Identification of Edible Oil Based on Multi-source Spectra Data Fusion

YU Yaru[1,a], TU Bing[1], WANG Jie[1], WU Shuang[1], ZHENG xiao[1*], HE Dongping[2]
1.School of Mechanical Engineering, Wuhan Polytechnic University, Wuhan 430023, China
2. School of Food Science and Engineering, Wuhan Polytechnic University, Wuhan 430023,China
[a]232732809@qq.com,*zhengxiao@whpu.edu.com

Keywords: Edible oil, Data fusion; Raman spectroscopy,Near-infrared spectroscopy,Support vector classification

Abstract: An approach based on multi-source spectra data fusion for identification of edible oil is proposed. A qualitative model based on fusion of Raman spectra and near-infrared spectroscopy (Raman-NIR) was established and compared with conventional single-spectra model. The spectra data was pre-processed using the moving average method (MA11), the Savitzky-Golay method (SG9), the adaptive iteratively reweighted penalized least squares method (airPLS), the normalization method (Nor), the multiplicative scatter correction method (MSC), and the standard normal variant and standard normal variant transformation de-trending method (SNV-DT). Then, optimized characteristic variables were selected using the competitive adapt[i]ive reweighted sampling method (CARS-SPA) and the backward interval partial least squares method (BiPLS). Based on that, a model for identification of edible oil was established using the support vector classification method (SVC). The results revealed that the SVC model established can accurately identify and classify eight different edible oil (soybean oil, peanut oil, rapeseed oil, tea seed oil, rice oil, corn oil, sunflower oil, and palm oil). The prediction accuracy for samples in calibration set and prediction set by the proposed model can be 100%, which is superior to that of conventional single-spectra model. The proposed model exhibits excellent generalization capability. Additionally, the study suggests that the Raman-NIR fusion shows improved efficiency in identification of edible oil and great potential for practical application.

1. Introduction

Owing to the differences in compositions of oil fatty acid, edible oils exhibits diversified nutritional characteristics. The combination of stoichiometry and molecular spectra has been widely applied for testing of edible oil. Zhang *et. al.* reported detection of classification and quality of oil using the stoichiometry method [1]. Chen *et. al.* demonstrated identification of edible oil using near-infrared (NIR) spectroscopy [2]. Data fusion refers to a process involving automatic testing, estimation, correlation, and combination of data from multiple sources. It is a multi-aspect, multi-level data processing process that generates significant information. Also, data fusion is regarded as a new target recognition method. Generally, data fusion consists of data level fusion, characteristic level fusion, and decision level fusion. In this article, identification of eight different edible oil (soybean oil, peanut oil, rapeseed oil, tea seed oil, rice oil, corn oil, sunflower oil, and palm oil) was achieved by fusion of the respective data levels and characteristic levels of Raman spectra and NIR spectra. The results were also compared with that of single spectrum method. This study provides references for identification of edible oil.

2. Materials and methods

2.1 Materials

157 samples of eight different edible oil were purchased from local markets and used in this study. These samples were divided into calibration set and prediction set by the ratio of 3:1 using the Sample set Portion based on joint x-y distances (SPXY) algorithm (118 calibration samples and 39 prediction samples).

2.2 Instrument

The key parameters of the RamTraceer-200 laser Raman system (OptoTrace, China) used in this study are as follows: wavelength = 785 nm, wave number range = 250~2340 cm^{-1}, resolution ≤ 8 cm^{-1}, laser power = 220 mW (maximum = 320 mW), and the integration time = 5 s. The key parameters of the AxsunXL410 laser NIR system (AXSUN, USA) used in this study are as follows: detection range = 1350~1800 nm, scan times = 32, resolution = 3.5 cm^{-1}, wavelength reproducibility = 0.01 nm, signal-noise ratio (250 ms, RMS) > 5500:1.

2.3 Sample collection

Three samples were collected for each category of edible oil at room temperature and characterized using the laser Raman system and the laser NIR system. The original spectra of each category of edible oil was determined based on those of the three samples. Fig. 1 shows the Raman spectra of different samples in 780~1800 cm^{-1} (with high signal-to-noise ratio) and Fig. 2 shows the original NIR spectra of different samples.

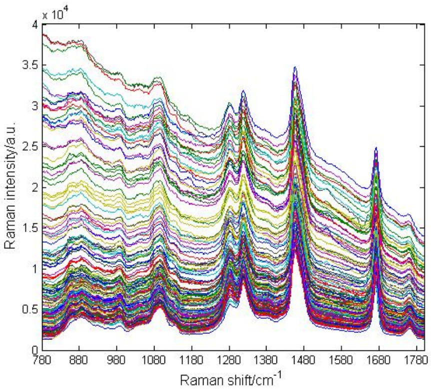


Fig. 1 Original Raman spectra.

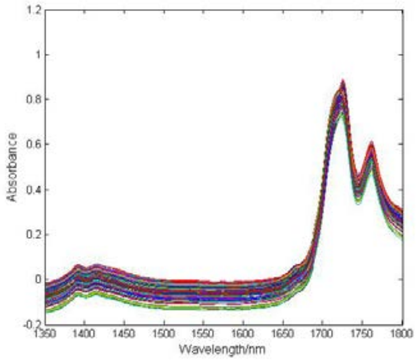
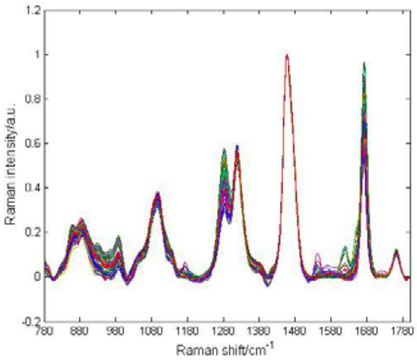


Fig. 2 Original NIR spectra.

2.4 Spectra pre-processing

Indeed, it is challenging to analyze the original Raman spectra obtained due to large noises and severe baseline shift, which were induced by factors such as the temperature. Therefore, the Raman spectra obtained were de-noised using the moving average method (MA11) and the Savitzky-Golay method (SG9); the baseline was calibrated using the adaptive iteratively reweighted penalized least squares (airPLS) algorithm; normalization was achieved with the intensity of the peak at 1454 cm^{-1} as the benchmark. The pre-processed Raman spectra are shown in Fig. 3.

The original NIR spectra, which are limited by severe information overlapping and poor information specificity, were calibrated using he multiplicative scatter correction (MSC) method to enhance the data correlation. Meanwhile, the baseline shift of Raman spectra was eliminated using the standard normal variant de-trending algorithm (SNV-DT). The pre-processed NIR spectra are shown in Fig. 4.



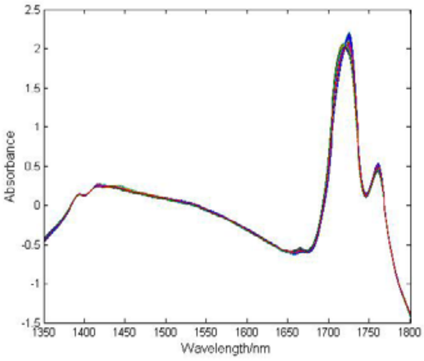


Fig.3 Sample of Raman's preprocessing using MA11-airPLS-Nor

Fig.4 Sample of NIR's preprocessing using SNV-DT

2.5 Determination of characteristic wavelength

In the competitive adaptive reweighted sampling (CARS), the subset corresponding to the minimum root mean square error of cross validation (RMSECV) in the partial least squares (PLS) model was determined using the adaptive reweighted sampling (ARS) technique and the Monte Carlo (MC) sampling technique [3] and is defined as the optimized subset [4]. The MC sampling time was set as 100 and the wavelength variable subset was selected according to the 10-fold RMSECV value in the PLS model. In the successive projections algorithm (SPA), the variable group with minimum redundant information was determined by the vector projection analysis so that the variable co-linearity was minimized to reduce variable quantity and model complexity [5].

The backward interval partial least squares (BiPLS) method can optimize the modeling zone by avoiding the zone with relatively large noises so that the model accuracy can be improved [6].

The characteristic variables of pre-processed Raman spectra and NIR spectra of different samples were determined using CARS-SPA and BiPLS, respectively. The characteristic variables obtained were then used for single spectra modeling and Raman-NIR characteristic layer fusion modeling. Table 1 summarizes characteristic variables of pre-processed Raman spectra and NIR spectra determined using CARS-SPA and BiPLS, respectively.

Table 1. Optimized characteristic wavelengths selected using CARS-SPA and BiPLS methods.

Method	Spectra	Pre-processing method	Variable quantity
CARS-SPA	Raman	MA11-airPLS-Nor	121
		SG9-airPLS-Nor	123
	NIR	MSC	47
		SNV_DT	44
BiPLS	Raman	MA11-airPLS-Nor	511
		SG9-airPLS-Nor	562
	NIR	MSC	112
		SNV_DT	112

3. Establishment of SVC model for edible oil testing

3.1 Single spectra modeling

SVC models (Raman-SVC) with Raman spectra pre-processed by SG9-airPLS-Nor and MA11-airPLS-Nor as input variables and full wavelength SVC models with NIR spectra pre-processed by MSC and SNV-DT as input variables were established in MATLAB. Also, SVC models with Raman and NIR characteristic variables optimized by CARS-SPA and BiPLS as input variables were established. The (C, g) of NIR-SVC model was optimized using the grid search (GS) method.

Table 2 summarizes the predictions by single spectrum SVC models. As observed, full wavelength and characteristic variable optimized single spectrum SVC models can effectively identify edible oil. The full wavelength SVC model pre-processed by SG9-airPLS-Nor and SNV_DT exhibited excellent overall identification capability. Raman-SVC models whose characteristic variables were optimized by CARS-SPA and BiPLS showed improved calibration set accuracy (up to 99.15%), indicating that CARS-SPA and BiPLS can enhance the model reliability by optimizing input variables of Raman spectra. The SG9-airPLS-Nor-BiPLS pre-processed model exhibited relatively low C and g, and its accuracy towards samples in calibration set and prediction set were 98.31% and 100%, respectively (optimized overall performance). The prediction accuracy of full wavelength and characteristic variable optimization NIR-SVC models showed no significant improvement compared with Raman-SVC models, while their modeling duration was significantly reduced. In other words, characteristic variable optimization can enhance the modeling efficiency. However, C and g of NIR-SVC models were significantly higher than those of Raman-SVC models, resulting in limited prediction performances. Fig. 5 shows predictions of Raman spectra by the SG9-airPLS-Nor-BiPLS-SVC model.

Table 2. Predictions by single spectra SVC model.

Spectra	Pre-processing method	Variable quantity	Parameter		Calibrati-on set	Prediction set
			C	g	Accuracy/%	Accuracy/%
Raman	MA11-airPLS-Nor	1221	8	2	96.61	100
	SG9-airPLS-Nor	1221	4	2	97.46	100
	MA11-airPLS-Nor -BiPLS	511	4	8	98.31	100
	SG9-airPLS-Nor-BiPLS	562	4	4	98.31	100
	MA11-airPLS-Nor-CARS-SPA	121	32	8	99.15	97.44
	SG9-airPLS-Nor-CARS-SPA	123	16	4	96.61	100
NIR	MSC	451	8	32	94.92	97.45
	SNV_DT	451	2	16	96.61	100
	MSC-BiPLS	112	256	8	95.76	97.44
	SNV_DT-BiPLS	112	2	32	96.61	100
	MSC-CARS-SPA	47	256	16	94.92	97.44
	SNV_DT-CARS-SPA	44	4	32	94.92	97.44

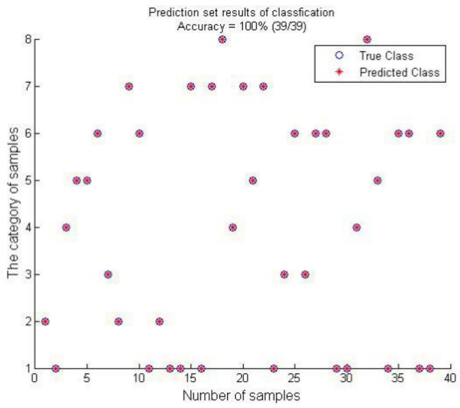


Fig. 5 Prediction about G9-airPLS-Nor-BiPLS-SVC model.

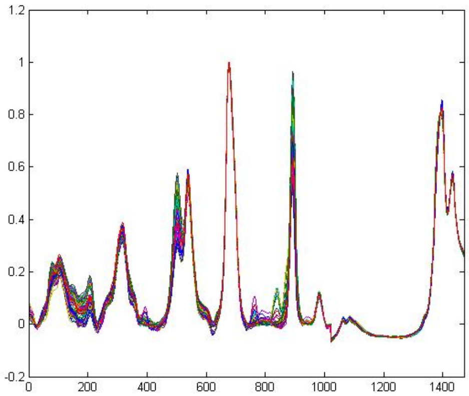


Fig. 6 Raman-NIR fusion spectra.

3.2 Raman-NIR SVC modeling based on data level fusion

The data level fusion refers to a process where the X coordinates of normalized Raman and NIR spectra were connected end-to-end while the Y-coordinates were standardized. The data level fusion can reflect variable of light intensity. Fig. 6 shows the fusion spectra. The pre-processed Raman and NIR spectra after data fusion were used as input variables for SVC model for identification of edible oil to establish data fusion Raman-NIR SVC edible oil identification model. Table 3 summarizes the parameters and prediction results by the data level fusion Raman-NIR SVC model.

The Raman-NIR SVC model by data level fusion of Raman and NIR spectra showed 100% accuracy for identification of edible oil. Specially, the MA11-airPLS-Nor-SNV_DT model exhibited 100% accuracy for both calibration set and prediction set; *g* of the MA11-airPLS-Nor-SNV_DT model was 2, indicating good generalization capability of this model. In summary, the MA11-airPLS-Nor-SNV_DT model exhibited optimized overall performances among the four models. Compared to Raman and NIR single spectrum model, the Raman-NIR SVC model exhibits several advantages. First, the Raman-NIR -SVC model showed improved prediction accuracy due to data fusion of Raman spectrum and NIR spectrum. Second, the Raman-NIR SVC model exhibited 100% accuracy for samples in the prediction set, indicating excellent model reliability. Third, *g* of the Raman-NIR SVC model was significantly reduced, indicating good generalization capability of this model.

Table 3. Predictions by Raman-NIR SVC based on fusion of characteristic layers.

Pre-processing method		Parameter		Calibration set	Prediction set
Raman	NIR	*C*	*g*	Accuracy/%	Accuracy/%
MA11-airPLS-Nor	MSC	32	2	100	100
MA11-airPLS-Nor	SNV_DT	16	2	100	100
SG9-airPLS-Nor	MSC	32	0.5	99.15	100
SG9-airPLS-Nor	SNV_DT	8	0.25	95.76	100

Fig. 7 shows the predictions by the model based on data level fusion of MA11-airPLS-Nor pre-processed Raman spectrum and SNV-DT pre-processed NIR spectrum.

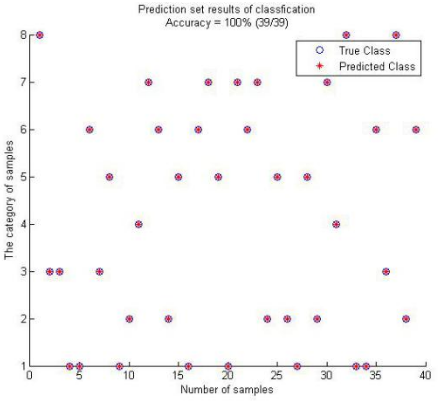


Fig. 7 Prediction by MA11-airPLS-Nor model.

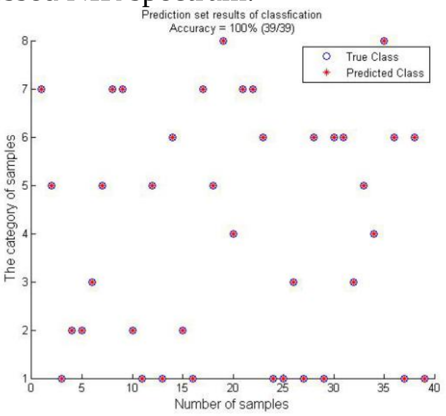


Fig. 8 Predictions by SG9-airPLS-Nor -BiPLS-SNV_DT-CARS-SPA model.

3.3 Raman-NIR SVC modeling based on characteristic level fusion

The characteristic level fusion refers to the data fusion of characteristic variables of the Raman spectrum and the NIR spectrum. The Raman and NIR spectra data optimized by the characteristic wavelength determination method were used as the input variables to establish the characteristic level fusion Raman-NIR SVC model for identification of edible oil. Table 4 summarizes the parameters and prediction results by the characteristic level fusion Raman-NIR SVC model.

Table 4. Predictions by of Raman-NIR SVC model based on fusion of characteristic layers.

Pre-processing method		Parameter		Calibrat-ion set	Predicti-on set
Raman	NIR	*C*	*g*	Accuracy/%	Accuracy/%
MA11-airPLS-Nor-BiPLS	MSC-BiPLS	16	4	100	100
MA11-airPLS-Nor-BiPLS	SNV_DT-BiPLS	4	8	100	100
MA11-airPLS-Nor-BiPLS	SNV_DT-CARS-SPA	16	2	100	100
MA11-airPLS-Nor-CARS-SPA	SNV_DT-BiPLS	128	0.5	99.15	100
SG9-airPLS-Nor-BiPLS	MSC-CARS-SPA	32	1	99.15	100
SG9-airPLS-Nor-BiPLS	SNV_DT-CARS-SPA	32	1	100	100
SG9-airPLS-Nor-CARS-SPA	MSC-CARS-SPA	64	4	100	100
SG9-airPLS-Nor-CARS-SPA	SNV_DT-BiPLS	64	1	99.15	100

Compared with Raman and NIR single spectrum models, the Raman-NIR SVC model established by characteristic level fusion of Raman spectrum and NIR spectrum showed significantly reduced *g*, indicating improved generalization capability and application potential. Compared with the Raman-NIR SVC model established by data level fusion of Raman spectrum and NIR spectrum, the Raman-NIR SVC model established by characteristic level fusion of Raman spectrum and NIR spectrum showed 100% accuracy for samples in the prediction set and almost 100% accuracy for samples in the calibration set for identification of edible oil. Indeed, the Raman-NIR SVC model established by characteristic level fusion of Raman spectrum and NIR spectrum showed improved prediction accuracy and reliability. Additionally, the information by characteristic level fusion is optimized compared with information by data level fusion, indicating that characteristic variable optimization can eliminate unnecessary information and facilitate modeling. Specially, *g* of the SG9-airPLS-Nor-BiPLS-SNV_DT-CARS-SPA-SVC model was 1 and prediction of accuracy for samples in calibration set and prediction set were 100%. Fig. 8 shows predictions by characteristic level fusion of characteristic wavelength determined Raman spectrum optimized by SG9-airPLS-Nor-BiPLS and NIR spectrum optimized by SNV_DT-CARS-SPA.

4. Summary

This study proposes a model based on data level and characteristic level data fusion of Raman spectrum and NIR spectrum for identification of eight different edible oil. Compared with conventional single spectrum models, the data fusion based model exhibited significantly improved prediction accuracy and application potential. Additionally, the characteristic level data fusion showed improved prediction accuracy and reliability.

Acknowledgement

Transformational projects of technological innovation and achievements about food in Hubei at 2016

Research Projects of Wuhan Science and Technology (2013010501010147)

Major projects of Food Nutrition and Safety in Wuhan Polytechnic University (2011Z06)

Innovative Fund Project about Master of Wuhan Polytechnic University (2014 cx 005)

References

[1] Y. Zhang, Y. Xie, S. Sun. Application of near-infrared spectroscopy combined with chemometrics in oil detection. Cereals & Oils, 2015 (1): 66-68.

[2] J. Chen, Y. Xu, Q. Huang. Application of near-infrared spectroscopy in edible oil testing. Guide of Food Safety, 2016 (9): 89-90.

[3] S. Nawar, H. Buddenbaum, J. Hill. Estimating the soil clay content and organic matter by means of different calibration methods of vis-NIR diffuse reflectance spectroscopy. Soil and Tillage Research, 2016, 155: 510-522.

[4] W. He. Application of near-infrared spectroscopy in edible oil testing. Guide on Consumption, 2015 (12): 355.

[5] M. Araújo, T. Saldanha, R. Galvão. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometrics & Intelligent Laboratory Systems, 200l, 57 (2): 65-73.

[6] Q. Yang, Q. Zhu, P. Ren. Wavelength selection for NIR spectroscopic analysis of chemical Oxygen demand based on Different Partial Least Squares models. Journal of Analytical Science, 2016, (4): 485-489.