Identification of peanut oil based on feature layer spectral data fusion method

Wang jie[1, a], Yu yaru[2,b], Wu shuang[3,c], Zheng xiao[4,d], He dongping[5,e]

[1]College of Mechanical Engineering of Wuhan Polytechnic University, China

[2]College of Mechanical Engineering of Wuhan Polytechnic University, China

[3]College of Mechanical Engineering of Wuhan Polytechnic University, China

[4]College of Mechanical Engineering of Wuhan Polytechnic University, China

[5]food science engineering at Wuhan Polytechnic University, China

[a]332653806@qq.com, [b]yuyaru1121@qq.com, [c]410451112@qq.com

Keywords: Peanut oil; Raman spectroscopy; Near infrared spectroscopy; Data fusion; Partial least squares-linear discriminant analysis

Abstract. The purpose of this study is to conduct qualitative analysis on the adulteration in peanut oil by combining data fusion of Raman and near infrared (NIR) spectral characteristics with chemometrics methods. With laser Raman and NIR spectrometer, the spectra of 134 adulterated oil samples and 24 pure peanut oil were collected. The spectra data of Raman and NIR were preprocessed. Competitive adaptive reweighted sampling（CARS）were used to extract the characteristic wavelengths of the spectra data. Combining data fusion technique and partial least squares linear discriminant analysis (PLS-LDA) method, the Ram-PLS-LDA model, NIR-PLS-LDA model and Ram-NIR-PLS-LDA model were established by using the obtained feature layer data. The calibration set and prediction set accuracy of the SG9-airPLS-Nor-CARS-SNV_DT-CARS-PLS-LDA model are 100%. According to the analysis, the prediction accuracy of Ram-NIR-PLS-LDA model is better than that of single spectral model, data fusion technology can enhance the ability to identify the model, which is conducive to practical application. It shows that the two kinds of spectra are complementary, and the using of spectral analysis and data fusion technology has great application value in the identification of edible oil.

Introduction

With the development of China's economy, people have been paying more and more attentions to life quality. Hence, high-quality edible oils are more and more popular. Unfortunately, driven by benefits, many unscrupulous merchants add inferior oils to high-grade edible oils, which do great harm to the interest and health of consumers[1]. In view of the various brands of commercial edible oils, it is valuable to develop a fast and accurate identification technique to evaluate the quality of edible oil.

Nowadays, spectral analysis used in the detection of edible oil quality has been widely studied, because it is fast, accurate and nondestructive compared with other detection methods such as sensory detection and physical–chemical detection. In particular, the near infrared spectroscopy (NIR)[2] and Raman spectroscopy (Raman)[3] have advantages including short detection time, no pretreatment, and online analysis. Thus, both have been widely applied in fields of medicine, petrochemical, food, agriculture, and so on[4]. In 2015, Liu *et al*[5] built a model, with the use of Raman, to fast determine the contents of rapeseed oil, peanut oil, and sesame oil (ternary components) in an oil. In 2014, Zeng *et al*[6], using NIR, determined the contents of inferior oils in a peanut oil, and proposed a model to qualitatively distinguish the species of inferior oils on the basis of the support vector machine classification (SVC) and another model to quantitatively determine the contents of inferior oils on the basis of the support vector machine regression (SVR). On the other hand, data fusion was also applied in our work, and it is a multi-stage and multi-level data-processing procedure, including data level, feature level, and decision level. Each level has its

advantages and disadvantages[7]. The feature level was selected in this paper because on this level, appraisable information compression can be realized, favorable to real-time processing. Moreover, because the extracted features are directly related to decision analysis, the fused results can provide featured information for decision analysis to the best advantage[8].

Aiming at the determination of inferior oils (rapeseed oil, sunflower oil, and soybean oil) in peanut oils, feature-level data fusion technology, combined with chemometrics, were applied to build a PLS-LDA model based on the fusion of feature levels of Raman and NIR data, to realize the fast identification of peanut oils.

Experimental section

Materials and oil samples.

24 peanut oils of different brands, rapeseed oil, sunflower oil, and soybean oil were purchased from the supermarkets in Wuhan City. 3 peanut oils were randomly selected as the matrix oil. Rapeseed oil, sunflower oil, and soybean oil, as inferior oils, were separately added to each kind of the peanut oil. Thus, 9 combinations of binary oils were obtained. The mass fraction of each of the inferior oil varied among the following 15 concentrations: 3%, 5%, 8%, 10%, 12%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, and 90%. Altogether, 135 binary samples were prepared and the 24 pure peanut oils were also sampled. Then, 10 g of each sample was homogenized through shaking, and placed at room temperature for 12 hours. The samples in calibration and prediction sets for the model to determine the contents of inferior oils were selected by using the sample set portioning based on joint x-y distances algorithm (SPXY). The numbers in the calibration and prediction sets were at a ratio of 3 : 1 (Table 1).

Table 1 Statistics of pure and doped peanut oil samples

Sample	Pure peanut oil	Doped peanut oil	Calibration set	Prediction set	Total
Number	24	135	120	39	159

Experimental instruments and spectral data acquisition.

A RamTraceer-200 laser Raman spectrometer (Optotrace Co., Ltd., China) was used, of which the laser wavelength was 785 nm; the resolution was ≤ 8 cm^{-1}; the maximum spectral range was 250–2340 cm^{-1} in frequency and only the spectra between 780–1800 cm^{-1}, at a high signal-to-noise ratio, were collected (Fig. 1); the integration time was 5 s; the laser power was 220 mW (the maximum power was 320 mW). The host of a custom-made laser NIR vegetable-oil-quality detector was Axsun XL410 laser near infrared spectrometer (AXSUN Co., Ltd., USA). The NIR spectral range was 1350–1800 nm in wavelength; each sample was scanned 32 times; the resolution was 3.5 cm^{-1}; the wavelength repeatability was 0.01 nm; the signal-to-noise ratio (250 ms, RMS) was > 5500:1; cuvettes with optical paths of 2, 5, and 10 mm were employed. Each sample was split into 3 equivalent parts for the spectral measurement at room temperature, and then their average spectrum was taken as the final spectrum (Fig. 2).

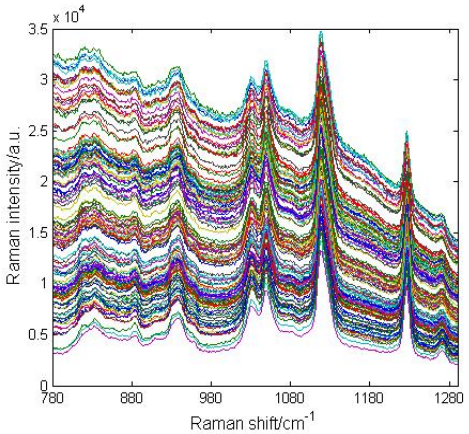

Fig. 1 Original Raman spectra

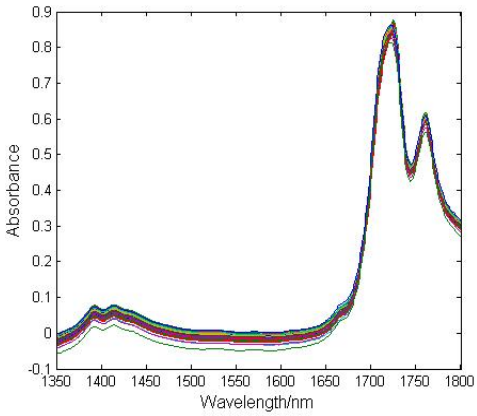

Fig.2 Original near infrared spectra

Data preprocessing.

The original Raman spectral data were processed successively according to the moving average 9-point algorithm (MA9) and Savitzky-Golay filter 11-point smoothing method (SG11), baseline correction adaptive iterative reweighted penalized least squares (airPLS), and asymmetric least squares algorithm (ALS). Then, the data were normalized (Nor) based on the intensity of the characteristic peak at ~1454 cm^{-1}. On the other hand, the original NIR spectral data were processed by using the standard normal variable transformation (SNV) and standard normal variable transformation–detrended algorithm (SNV_DT).

Extraction of featured wavelength spectra and data fusion.

Featured variables in the preprocessed Raman and NIR spectra were extracted, establish the feature spectral extraction were applied to laser Raman spectroscopy, and the extracted featured spectra were used to fit a Raman–NIR feature-level fusion model. The feature levels were selected based on the fusion conditions, the accuracy and practicability of model, and the facts that Raman is susceptible to interference and the bands in NIR are frequently overlapped.

PLS-LDA model for the identification of peanut oil

The partial least squares–linear discriminant analysis method (PLS-LDA) was used to build a peanut oil identification model. In the model, the pure peanut oil and inferior oil were denoted by 1 and -1, respectively. If the predicted value is higher than 0, the sample will be classified as a pure peanut oil; if the predicted value is lower than 0, the sample will be classified as an inferior oil. In detail, there are two parameters in the model: sensitivity (Sen, the possibility that a pure peanut oil samples is correctly identified.) and specificity (Spe, the possibility that an inferior peanut oil sample is correctly identified.).

PLS-LDA model based on Raman or NIR for peanut oil identification.

The PLS-LDA model was fitted to the CARS-optimized feature-level data, resulting in Raman- and NIR-PLS-LDA models. The numbers of optimal principal component in the PLS-LDA models were ascertained with the double cross validation method (DCV). The maximum principal component was set to 20, and the model was built following the 10-fold cross validation. The optimal principal component (optPC) was obtained at the lowest error rate of cross validation (ERCV). The identification results based on the Raman- and NIR-PLS-LDA models are shown in Table 2.

Table 2 Identification results based on the Raman- and NIR-PLS-LDA models

Spectrum	Data preprocessing method	Number of variables in models	optPC	Min ERCV	Calibration set			Prediction set		
					Sen/%	Spe/%	Accuracy rate/%	Sen/%	Spe/%	Accuracy rate/%
Raman	SG9-airPLS-Nor-CARS	88	19	0.0583	100	100	100	100	64.29	74.36
	SG11-airPLS-Nor-CARS	113	20	0	100	100	100	100	83.33	87.18
NIR	SNV_DT-CARS	48	17	0.025	100	97.92	98.33	100	87.18	87.18
	SNV-CARS	60	19	0.0083	100	100	100	100	79.49	79.49

As described in Table 2, the Raman- and NIR-PLS-LDA models can realize the correct identification of peanut oils. Both the Sen and Spe values of the calibration set in the Raman-based model reached 1 and the accuracy rate of calibration set was 100%. The results of the prediction set clearly show that the SG11-airPLS-Nor-CARS model had a higher accuracy rate of 87.18%. In the NIR model, the accuracy rate of the calibration set in the SNV_DT-CARS model reached 100%, but the accuracy rate of SNV-CARS prediction set was higher (87.18%) and had a better identification performance. In Table 2, the best comprehensive performance was present in the SG11-airPLS-Nor-CARS-Ram model.

Feature-level fusing PLS-LDA model for peanut oil identification.

The CARS-optimized NIR and Raman spectral feature-level data were simultaneously selected to fit the PLS-LDA model, resulting in a feature-level fusing Raman-NIR-PLS-LDA model. The prediction results of this model are shown in Table 3.

Table 3 Prediction results of the feature-level fusing Raman-NIR-PLS-LDA model.

Data preprocessing method		optPC	Min ERCV	Calibration set			Prediction set		
Raman	NIR			Sen/%	Spe/%	Accuracy rate/%	Sen/%	Spe/%	Accuracy rate/%
SG9-airPLS-Nor-CARS	SNV_DT-CARS	17	0.0333	100	100	100	100	100	100
SG11-airPLS-Nor-CARS	SNV-CARS	19	0.0083	100	100	100	100	100	100
SG9-airPLS-Nor-CARS	SNV-CARS	15	0.0333	100	100	100	100	93.94	94.87
SG11-airPLS-Nor-CARS	SNV_DT-CARS	18	0.0083	100	100	100	100	100	100

The results in Table 3 indicate that the feature-level fusing Raman-NIR-PLS-LDA model can accurately identify the quality of peanut oil. Based on the fusion of extracted Raman and NIR spectral data at featured wavelengths, the accuracy rate of the prediction set reached 100%, and the accuracy rates of all the calibration sets were 100%, demonstrating that the model has a great prediction performance and good stability.

The prediction results of the feature-level fusing SG9-airPLS-Nor-CARS-SNV_DT-CARS-PLS-LDA and SG11-airPLS-Nor-CARS-SNV-CARS-PLS-LDA model are illustrated in Figure 3 and 4, respectively. It is demonstrated that the identification accuracy rates of pure and inferior peanut oils were both 100%, and the minimum prediction value of pure peanut oil and maximum prediction value of inferior peanut oil were significantly different, indicating that the feature-level fusing PLS-LDA model can effectively identify the quality of peanut oils.

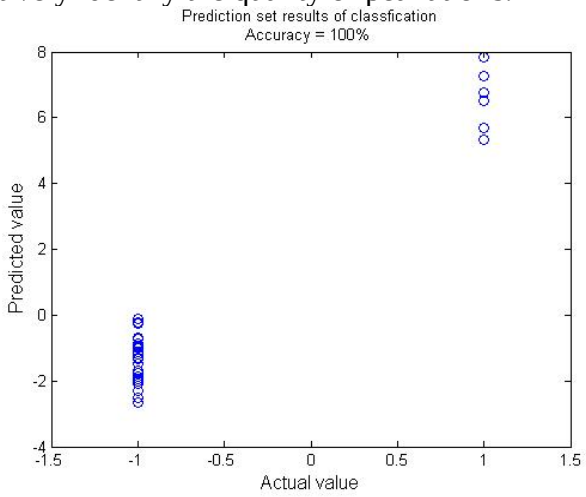

Fig. 3 Prediction results of the samples in the prediction set using the SG9-airPLS-Nor-CARS-SNV_DT-CARS-PLS-LDA model.

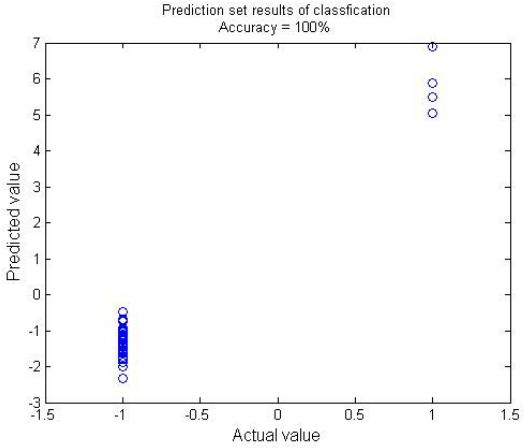

Fig. 4 Prediction results of the samples in the prediction set using the SG11-airPLS-Nor-CARS-SNV-CARS-PLS-LDA model.

Comparing the results in Table 2 and 3, the feature-level fusing Ram-NIR-PLS-LDA model is superior to the single-spectrum-based PLS-LDA model in the specificity (Spe) of identifying inferior peanut oils, which is increased from 83.33% to 100%. The comprehensive accuracy rate is increased from 87.18% to 100%. Compared with the single-spectrum-based NIR-PLS-LDA model, the accuracy rate of the calibration set of the Ram-NIR-PLS-LDA model is 100% and the model is more stable.

Conclusions

The identification model based on data fusion technology and spectral analysis can fast and qualitatively identify inferior peanut oils. The accuracy rate of identifying the samples in the prediction set, with the use of the feature-level fusing Raman-NIR-PLS-LDA model, reached 100%, and the accuracy rates of all the feature-level fusing models in the identification of calibration sets were also 100%. The feature-level fusing models have a better comprehensive performance and higher accuracy rate, compared with Raman- or NIR-based models. In the preprocessing part, the extraction of featured wavelength with the CARS algorithm reduced the amount of data and the

difficulty of modeling, but it did not negatively affect the identification accuracy of the single-spectrum-based PLS-LDA models, indicating that the CARS method could extract the Raman and NIR spectral variables relevant to the identification of samples. However, the preprocessing of the feature-level fusing modeling is more complex and hence the computation amount will be increased, increasing the modeling difficulty. Hence, a better data processing method is required and the work on this algorithm is under investigation.

References

[1] Wang Ruiyuan. The development trend of oil industry in our country -- China cereals and Oils Association branch "twenty-third session of the annual meeting and exhibition at the [J]. grain and oil processing (electronic version), 2014 (11): 24-25.

[2] ZHOU Yang,LIU Tie-bing,LI Jin-rong．Rapid identification between edible oil and swill-cooked dirty oil by using a semi-supervised support vector machine based on graph and near-infrared spectroscopy［J］． Chemometrics and Intelligent Laboratory Systems,2015,143: 1－6．

[3] Mar IA．C,Fernando L,C Esar J S,et al．Raman spectroscopy study of edible oils and determination of the oxidative stability at frying temperatures［J］.European Journal of Lipid Science and Technology,2014,116(11) : 1451 －1456．

[4] Liu Yi, Wang Guoqing, et al., et al. Laser Raman spectroscopy and its application in the petrochemical field [J]. Petroleum chemical engineering, 2014, 43 (10): - 1220

[5] Liu Yande, Jin Tan Tan, Wang Haiyang. Three groups of Raman spectroscopy into edible oil rapid detection of [J]. optics and precision engineering, based on the 2015, 23 (9):2490-2496.[6] Lulu Zeng, Zhiqiang Song, Xiao Zheng, Bin Tu, Cheng Yin, Dongping He2, Peishi QiNIR Spectroscopy-Based Qualitative and Quantitative Detection of Adulteration of Peanut Oil[J]. Applied Mechanics and Materials, Vols. 687-691 (2014) pp 795-801.

[6] Teng Zhaosheng, Yu Wenxian. Instrument [J]. Journal of electric technology, the rapid determination of intelligent moisture based on data fusion 1999, 14 (5).

[7] Wang Gang, Zhang Zhiyu. Research status of multisensor data fusion [J]. electrical measurement and instrumentation, 2006,43 (2): 1-5.