smell information of the sample. The signal processing performs feature extraction and processing to remove redundant information, and the pattern recognition makes the classification decision. Owing to its use of sensor detection technology, the e-nose has the advantages of high stability, rapid processing, and simple operation, and it has been widely used in food engineering,(3,4) electrical

*Corresponding author: e-mail: wangchao751215@163.com
https://doi.org/10.18494/SAM.2021.3363

ISSN 0914-4935

engineering,[5] and medical engineering.[6,7] After obtaining the detection data, a feature reduction method will affect the detection performance of the e-nose.

The traditional feature processing method mainly includes feature dimensionality reduction and feature selection. For the feature dimensionality reduction method, Shan *et al.* used principal component analysis (PCA),[8] Kim *et al.* used linear discriminant analysis (LDA),[9] and Peng *et al.* used kernel principal component analysis (KPCA)[10] to reduce the dimensionality of multiple features. These methods mainly convert linear or nonlinear original features into several comprehensive features to remove redundant information, but they do not consider the relationship between each feature and the output category. Feature selection methods include the filter, wrapper, and embedded methods.[11] Fadi *et al.* used the information gain,[12] Rehman *et al.* used recursive feature elimination,[13] and Yun *et al.* used the decision tree[14] to delete redundant information that affects gas identification. These methods fully consider the relationship between each feature and the output category but cannot reduce the correlation between features. Therefore, an effective feature processing method is required, which should consider the relationship between each feature and the output category and reduce the correlation between features to obtain the optimal feature set. In this work, a feature reduction method is proposed. Firstly, the importance of the relationship between each feature and the output category is considered, and the original feature set is selected using the variable importance of projection (VIP) feature selection method. Secondly, the feature reduction methods of KPCA and KECA are used to reduce the correlation between features to obtain the optimal feature set. Finally, a support vector machine (SVM) is used to verify the effectiveness of the feature reduction method, and it is applied to identify the quality of different grades of tea.

Tea is rich in a variety of chemical components with nutritional value and health functions, and is one of the most popular beverages worldwide.[15] Tea is classified into high and low grades. In the tea market, low-grade tea is often fraudulently sold as high-grade tea. At present, the inspection method of the tea quality mainly relies on artificial sensory evaluation, but the method has the disadvantages of strong subjectivity and poor reproducibility.[16] Meanwhile, gas chromatography,[17] liquid chromatography,[18] near-infrared spectroscopy,[19] and other technologies can qualitatively analyze the chemical composition of tea, but the detection of a single chemical composition cannot characterize the overall tea quality. When people choose tea, the smell information, which is the external manifestation of the internal chemical composition of the tea, directly affects their sensory experience. Therefore, the detection and analysis of the smell information of tea based on advanced sensor detection technology can provide an effective method for tea quality monitoring.

In this work, to improve the detection performance of the e-nose and provide a new technology for tea quality detection, a feature reduction method was proposed. Using the VIP method, the time-domain features are preliminarily screened, and the optimal feature set is determined by kernel entropy component analysis (KECA). A grey wolf optimization (GWO) method is also introduced to optimize the two important parameters that affect the classification performance of SVM. The process of feature selection and the algorithm are discussed in detail.

2. Materials and Methods

2.1 Sample preparation

Five different grades of Shucheng Xiaolanhua tea were collected as the samples for smell information detection. Five grams of tea was placed in a 200 ml beaker, and 150 ml of boiled distilled water was added to the beaker, which was covered with a watch glass. After soaking for 5 min, the tea leaves were removed by filtering, and the filtrate was left to cool to room temperature for analysis.

2.2 Electronic nose

The PEN3 e-nose (AIRSENSE Analytics) was used to obtain the smell information of tea. The e-nose system consists of an array of metal oxide gas sensors, a gas sampling device, and a signal processing unit. Figure 1 shows the structure of the system. Ten different metal oxide sensors with specific reactions to different volatile substances are used for sampling. The sensor parameters are shown in Table 1. The response value of the sensor is defined as G/G_0, where G is

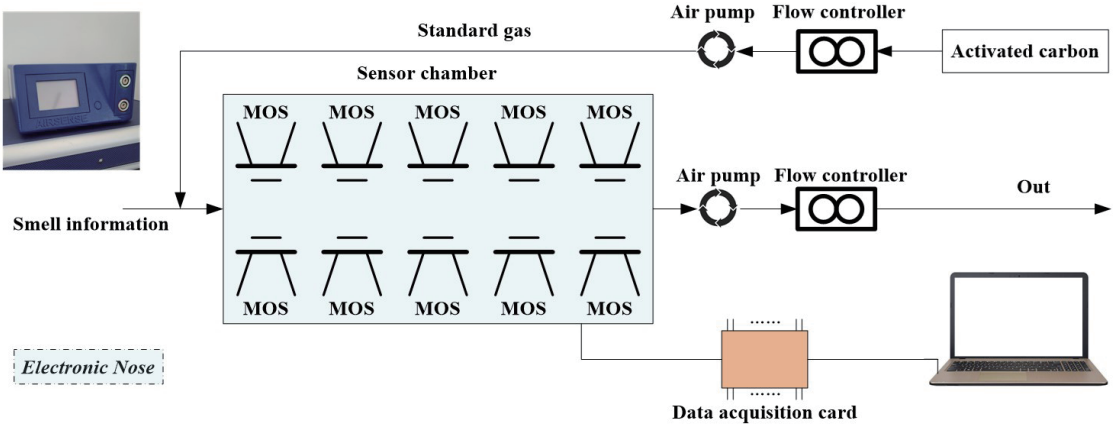


Fig. 1. (Color online) Structure diagram of e-nose system.

Table 1
Sensor information of PEN3 e-nose.

Sensor	Substances detected	Threshold value ($ml \cdot m^{-3}$)
W1C	Aromatics	10
W5S	Nitrogen oxides	1
W3C	Ammonia and aromatic molecules	10
W6S	Hydrogen	100
W5C	Methane, propane, and aliphatic nonpolar molecules	1
W1S	Broad methane	100
W1W	Sulfur-containing organics	1
W2S	Broad alcohols	100
W2W	Aromatics and sulfur- and chlorine-containing organics	1
W3S	Methane and aliphatics	10

the conductance of the sensor when the measured gas enters the air chamber and G_0 is the conductance of the sensor when pure air enters the air chamber.

2.3 Experiment

The environmental temperature of the experiment was 25 ± 0.5 °C, and the humidity was 35 ± 2% RH. The experimental steps were as follows:

(1) Place 50 mL of tea sample in a 200 mL volumetric flask, seal it with plastic wrap, and allow it to stand for 20 min to ensure a sufficient headspace of air.
(2) Pass clean air treated with activated carbon into the sensor air chamber for 60 s with a flow rate of 300 mL/min.
(3) After cleaning and calibrating the sensor array, measure the smell information of tea for a sample for 80 s with a sampling frequency of 1 Hz. Figure 2 shows the response curve of the e-nose.
(4) Repeat steps (1)–(3) for a different sample. Forty samples of each tea were prepared in parallel, and 200 samples were obtained for the five different grades of tea.

2.4 VIP

As an important analysis technique, the VIP score reflects the explainability of independent variables to dependent variables in partial least squares regression (PLSR).[20] The contribution degree of the independent variable (X) to the dependent variable (Y) is called VIP_j, as shown in Eq. (1).

$$VIP_j = \sqrt{\frac{p}{Rd(Y;t_h)}\sum_{h=1}^{m} Rd(Y;t_h)w_{hj}^2} \tag{1}$$

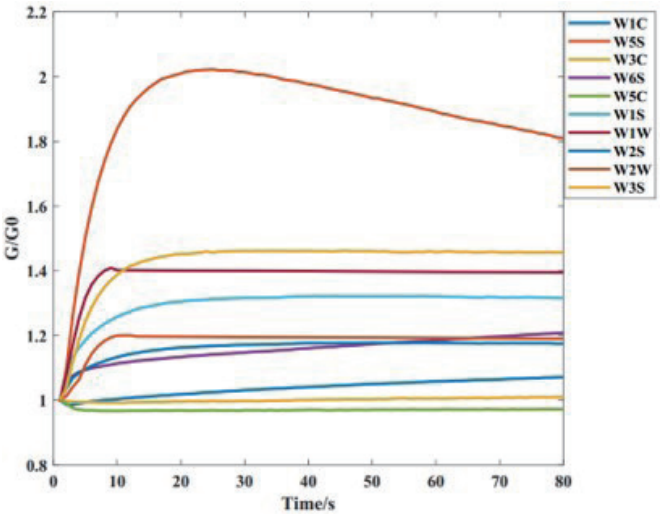


Fig. 2. (Color online) Response curve of e-nose sensor.

Here, t_h is the transfer factor, w_{hj} denotes the principal component of the jth input feature, and $Rd(Y; t_h)$ represents the interpretive ability of t_h for Y. The interpretive of X to Y is transmitted through t_h. If t_h has a strong ability to interpret Y, and X plays an important role in the construction of t_h, it can be considered that the interpretability of X to Y is amplified.

In this work, the independent variable was the time-domain feature extracted from the original information of the e-nose, and the dependent variable was the category label of the different grades of tea. The higher the VIP score, the greater the explanatory effect of the feature on the dependent variable. By using the VIP score, the feature sets $L = \{L_1, L_2, ..., L_k\}$ were generated, where L_i is a subset with i features.

2.5 KPCA and KECA

PCA transforms multidimensional features into several linearly independent comprehensive features by a linear transformation.[21,22] However, the multidimensional features of the e-nose sensor have complicated and nonlinear relationships. Thus, in KPCA, the original features are mapped to a high-dimensional space using a kernel function to make the original features as linear as possible.[23] Equation (2) shows the mapping:

$$K(x,x') = \exp\left(\frac{-\|x-x'\|^2}{\tau}\right), \tag{2}$$

where $K(x, x')$ denotes the kernel matrix, x and x' denote the observation vectors, and τ is the kernel parameter. In KPCA, the dimension of the high-dimensional space depends on the number of samples. Then, K is decomposed, and the principal components with the highest eigenvalues are considered as the eigenvectors. In this work, the number of kernel principal components (KPCs) was determined as that giving a cumulative contribution rate of eigenvalues (PACR) exceeding 95% and τ was set to 100.

In contrast to KPCA, KECA does not use eigenvalues to reduce the feature dimension. Instead, it uses the Renyi entropy to find the direction that reduces the dimensionality in the high-dimensional space. To ensure the minimum information loss, KECA uses eigenvalues and eigenvectors to determine the projection direction.[10] Previous research has shown that the features transformed by KECA will be more beneficial for identification. In the process of dimensionality reduction, the eigenvector is calculated from the Renyi entropy. Similarly to KPCA, the number of kernel entropy components (KECs) was determined as that giving a cumulative contribution rate of Renyi entropy (EACR) of greater than 95%, and the same kernel function and kernel parameter were applied.

2.6 GWO-SVM

The SVM was proposed by Cortes and Vapnik on the basis of statistical theory.[24] It is based on the principle of structural risk minimization and has many advantages for pattern recognition problems, such as a small sample size, nonlinearity, and a high-dimensional feature space.[25] In

the process of pattern recognition, the kernel function maps the feature vector to a space with a higher dimension and establishes a hyperplane in this space. Previous studies have shown that the Radial Basis Function (RBF) kernel function exhibits a good classification performance.[11,20] Therefore, the RBF was applied as a kernel function to map low-dimensional data. The penalty factor c and the kernel function parameter g are important parameters that affect the classification performance of the SVM. Therefore, a GWO was introduced to calculate the important parameters. The fitness function was the highest accuracy of the training set under fivefold cross-validation (CVAccuracy). The best parameters were obtained when the highest CVAccuracy was achieved. In the GWO, the initial number of wolves was 30 and the number of iterations was 100.

3. Results and Discussion

3.1 Feature extraction

Figure 3 shows the sensor response radar chart at 40 s during the detection process for the five different grades of tea, where the central axis of each radar chart shows the sensor response value. The smell information of the five different grades of tea was similar. Therefore, feature

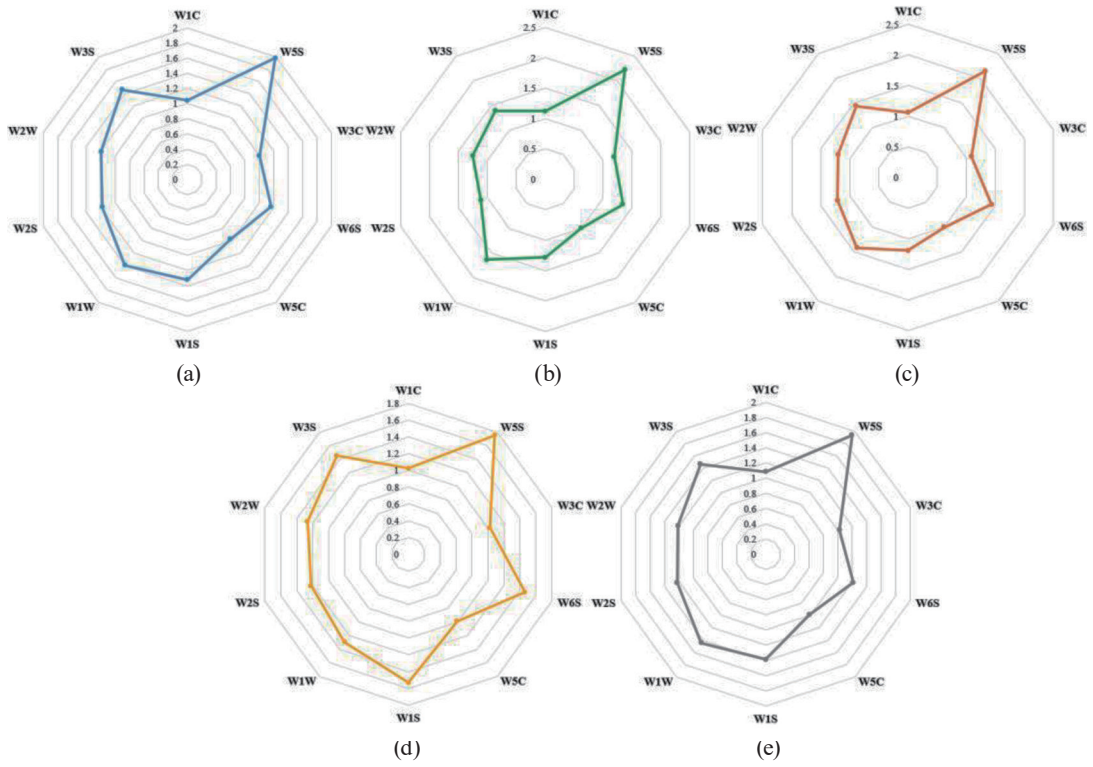


Fig. 3. (Color online) Radar charts of e-nose response information. (a) Super, (b) Grade 1, (c) Grade 2, (d) Grade 3, and (e) Grade 4.

extraction was necessary to represent the original detection signal, which is beneficial for improving the ability to identify the tea quality.

According to Fig. 2, the response signal of the e-nose was stable and the sampling frequency was low. Therefore, the time-frequency domain feature was not considered, and we extracted the time-domain features, which were the maximum value over 1–60 s (MAX), the steady-state average value over 50–60 s (ME), the integrated value over 1–60 s (IN), and the peak factor of over 1–60 s (PF), as expressed by Eqs. (3)–(6), respectively.

$$T_1 = \max\left|f(t)\right| \tag{3}$$

$$T_2 = \frac{1}{10}\sum_{t=50}^{60} f(t) \tag{4}$$

$$T_3 = \int_1^{60} f(t)dt \tag{5}$$

$$T_4 = \max\left|f(t)\right| \Big/ \sqrt{\left[\sum_{t=1}^{60}(f(t))^2\right]\Big/60} \tag{6}$$

Here, $f(t)$ is the response value at time t, T_1 represents the dynamic balance of gas volatilization in the steady detection state, T_2 is the peak value of the gas volatility concentration, T_3 is the dynamic characteristic of gas volatilization in the detection process, and T_4 is the stability of the gas detection process. We used these four features to represent the original detection signal.

3.2 Feature selection based on VIP score

The importance ranking of the 40 features contributing to the classification label was calculated using the VIP score, and the ranking result is shown in Fig. 4. The contribution of the sixth and eighth sensors' integral values to the classification result was high, whereas the

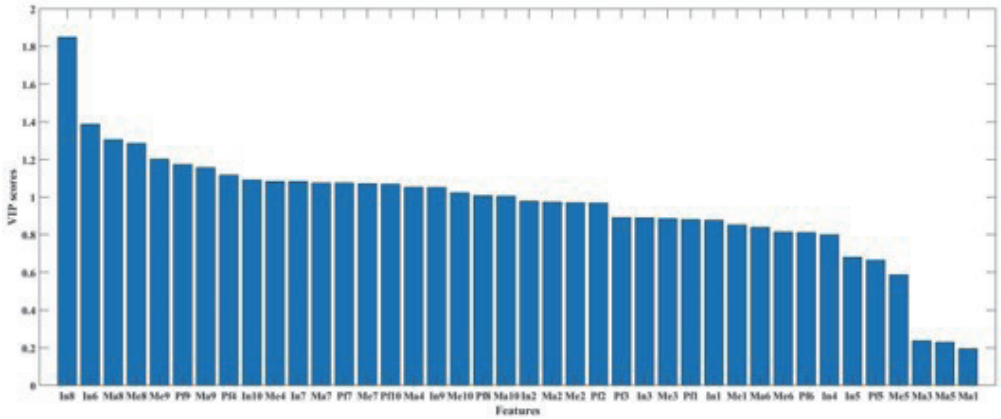


Fig. 4. (Color online) Ranking result of VIP score.

contribution of the first and fifth sensors' maximum values to the classification result was low. However, from the ranking of the VIP scores, we still cannot determine the main features that affect the classification performance. Therefore, 40 feature subsets were obtained by feature accumulation using the VIP score, which were combined with the classification results of GWO-SVM to select the main features that affect the tea quality.

By the Kennard–Stone (KS) method, the training set and testing set were divided, with 30 of the 40 samples of each grade of tea used as the training set and the remaining 10 samples used as the testing set. Therefore, the training set contained 150 samples and the testing set contained 50 samples. The KS method can make the sample distribution of the training set wider and help strengthen the generalization ability of a model.[16] Table 2 shows the classification results of the 40 feature subsets in the GWO-SVM algorithm. Overall, as the number of features increased, the classification accuracy increased while the accuracy fluctuated in a small range, indicating

Table 2
Classification results of different feature subsets based on GWO-SVM.

No.	Features	GWO-SVM	
		Training set (%)	Testing set (%)
#1	In8	67.33	48
#2	In8+In6	67.33	46
#3	In8+In6+Ma8	68	54
#4	In8+In6+Ma8+Me8	68	38
#5	In8+In6+Ma8+Me8+Me9	67.33	58
#6	In8+In6+Ma8+Me8+Me9+Pf9	73.33	60
#7	In8+In6+Ma8+Me8+Me9+Pf9+Ma9	72.67	58
#8	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4	75.33	62
#9	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10	75.33	64
#10	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4	75.33	62
#11	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7	73.33	64
#12	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7	76	78
#13	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7	88	82
#14	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7	88	78
#15	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10	87.33	78
#16	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4	88	80
#17	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9	88	78
#18	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10	91.33	84
#19	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8	91.33	86
#20	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10	90	86
#21	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2	90	86
#22	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2	91.33	84
#23	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2	93.33	90
#24	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2+Pf2	93.33	86
#25	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2+Pf2+Pf3	91.33	90

Table 2
(Continued) Classification results of different feature subsets based on GWO-SVM.

No.	Features	GWO-SVM	
		Training set (%)	Testing set (%)
#26	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2+Pf2+Pf3+In3	91.33	88
#27	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2+Pf2+Pf3+In3+Me3	90	90
#28	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2+Pf2+Pf3+In3+Me3+Pf1	93.33	90
#29	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2+Pf2+Pf3+In3+Me3+Pf1+In1	93.33	92
#30	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2+Pf2+Pf3+In3+Me3+Pf1+In1+Me1	93.33	92
#31	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2+Pf2+Pf3+In3+Me3+Pf1+In1+Me1+Ma6	91.33	92
#32	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2+Pf2+Pf3+In3+Me3+Pf1+In1+Me1+Ma6+Me6	93.33	90
#33	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2+Pf2+Pf3+In3+Me3+Pf1+In1+Me1+Ma6+Me6+Pf6	93.33	92
#34	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2+Pf2+Pf3+In3+Me3+Pf1+In1+Me1+Ma6+Me6+Pf6+In4	93.33	90
#35	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2+Pf2+Pf3+In3+Me3+Pf1+In1+Me1+Ma6+Me6+Pf6+In4+In5	93.33	90
#36	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2+Pf2+Pf3+In3+Me3+Pf1+In1+Me1+Ma6+Me6+Pf6+In4+In5+Pf5	91.33	92
#37	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2+Pf2+Pf3+In3+Me3+Pf1+In1+Me1+Ma6+Me6+Pf6+In4+In5+Pf5+Me5	93.33	92
#38	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2+Pf2+Pf3+In3+Me3+Pf1+In1+Me1+Ma6+Me6+Pf6+In4+In5+Pf5+Me5+Ma3	91.33	90
#39	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2+Pf2+Pf3+In3+Me3+Pf1+In1+Me1+Ma6+Me6+Pf6+In4+In5+Pf5+Me5+Ma3+Ma5	91.33	90
#40	In8+In6+Ma8+Me8+Me9+Pf9+Ma9+Pf4+In10+Me4+In7+Ma7+Pf7+Me7+Pf10+Ma4+In9+Me10+Pf8+Ma10+In2+Ma2+Me2+Pf2+Pf3+In3+Me3+Pf1+In1+Me1+Ma6+Me6+Pf6+In4+In5+Pf5+Me5+Ma3+Ma5+Ma1	91.33	90

that there was a correlation between the features. When the number of features reached 25, the classification accuracy was the same as that of the original feature fusion set, indicating that the original fusion feature contained a large amount of redundant information. When the number of features reached 29, the classification accuracy reached a maximum of 92%, above which it saturated. It was preliminarily determined that these 29 features were the main features that affect the tea quality.

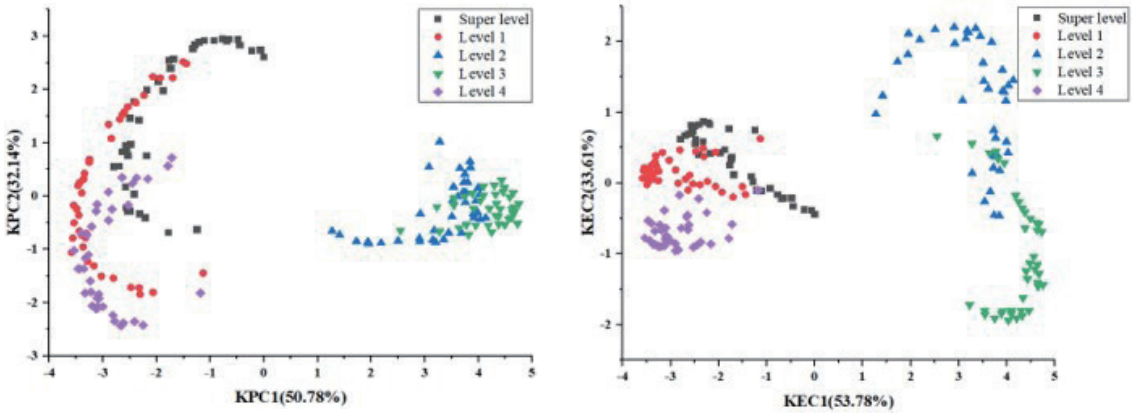
However, the VIP method mainly considers the importance of the relationship between the features (independent variables) and category labels (dependent variables), and ignores the mutual influence between features. Therefore, the selected features may have a strong

correlation, which is not beneficial for improving the classification performance. Therefore, KPCA and KECA were used to process the 29 features to obtain the optimal feature set.

3.3 Dimensionality reduction analysis of KPCA and KECA

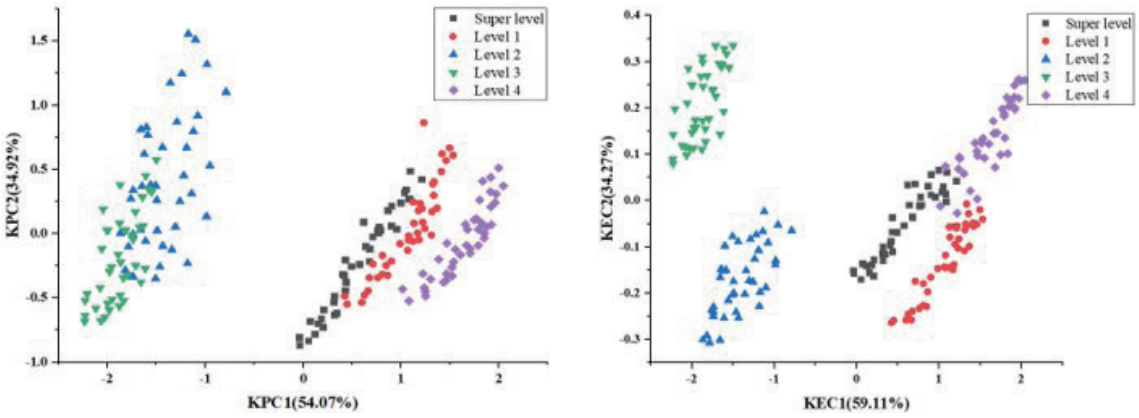
The dimension of the original fusion features was 40. Figure 5(a) shows the dimensionality reduction result of the original fusion features using KPCA. The cumulative contribution rate of the first two principal components was 82.92%, and the super, grade 1, and grade 4 teas had a large overlap, which shows that the smell information of the grade 4 tea was similar to that of the super tea. There was also a large overlap between the grade 2 and grade 3 teas, indicating that the smell information was similar. In contrast to KPCA, KECA uses the Renyi entropy to find the best projection direction. Figure 5(b) shows the dimensionality reduction result of the original fusion features using KECA. The cumulative contribution rate of the first two principal components was 87.39%, and the overlap between the smell features of the different grades of tea was less than that of KPCA, indicating the advantageousness of KECA in dimensionality reduction. However, there was still feature crossover between the super, grade 1, and grade 4 teas and between the grade 2 and grade 3 teas.

The dimension of the feature set selected by the VIP score was 29. Figure 6(a) shows the dimensionality reduction result of the KPCA dimensionality reduction method for the feature set selected by the VIP score. The cumulative contribution rate of the first two principal components was 88.99%. Compared with the dimensionality reduction effect of the original fusion features, the clustering effect of the super, grade 1, and grade 4 teas, was obvious, and the overlap of the grade 2 and grade 3 teas was reduced. Figure 6(b) shows the dimensionality reduction result of the KECA dimensionality reduction method for the feature set selected by the VIP score. The cumulative contribution rate of the first two principal components was 93.38%, and a clear clustering effect was observed in each category, showing the effectiveness of VIP-KECA.



(a) (b)

Fig. 5. (Color online) Dimensionality reduction effect of the original fusion feature set for different methods. (a) KPCA. (b) KECA.



(a) (b)

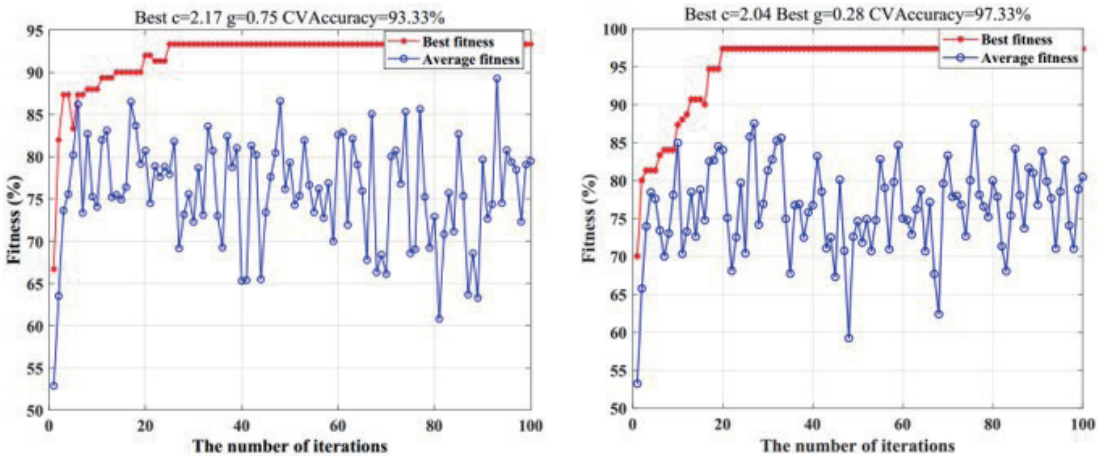
Fig. 6. (Color online) Dimensionality reduction effect of the VIP feature set for different methods. (a) KPCA. (b) KECA.

3.4 Classification results for different feature sets

Table 3 shows the GWO-SVM classification performance for different feature sets. Overall, the performance of the feature subset selected by the VIP score was improved slightly and the number of feature dimensions was reduced slightly. The original fusion features were directly processed by KPCA and KECA, the number of feature dimensions was reduced to 10 and 9, respectively, and the classification performance was improved slightly. The feature set selected by the VIP score was further reduced by KPCA and KECA, the number of feature dimensions was reduced to 6 and 5, respectively, the classification performance was improved significantly, and GWO-SVM obtained classification accuracies of 96% and 98%, respectively. Meanwhile, the feature processing methods of recursive feature elimination (RFE) and Max-Relevance and Min-Redundancy (mRMR) were compared to verify the effectiveness of the feature reduction method. The original fusion features were directly processed by RFE and mRMR, the number of feature dimensions was reduced to 26 and 25, respectively, and the classification performance was improved slightly. Figure 7(a) shows the result of parameter optimization based on the GWO using the VIP feature set. As the number of iterations increased, the information of individual wolves continued to interact, and the optimal fitness function continued to rise. When the number of iterations was 25, the best fitness function of the wolf population was 93.33%, and the best parameters c and g were obtained. That is, the training set achieved the highest classification accuracy of 93.33% under fivefold cross-validation when the optimal parameter c was 2.17 and g was 0.75. Figure 7(b) shows the GWO parameter optimization process under the VIP-KECA feature set. When the number of iterations was 20, the best fitness function of the wolf population was 97.33%, and the best parameters c and g were obtained. That is, when CVAccuracy reached 97.33%, the optimal parameter c was 2.04 and g was 0.28.

Table 3
GWO-SVM classification performance for different feature sets.

Feature set	Dimension	Best c	Best g	Training set accuracy (%)	Testing set accuracy (%)
Original	40	3.21	0.34	86.67	90.00
VIP	29	2.17	0.75	93.33	92.00
KPCA	10	3.14	0.24	90.00	92.00
KECA	9	1.34	0.22	90.00	92.00
VIP-KPCA	6	2.07	0.02	96.00	96.00
VIP-KECA	5	2.04	0.28	97.33	98.00
RFE	26	2.01	0.63	93.33	92.00
mRMR	25	4.62	0.92	90.00	92.00



(a) (b)

Fig. 7. (Color online) Parameter optimization result of GWO-SVM for VIP and VIP-KECA feature sets. (a) VIP. (b) VIP-KECA.

4. Conclusion

In this work, a feature reduction method was proposed to reduce the number of feature dimensions, which was combined with the SVM to identify the smell information of five different grades of tea. The preliminary features that affect tea smell information were screened using the VIP score. Furthermore, KECA was applied to eliminate the correlation between features based on the Renyi entropy. KECA shows a good dimensionality reduction result for the VIP feature set. The GWO was used to optimize the important parameters that affect the classification performance of the SVM, and a high classification accuracy of 98% was obtained.

Acknowledgments

This work was supported by the Pre-Research Project of General Armaments Department (41419070201).

References

1 F. Fedorov, S. Fedor, A. Yaqin, and D. Krasnikov: Food Chem. **345** (2021) 128747. https://doi.org/10.1016/j.foodchem.2020.128747
2 H. Zhu, F. Liu, Y. Ye, L. Chen, J. Liu, A. Gui, J. Zhang, and C. Dong: J. Food Eng. **263** (2019) 165. https://doi.org/10.1016/j.jfoodeng.2019.06.009
3 Y. Shi, X. Jia, H. Yuan, S. Jia, J. Liu, and H. Men: Meas. Sci. Technol. **32** (2021) 025107. https://doi.org/10.1088/1361-6501/abb9e7
4 Y. Shi, H. Yuan, C. Xiong, Q. Zhang, S. Jia, J. Liu, and H. Men: Sens. Actuators, B **333** (2021) 129546. https://doi.org/10.1016/j.snb.2021.129546
5 M. Karim, A. Ganose, and L. Pieters: Chem. Mater. **31** (2019) 9430. http://doi.org/10.1021/acs.chemmater.9b03267
6 T. Saidi, M. Moufid, and K. De Jesus: Sens. Actuators, B **311** (2020) 127932. http://doi.org/10.1016/j.snb.2020.127932
7 L. Giuseppe and H. L. M: Clin. Chem. Lab. Med. **58** (2020) 958. http://doi.org/10.1515/cclm-2019-1269
8 Q. Shan, W. Jun, T. Chen, and D. Dong: J. Food Eng. **166** (2015) 193. https://doi.org/10.1016/j.jfoodeng.2015.06.007
9 H. Kim, B. L. Drake, and H. Park: Pattern Recognit. **40** (2007) 2939. http://doi.org/10.1016/j.patcog.2007.03.002
10 X. Peng, L. Zhang, F. Tian, and D. Zhang: Sens. Actuators, A **234** (2015) 143. http://doi.org/10.1016/j.sna.2015.09.009
11 K. Qian, Y. Bao, J. Zhu, J. Wang, and Z. Wei: J. Food Eng. **290** (2021) 110250. http://doi.org/10.1016/j.jfoodeng.2020.110250
12 T. Fadi, K. Firuz, H. Suhel, and S. S. Reza: Inf. Sci. **534** (2020) 1. http://doi.org/10.1016/j.ins.2020.05.017
13 A. Rehman, and A. Bermak: IEEE Sens. J. **18** (2018) 320. http://doi.org/10.1109/JSEN.2017.2771388
14 L. Yun, Z. Jin, and W. Yuan: Anal. Bioanal. Chem. **410** (2018) 91. http://doi.org/10.1007/s00216-017-0692-0
15 T. Kolackova, D. Sumczynski, V. Bednaril, S. Vinter, J. Orsavova, and K. Kolofikova: J. Food Compos. Anal. **97** (2021) 103792. http://doi.org/10.1016/j.jfca.2020.103792
16 Y. Shi, X. Liu, C. Yin, J. Liu, and H. Men: Anal. Methods **11** (2020) 1460. http://doi.org/10.1039/C9AY02408E
17 L. Tuan, H. Tuan, B. Philippe, and N.-T. Tran-Thi: Food Chem. **326** (2020) 126928. http://doi.org/10.1016/j.foodchem.2020.126928
18 V. Cardoso and R. Poppi: Microchem. J. **164** (2021) 106052. http://doi.org/10.1016/j.microc.2021.106052
19 W. Yu, J. Shan, L. Meng, L. Ying, L. Lu, N. Jing, and Z. Zheng: Comput. Electron. Agric. **175** (2020) 105538. http://doi.org/10.1016/j.compag.2020.105538
20 Y. Li, J. Zhang, T. Li, H. Liu, J. Li, and Y. Wang: Spectrochim. Acta. Part A **177** (2017) 20. https://doi.org/10.1016/j.saa.2017.01.029 http://doi.org/10.1016/j.saa.2017.01.029
21 S. Hod: Phys. Rev. Lett. **105** (2010) 1. http://doi.org/10.1103/PhysRevLett.105.208701
22 J. Collie, E. Hudson, A. Deane, R. Bellomo, and R. Greaves: Ann. Lab. Med. **41** (2021) 414. http://doi.org/10.3343/alm.2021.41.4.414
23 Y. Shi, M. Liu, A. Sun, J. Liu, and H. Men: IEEE Sens. J. **12** (2021) 1. http://doi.org/10.1109/JSEN.2021.3079424
24 C. Cortes, and V. N. Vapnik: Mach. Learn. **20** (1995) 273. http://doi.org/10.1023/A:1022627411411
25 J. Liu, S. Liu, S. Shin, F. Liu, T. Shi, C. Lv, Q. Qiao, H. Fang, W. Jiang, and H. Men: Sens. Mater. **32** (2020) 1767. https://doi.org/10.18494/sam.2020.2715

About the Authors

Chao Wang received his B.S. degree from Shenyang Ligong University, China, in 2001 and his M.S. degree from Huazhong University of Science and Technology, China, in 2011. Since 2001, he has served as a teacher of the Faculty of Electronic Information Engineering, Henan Polytechnic Institute, where he has been an associate professor since 2015. His research interests are in sensors and computer application technology. (wangchao751215@163.com)

Jizheng Yang received her B.S. degree from Henan University, China, in 2014 and her M.S. degree from Northwest University, China, in 2017. Since 2019, she has been a teaching assistant in Henan Polytechnic Institute. Her research interests are in data processing and artificial intelligence. (yangjz1991@163.com)

Junhui Wu received his B.S. degree from Nanchang Hangkong University, China, in 2005 and his M.S. degree from Nanchang University, China, in 2017. Since July 2005, he has been a lecturer in Jiangxi Water Resources Institute, where he has been an associate professor since 2018. His research interests are in intelligent detection and applications. (wujunhui2002@sohu.com)