Rapid Identification of Waste Cooking Oil with Near Infrared Spectroscopy Based on Support Vector Machine

Xiong Shen, Xiao Zheng, Zhiqiang Song, Dongping He, Peishi Qi

Rapid Identification of Waste Cooking Oil with Near Infrared Spectroscopy Based on Support Vector Machine

Xiong Shen[1,a], Xiao Zheng[1,b], Zhiqiang Song[1,c], Dongping He[2,d], Peishi Qi[3,e]

[1]Institute of Mechanical Engineering, Wuhan Polytechnic University, Wuhan 430023,China;
[2]Institute of Food Science and Engineering, Wuhan Polytechnic University, Wuhan 430023, China;[3]PASHUN GROUP, Wuhan 430023,China
[a]sx198711@yahoo.com.cn, [b]zhengxiao@whpu.edu.cn, [c]327463922@163.com, [d]hedp123456@163.com, [e]qps@vip.sina.com

Abstract. The qualitative model for rapidly discriminating the waste oil and four normal edible vegetable oils is developed using near infrared spectroscopy combined with support vector machine (SVM). Principal component analysis (PCA) has been carried out on the base of the combination of spectral pretreatment of vector normalization, first derivation and nine point smoothing, and seven principal components are selected. The radial basis function (RBF) is used as the kernel function; the penalty parameter C and kernel function parameter γ are optimized by K-fold Cross Validation (K-CV), Genetic Algorithm (GA), Particle Swarm Optimization (PSO), respectively. The result shows that the best classification model is developed by GA optimization when the parameters C = 911.33, γ= 2.91. The recognition rate of the model for 208 samples in training set and 85 samples in prediction set is 100% and 90.59%, respectively. By comparison with K-means and Linear Discriminant Analysis (LDA), the result indicates that the SVM recognition rate is higher, well generalization, can quickly and accurately identify the waste cooking oil and normal edible vegetable oils.

Keywords: near infrared spectroscopy, waste cooking oil, support vector machine, parameters optimization

1 Introduction

Catering waste oils include drainage oil (in narrow sense), hogwash fat (waste cooking oil) and fried old oil. After pickling, washing, decoloration, deodorization and other processing, the catering waste oils often close to or completely achieve the national Hygienic Standard of Edible Vegetable Oil in sensory index and conventional typical properties, which consumers and government supervisors are difficult to identify by the sense of the sights and smell. At present, a complete set of testing technology standard of identification of the catering waste oil hasn't been established domestically or abroad. The Ministry of Health is requesting proposals for proposals from the public. Near Infrared Spectroscopy (NIR) technology is a nondestructive testing technique rapidly developed in recent years [1]. The domestic

scholars make use of NIR qualitative analysis to research the types of edible oil [2-4], however, qualitative analysis for catering waste oil is still limited.
Support Vector Machine (SVM) is a new kind of machine learning algorithm based on the minimum principle of statistical learning theory and structural risk, which has advantages of simple structure, strong generalization ability and others. It presents many unique advantages in solving problems of pattern recognition in small sample, nonlinear, high dimension, local minimum [5]. The methods combined SVM with NIR have been applied successfully in identifying the category of tea, milk powder, apple and others [6-9]. The objective of this study is to develop a classified model for catering waste oil and four normal edible vegetable oils by combining SVM with NIR. This model provides a new approach to fast and effective identification of catering waste oil.

2 Experiments and Methods

2.1 Experimental Samples

Catering waste oils used in this experiment include drainage oil and hogwash fat obtained through different degree of refining of decoloration, deodorization, and normal edible vegetable oil which are of different brands or the same brand of different batches in major supermarkets. The samples make up of the following table 1:

Table 1. Composition of the experimental samples

	Training set	Predicting set	In total
The first category: drainage oil and hogwash fat	99	47	146
The second category: soybean oil	40	19	59
The third category: peanut oil	26	7	33
The forth category: olive oil	23	6	29
The fifth category: blend oil	20	6	26
In total	208	85	293

2.2 Experimental Methods

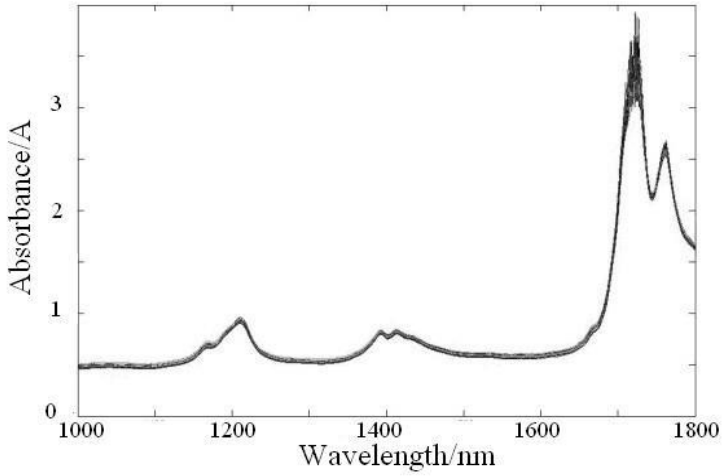
Adopt SupNIR-5700 NIRS (Focused Photonics (Hangzhou), Inc.) to collect NIR spectra of all samples. Spectral measurement of samples uses random RIMP software and its testing method is: transmission, measurement range: 1000~1800nm, scanning speed:10 times/sec, spectral resolution: 6nm, temperature of sample cell: 60°C, testing method: load the sample into the three-quarters of sample bottle, and then place the sample bottle into the sample cell. Stabilized in constant temperature for 5min, the bottle is taken out to check if there exist bubbles. It starts to collect spectrogram if there is no bubble, and each sample averages out three times.

Use NIRS random RIMP software and MATLAB7.8 to collect spectra and convert data format, use chemometrics software Unscrambler X 10.1 to pretreatment the spectral data and analyse principal component, and use SVM pattern recognition and regression software package designed by a professor Lin Zhiren from National Taiwan University to build SVM models in MATLAB7.8 and parameters optimization.

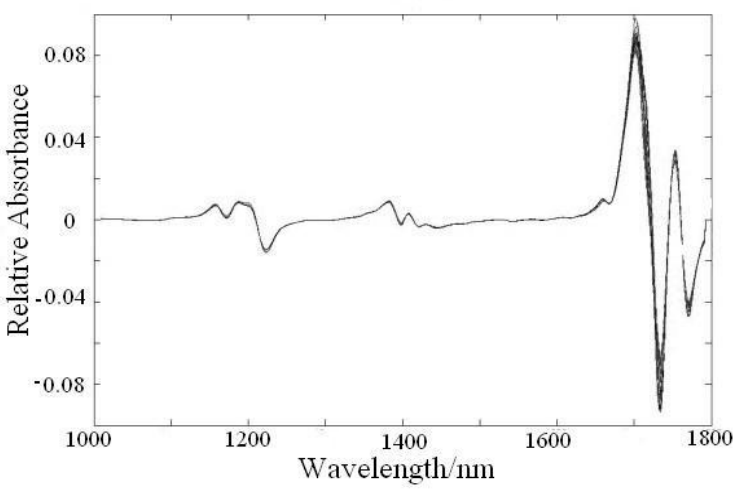
3 Results and Discussion

3.1 Pretreatment for Spectral Data

Besides samples' information collected through NIRS, it contains other irrelevant information and noise, therefore, it is very important and necessary to pretreatment spectra before developing model. Many kinds of methods for spectral pretreatment, including mean centralization, normalization, Savitzky-Golay smoothing, Savitzky-Golay first derivation and second derivation and so on, have been tried in this study. The attempted result indicates that NIR obtains the best pretreatment effect by combining vector normalization with Savitzky-Golay first derivation and nine-point smoothing. Fig.1 shows raw and spectra after pretreatment respectively.



(a) Raw spectra

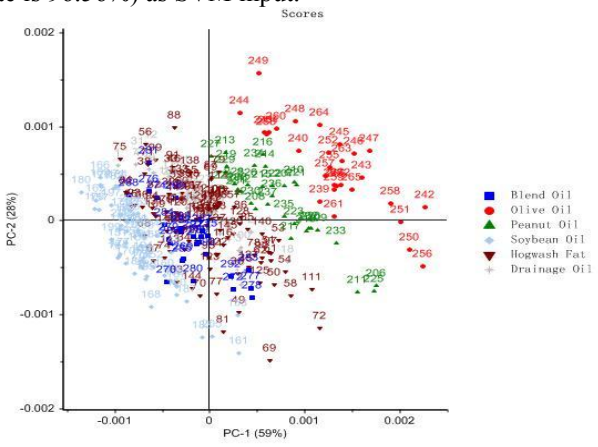


(b) Pretreatment spectra

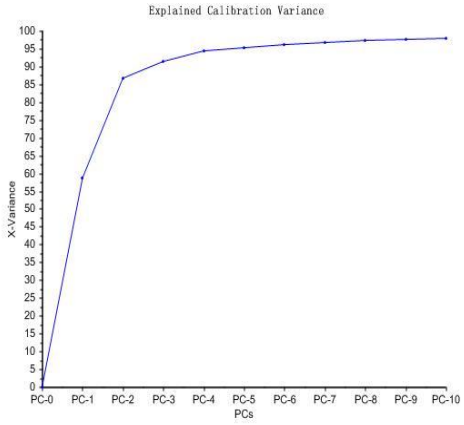
Fig. 1. Conventional and spectra after pretreatment

3.2 Extraction of Spectral Principal Component

Analyze the principal component of spectra after pretreatment, as shown in Fig.2-a, the X-axis stands for the first principal component (PC1), Y-axis represents the second principal component (PC2). The figure shows the good effect of sample distribution. This experiment proves that principal component can reflect most of information when principal component's accumulative contributing rate is above 95% and principal component scree plot (as shown in Fig.2-b) is quite smoothing. Therefore, this paper selects the previous seven principal components (accumulative contributing rate is 96.56%) as SVM input.



(a) PCA SCORE



(b) Explained Variance

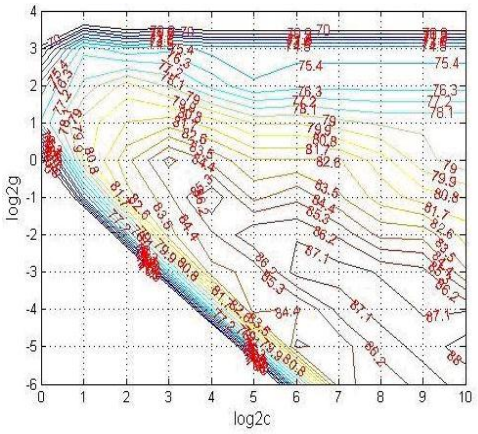
Fig. 2. PCA SCORE and explained variance

3.3 SVM Model Building and Parameter Optimization

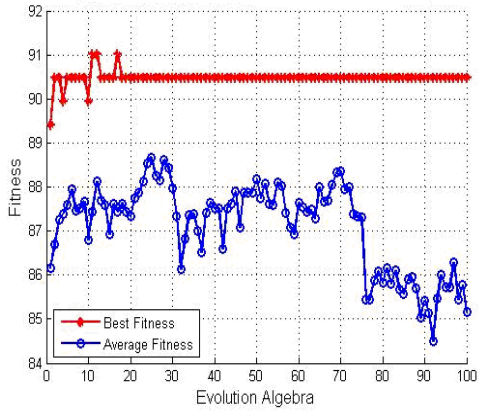
Libsvm includes two classification models: C-SVC and nu-SVC. Based on one-against-one algorithm solving multi-classes pattern recognition, this paper uses C-SVC to establish classification modeling. It needs to select kernel function and parameters when using SVM for pattern recognition. At present there is no unified international model, so we could only use experience or experimental comparison. Typically, using RBF kernel function often gets better simulation results [9], and reduces complexity of computation during the training process. Therefore, this paper makes use of RBF kernel function to establish identification model.

It is very important to select penalty parameter C and kernel function parameter γ in RBF kernel function. C is used to measure the size of the penalty, γ is used to control function regression error and directly influence the initial characteristic value and feature vector. The research respectively uses K-CV, GA and PSO algorithm to optimize the models of C and γ to reach the highest accuracy of classification of training set under the best parameters C and γ. However, it cannot guarantee the testing set to reach the highest accuracy of classification. Fig.3 shows the results of three parameters optimization. Fig.3-a gives the optimization results using K-CV parameter optimization. Fig.3-b gives the optimization results of fitness curve using GA parameter optimization, where the maximum number is 100, the population size is 20, the crossover probability is 0.8, the range of parameters C and γ are 0-1000, other parameters are by default. Fig.3-c gives the optimization results of fitness curve using PSO parameter optimization, where the maximum number of iterations is 100, the initial population size is 20, the learning factor c_1=1.5, c_2=1.7, the range of parameters C and γ are 0-1000, other parameters are by default.

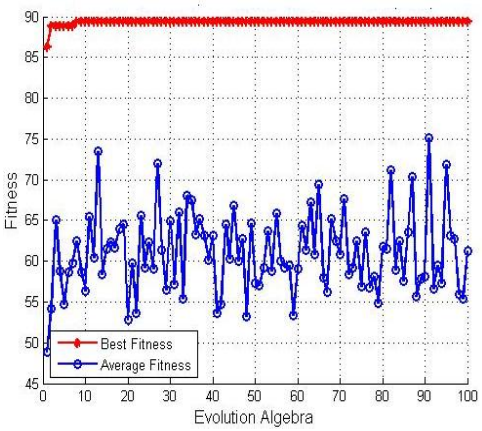
Use the default parameters (C = 1, γ= 1 / K = 0.1429) and optimal results of three different parameters to respectively establish the SVM recognition model, which are analyzed in Table 2.



(a) K-CV



(b) GA



(c) PSO

Fig. 3. The results of three parameters optimization

From the table 2, it is clear that SVM model recognition rate of the default parameters is very low, and almost four kinds of normal edible vegetable oils can be classified as catering waste oils; recognition rate of SVM model increases significantly about 90% after optimal results of different parameters of K -CV, GA and PSO. The learning ability and generalization ability of SVM classifier with optimal parameters C and γ can keep a balance and avoid the occurrence of learning state and non-learning state. Examples show that SVM classification model established when GA optimal parameters C = 911.331, γ = 2.91045, recognition rate of the 208 training sets and 85 predicting sets is 100% and 90.59% respectively, only occurs four blend oils mistaken for catering waste oil, four hogwash oils for blend oils. In the meantime, compared with methods of k-means clustering and LDA, the recognition rate of GA-SVM model is higher than those about 10%. Therefore, SVM model is superior to the methods of k-means clustering and LDA.

Table 2. Different parameters—analysis of SVM modeling results

	Default (C=1, γ=0.1429)		K-CV (C=1024, γ=0.03125)		GA (C=911.331, γ=2.91045)		PSO (C=2287.16, γ=0.01)	
	Returning error number	Predicting error number	Returning error number	Predicting error number	Returning error number	Predicting error number	Returning error number	Predicting error number
The first category	0	0	2	0	0	4	2	1
The second category	40	19	1	0	0	0	1	0
The third category	26	7	0	0	0	0	0	0
The forth category	15	5	0	0	0	0	0	0
The fifth category	20	6	20	6	0	4	20	6
Recognition rate	51.44%	56.47%	88.94%	92.94%	100%	90.59%	88.94%	91.76%

5 Conclusions

The research uses GA-SVM to establish NIR classification model for catering waste oil and four normal edible vegetable oils, and determines the appropriate model parameters. The recognition rate of the established models is achieved respectively 100% for training set and 90.59% for predicting set, the recognition rate and generalization ability of GA-SVM of NIR classification model is higher than conventional analysis model, which can rapidly and accurately identifies the catering waste oil.

The sample source of catering waste oil in the research is limited and cannot completely represent diversity and complexity of catering waste oil. In addition, the law breakers usually add catering waste oil to qualified edible vegetable oil according to a certain proportion, and then sell the fake oil, therefore, it needs to further collect representative adulterated samples in the future.

It is essential to keep developing new methods of qualitative classification to research, and constantly strengthen the maintenance for the models of qualitative classification; in addition, a rapid portable detecting instrument for testing catering waste oils based on the models of NIR quantitative classification needs to be developed in order to protect the security of food production, to provide a more reliable basis for food supervisions and to prevent catering waste oils back to the table.

Acknowledgment

Funds for this research was provided by the National Science and Technology Plan Projects (2009BADB9B08), the major projects foster special of food nutrition and safety of Wuhan Polytechnic University (2011Z06), the entrust projects of Wuhan PASHUN Group green energy technology Co., LTD, and the postgraduate 2010 innovation fund of Wuhan Polytechnic University(2010cx005).

References

1. Lu Wanzhen. Modern Near Infrared Spectroscopy Analytical Technology (Second Edition) [M]. Beijing: Chinese Oil and Chemical Press, 2006, 19-36(in Chinese)
2. Wu Jingzhu, Liu Cuiling, Li Hui et al. Application of NIR technology on identifying types and determining main fatty acid content of edible vegetable oil [J]. Journal of Beijing Technology and Business University (Natural Science Edition), 2010, 28(5):56-59.
3. Liu Fuli, Chen Huacai, Jiang Liyi et al. Rapid discrimination of edible oil by near infrared transmission spectroscopy using clustering analysis [J]. Journal of China Jiliang University, 2008, 19(3):278-282.
4. Li Juan, Fan Lu, Deng Dewen et al. Principal component analysis of 6 kinds of vegetable oils and fats by near infrared spectroscopy. Journal of Henan University of Technology (Natural Science Edition), 2008, 29(5):18-21.
5. Zhang Xuegong. Introduction to Statistical Learning Theory and Support Vector Machines [J]. Acta Automatica Sinica, 2000, 26(1):32-34.
6. Chen QuanSheng, Zhao Jiewen, Zhang Haidong et al. Indentification of Authenticity of Tea with Near Infrared Spectroscopy Based on Support Voctor Machine [J]. Acta Optica sinica, 2006, 26(6):933-937.
7. Zhao Jiewen, Hu Huaiping, Zhou Xiaobo. Application of Support Vector Machine to apple classification with near—infrared spectroscopy [J]. Transactions of the CSAE, 2007, 23(4):149-152.
8. Wu Jingzhu,Wang Yiming, Zhang Xiaochao et al. Applied Study on Support Vector Machines in Identifying Standard and Sub-standard Milk Powder with NIR Spectrometry [J]. Agricultural Mechanization Sciences, 2001, 1(1):155-158.
9. Ye Meiying, Wang Xiaodong. Identification of Chaotic Optical System Based on Support Vector Machine [J]. Acta Optica sinica, 2004, 24(7):953-956.