Three different SVM classification models in Tea Oil FTIR Application Research in Adulteration Detection

ZHANG Yu¹, LI Tinghui*¹
¹College of Electronic Engineering, Guangxi Normal University, Guilin Guangxi 541004, China  
*Corresponding author’s e-mail: tinghuilii@gxnu.edu.cn

Abstract. Fourier transform infrared spectroscopy (FTIR), as a new type of rapid environmental detection technology, has attracted extensive attention from researchers. In this study, FTIR combined with stoichiometry was used to establish and study the qualitative detection model of tea oil-doped soybean oil and corn oil. Different spectrum preprocessing and feature value extraction were performed on the infrared spectra of 105 samples of tea oil and adulterated oil in the spectral band of 600-4000 cm⁻¹. The KS sample selection method was used to divide the training set and the test set, and the training set was used to construct Support Vector Machine (SVM) Classification model, select the best model construction method based on the accuracy of the test set. The results show that the SVM tea oil adulteration detection model with convolution smoothing (SG) spectral preprocessing and random forest (RF) feature extraction achieves the best accuracy of 0.93, which is simple, fast, and pollution-free for the market to detect adulterated tea oil. Provide technical reference.

1. Introduction
Camellia oil is a kind of high-end woody edible oil unique to our country. Because of its reasonable fatty acid composition, rich nutrient content and various health care values, its market is sought after by consumers [1]. With the increasing price of tea oil, some illegal operators in the market mixed tea oil with other relatively cheap vegetable oils in order to make huge profits, which seriously harmed the personal legitimate interests of consumers.

Traditional methods for detecting adulteration of edible oils mainly include: chromatography [2-4], mass spectrometry [5-6], nuclear magnetic resonance method [7-8] and electronic nose method [9]. Compared with traditional detection methods, infrared spectroscopy technology has the characteristics of fewer pre-processing steps, short detection cycle, suitable for on-site rapid detection, less damage to the detection sample, and simple operation. Therefore, it has become a popular new detection technology [10-11]. At present, NIRS (Near Infrared Spectroscopy) is the most researched aspect of the identification of adulterated tea oil types and the quantitative detection of tea oil quality [12-14]. But compared with NIRS, FTIR has more advantages in terms of resolution, signal-to-noise ratio, sensitivity, result reproducibility and accuracy. In recent years, FTIR has been widely reported in infrared spectroscopy detection technology [15-18], but there are relatively few studies on tea oil adulteration detection. Researcher Liu Lingling [19] used FTIR combined with cluster independent soft pattern recognition technology (SIMCA) to establish a rapid identification method for the authenticity of edible oil. Researcher Zhu Qisi [20] et al. studied the method of quickly identifying camellia oil by FTIR spectroscopy.
Due to the complexity of the information of infrared spectroscopy data, the dimensional disaster of the data and other reasons, different spectral preprocessing and feature extraction methods have different effects on the stability and accuracy of the model. This paper takes tea oil and adulterated oil soybean oil and corn oil as the research objects. By collecting the FTIR spectra of tea oil and two adulterated tea oils, three different SVM tea oil adulteration detection models are established. Provide technical reference for classification and identification of adulterated oil.

2. Materials and Methods

2.1. Instruments and materials

Using Agilent Technologies Co., Ltd. Cary 630 FTIR spectrometer, the test method is attenuated total reflection. The tea oil samples were from Sanjiang Dong Autonomous County, Guangxi, and the other edible oil samples were commercially available Fulinmen soybean oil and longevity corn oil.

Preparation of adulterated oil samples in this experiment: Soybean oil, corn oil and tea oil were mixed in volume ratios of 2:8, 3:7, 4:6, 5:5, 6:4, 7:3, 8:2. Into 3 kinds of tea oil of different brands. Configure adulterated tea oil with 6 adulterated combinations and 42 adulterated samples. For each adulterated sample, randomly take 2-4 sets of parallel samples. In the end, a total of 105 samples were allocated, including 20 tea oil samples, 42 adulterated tea oil samples containing soybean oil, and 43 adulterated tea oil samples containing corn oil.

2.2. Experimental Method

2.2.1. FTIR Spectral Acquisition

First put 105 prepared samples at room temperature for 24 hours, and then perform FTIR spectrum collection. The parameters of the spectrometer are: the spectrum collection range is 400-4000 cm⁻¹, the number of background spectral scans is 64, the number of sample scans is 64, and the spectral resolution is 4 cm⁻¹. Use a rubber-head dropper to suck a sample of about 20uL, place it on the surface of the ATR appendage of FTIR, and scan to obtain a spectrum. In order to eliminate the influence of the background, the background spectrum is collected before each sample is scanned; in order to make the spectrum data of the sample more suitable for the subsequent data model analysis, the spectrum is collected in 3 days and is operated by 2 experimenters separately.

2.2.2. Data Processing

Use Matlab2017b software to import 105 sample FTIR spectrum data, data processing, modeling and classification research. Considering that the main noise of the instrument itself is distributed in the low wave number range, the data removes the wave number range of 400-600 cm⁻¹; through the outlier detection method of normal distribution, 10 abnormal data samples are found, so they are removed; for the remaining 95 The preprocessing methods selected for effective sample data are: window smoothing, convolution smoothing (SG), first derivative method, multivariate scattering correction (MSC), and standard normal variable transformation (SNV). In addition, in order to improve the running speed of the model, it is necessary to select a suitable feature extraction method to reduce the dimensionality of the spectral data. At the end of the experiment, the best preprocessing method and feature extraction method are selected by comparing the modeling results.

2.2.3. Model Establishment and Evaluation

Support vector machine (SVM), as a traditional machine learning classification algorithm, has unique advantages in solving small sample, nonlinear and high-dimensional pattern recognition, so it is used as the main modeling method of this research. The modeling process: the pure tea oil sample is numbered 1; the tea oil mixed with soybean oil is numbered 2; the tea oil mixed with corn oil is numbered 3. Divide the preprocessed spectral data into training set and test set by KS selection algorithm, use training set to build SVM tea oil adulteration classification model, analyze 10-fold cross-validation accuracy of the model.
training set, and select the best model parameters; finally use the classification accuracy $e_1$ of the test set as the model evaluation index, and select the best combination of spectral preprocessing-feature extraction methods.

3. Results and Analysis

3.1. FTIR Spectral analysis

FTIR spectroscopy was performed on tea oil, soybean oil, corn oil, and adulterated oil samples. The bands with large noise at both ends and 10 abnormal samples were removed, and finally 95 samples with 600-4000 cm$^{-1}$ spectral data were retained. The FTIR spectrum curves of three different types of edible oils, tea oil, soybean oil, and corn oil, are shown in Figure 1. From the spectral curves of three different types of edible oils, it is found that the overall change trend is basically the same. The main reason is that the chemical composition of edible oils is similar to that of C-H groups. The three spectral curves all have obvious characteristic absorption peaks at 722, 1096, 1120, 1159, 1237, 1377, 1457, 1743, 2853, 2922 and 3010 cm$^{-1}$, and their absorption intensity has slight differences; and at 600 within the fingerprint spectrum range of -1500 cm$^{-1}$, the spectral curves are slightly different, and it is found that the spectral curves of soybean oil and corn oil at 900-920 cm$^{-1}$ are higher than those of crude tea oil and refined tea oil.

![Figure 1. FTIR spectra of tea oil, soybean oil, and corn oil](image1)

![Figure 2. FTIR spectra of tea oil and soybean oil doped in different proportions](image2)
Figure 2 shows the FTIR spectra of seven different ratios of tea oil mixed with soybean oil. Compared with Figure 2, the spectral curves of different adulteration ratios are more similar. It is difficult to effectively analyze the adulterated tea oil from the spectra. And there is no obvious distinguishable feature in the fingerprint area, so it is necessary to use chemometrics and machine learning methods to establish a detection model that can identify adulterated tea oil.

3.2. The influence of different pretreatments on tea oil adulteration model
Spectral noise and interference factors have always been one of the reasons that affect the accuracy and stability of the model. Choosing a suitable preprocessing method can effectively filter the noise information in the FTIR spectrum and retain effective information. The preprocessing methods mainly used in this experiment are: smoothing, derivative, MSC and SNV. Use the K-S selection method to divide the training set and the test set for the processed spectral data, use the training set for SVM modeling, 10-fold cross-validation for model parameter tuning, and test set for model evaluation.

| Pretreatment Method | e_0 | e_1 | Time (s) |
|---------------------|-----|-----|---------|
| Original            | 0.72| 0.72| 90.573  |
| Moving Smooth       | 0.75| 0.65| 87.659  |
| SG                  | 0.95| 0.82| 80.165  |
| 1st                 | 0.59| 0.27| 92.339  |
| MSC                 | 0.77| 0.86| 100.553 |
| SNV                 | 0.77| 0.86| 100.553 |

It can be found from Table 1 that for the SVM tea oil adulteration detection model, the accuracy of the test set of different preprocessing methods is significantly different. Among them, the SG smooth 10-fold cross-validation accuracy e_0 is the highest in the test set. However, the performance of MSC and SNV processed spectral data is the best in the performance of the test set, and the accuracy rate e_1 is 0.86. Compared with the original data, the accuracy of the first derivative processing has dropped significantly. From this, it can be concluded that using the spectra processed by SG, MSC and SNV for modeling research is better.

3.3. The influence of feature extraction on tea oil adulteration model
Feature extraction is a key step in the establishment of infrared spectroscopy models. Eliminating variables with less correlation will greatly improve the performance and running time of the model. Feature extraction methods have always been the focus of researchers. The main feature extraction used in this research are principal component analysis (PCA), continuous projection algorithm (SPA) and random forest (RF).

3.3.1. PCA-SVM Camellia oil adulteration model
PCA is one of the commonly used methods of dimensionality reduction. Its main idea is to make full use of data, replace more old variables with fewer new variables, and make the new variables reflect the information of the original variables as much as possible. Perform PCA processing on the original spectra to obtain the cumulative reliability of the principal components as shown in Figure 3. In order to better reflect the original spectral information of the sample, select the principal components with a cumulative feasibility of 99%.
According to 2.2 experimental research and analysis, the prediction results of the SVM tea oil adulteration detection model using different pretreatments and PCA are shown in Table 2. It can be seen from Table 2 that after PCA is used for dimensionality reduction operations, the distribution range of the classification accuracy $e_1$ of the model under different preprocessing on the test set is 0.65-0.79. Compared with Table 1, the accuracy is reduced and the model time is greatly reduced. The reason for the research and analysis is that PCA fusion transforms the original data, which changes the original data information, and some noise information produces greater interference during the fusion and transformation, so it may affect the accuracy of the model.

| Method | N-PCA | $e_0$ | $e_1$ | Time (s) |
|--------|-------|-------|-------|---------|
| Original | 63 | 0.74 | 0.65 | 8.849 |
| SG | 41 | 0.95 | 0.79 | 7.146 |
| MSC | 84 | 0.77 | 0.79 | 9.965 |
| SNV | 84 | 0.80 | 0.79 | 10.540 |

3.3.2. SPA-SVM Camellia oil adulteration model

Compared with PCA, SPA's biggest difference is that SPA selects variables from the original data without changing the original data. SPA can fully find the variable group containing the minimum redundant information from the spectral information, so that the collinearity between the variables is minimized.

The key of the SPA algorithm is the choice of the initial iteration vector $x_{k(0)}$ and the number of variables $N$. Due to the existence of collinearity between variables, the value of $N$ cannot be greater than the number of samples in the training set, otherwise all the projections of the spectral values will become 0. Therefore, the value range of $x_{k(0)}$ is 1–J, the value range of $N$ is $11–M_{tra}$ (the number of samples in the training set), and the minimum value is 11 because there are 11 obvious absorptions in the spectrum peak. After determining the parameter selection of SPA, perform projection calculation on the original spectrum, save the cross-validation accuracy rate of each group of $x_{k(0)}$ and N SVM model, and determine the optimal $x_{k(0)}$ and $N$ values through comparison. As shown in Figure 4 and Figure 5, the optimal $x_{k(0)}$ of the original spectrum is 1746cm$^{-1}$ and $N$ is 30.
By comparing the SVM models combined with different preprocessing and SPA, the results are shown in Table 3. It is found that the model after SPA feature extraction on different preprocessing data has a distribution range of 0.66-0.79 for the classification accuracy of the test $e_1$. Compared with the PCA-SVM model and the SVM model, its accuracy is not improved, so it can be judged that SPA is not suitable for this research.

| Method  | $x_{k(0)}$ (cm$^{-1}$) | N  | $e_0$ | $e_1$ | Time (s) |
|---------|------------------------|----|-------|-------|----------|
| Original| 1746                   | 30 | 0.64  | 0.76  | 7.637    |
| SG      | 1981                   | 23 | 0.77  | 0.79  | 5.050    |
| MSC     | 1716                   | 64 | 0.72  | 0.72  | 7.906    |
| SNV     | 2068                   | 52 | 0.73  | 0.76  | 6.763    |

Figure 4 SPA Verification accuracy rate for different locations

Figure 5 SPA The verification accuracy rate corresponding to the number of variables N
3.3.3. RF-SVM Camellia oil adulteration model

RF is a classifier that contains multiple decision trees. It can efficiently process thousands of input attributes, and calculate the importance of different variable attributes in classification training, which is also the key to the use of RF as variable feature extraction. When RF is used for feature selection, it mainly compares the contribution of each feature on each decision tree in RF, and the Gini index (gini) is often used as a measure of contribution.

The following introduces the method of using gini as an evaluation: the importance score of the variable is represented by VIM, and the gini value is represented by GI, assuming there are m feature values x1, x2, ..., xc. Calculate the gini index score VIM of each feature xj, that is, the average of the impurity of the node split of the j-th feature in all decision trees in RF. The calculation formula of the gini index is:

\[
G_{I_m} = 1 - \sum_{k=1}^{K} p_{mk}^2
\]

Where K indicates that there are k categories, \( p_{mk} \) indicates the proportion of category k in node m.

The importance of feature xj at node m is calculated, that is, the change in gini index before and after the branch of node m is:

\[
VIM_{jm} = GI_m - GI_l - GI_r
\]

GI_l and GI_r respectively represent the gini index of the two new nodes after branching.

If the node of feature xj in decision tree i is in the set and M, then the importance of xj in the i-th tree is:

\[
VIM_{i}^{gini} = \sum_{j \in M} VIM_{jm}^{gini}
\]

Suppose there are n trees in the random forest, then:

\[
VIM_j^{gini} = \sum_{i=1}^{n} VIM_{i}^{gini}
\]

Finally, all the obtained importance scores are normalized to get the importance score:

\[
VIM_j = \frac{VIM_j}{\sum_{i=1}^{c} VIM_i}
\]

As shown in Figure 6, after the RF feature extraction method of the original data, the gini coefficient of each variable affects the heterogeneity of the observed value at each node, so as to compare the importance of each variable. The larger the value, the greater the value of the variable. The more important. After the importance score of the gini coefficient of each variable is obtained, the variable whose score is greater than the average of all the scores is selected as the new variable for the construction of the tea oil adulteration detection model.
Table 4 is the model evaluation after different preprocessing combined with RF-SVM. It is found that the test set accuracy $e_0$ of the RF-SVM model has a distribution range of 0.79-0.93, which is significantly improved compared with the PCA-SVM and SPA-SVM models. And the accuracy of the SG-RF-SVM model is the highest, reaching 0.93, which meets the requirements of tea oil adulteration detection.

Table 4  RF-SVM model cross-validation and test set accuracy under different preprocessing

| Method | e0   | e1   | Time (s) |
|--------|------|------|----------|
| Original | 0.79 | 0.79 | 21.437   |
| SG      | 0.93 | 0.93 | 17.387   |
| MSC     | 0.90 | 0.86 | 19.923   |
| SNV     | 0.89 | 0.86 | 22.628   |

4. Conclusion

In this paper, by using Fourier transform infrared spectroscopy, the FTIR spectral characteristics of tea oil, tea oil mixed with soybean oil and corn oil are analyzed, combined with stoichiometric methods, through experiments, a variety of tea oil mixed Pseudo rapid detection model, the specific conclusions are as follows:

1) The FTIR spectral absorbance curve trends of the three different types of camellia oil and adulterated oil are basically the same, and the wavenumber ranges of the main absorption peaks are between 600~800, 1000~1500, 1500~2000, 2500~3000cm$^{-1}$, and The appearance of the absorption peak is mainly caused by the vibration of the CH group of the unsaturated fatty acid in the edible oil.

2) By using the spectrum data obtained after different spectral preprocessing methods, SVM modeling is performed, and the model prediction results are compared through the training set, and it is concluded that the SVM model processed by MSC and SNV is the optimal model, and its accuracy rate is 0.86.

3) Different feature extraction algorithms are used, and some spectral data are selected to improve the model. By comparing PCA-SVM, SPA-SVM and RF-SVM models, the prediction effects of the camellia oil detection model preprocessed by SG, MSC and SNV on the test set were compared, and the optimal model was convolutional smoothing and random forest. Combined with support vector machine (SG-RF-SVM), the 10-fold cross-validation accuracy $e_1$ of the training set and the test set accuracy $e_0$ are 0.94 and 0.93, respectively, which meets the detection classification requirements and can provide a way for tea oil adulteration Fast detection method.
References

[1] Li Qiuting, Lu Shunzhong. A promising health-care edible oil-tea oil [J]. Guangxi Forestry Science, 2003, 32(3): 155-158.
[2] Li Zhuoxin. Study on the determination of peanut oil adulteration by gas chromatography [J]. Grain Storage, 2001, 30(3): 41-43.
[3] Kapoulas VM, Andrikopoulos NK. Detection of olive oil adulteration with linoleic acid-rich oils by reversed-phase high-performance liquid chromatographyI [J]. Chromatogr, 1986, (366): 311-320.
[4] Merchak N, Rizk K, Silvestare V. Olive oil characterization and classification by 13C NMR with a polarization transfer technique: a comparison with gas chromatography and IH NMR [J]. Food Chemistry, 2018, 245: 717-723.
[5] Jin Qingzhe, Xie Feng, Ding Zhihua, et al. Study on the detection of peanut oil and corn oil blend by carbon isotope ratio mass spectrometry [J]. Journal of the Chinese Cereals and Oils Association, 2010(05): 95-99.
[6] Huang Biao, Liu Wenjing, Fu Jianwei, et al. Simultaneous determination of 7 nicotine pesticide residues in tea oil by QuEChERS pretreatment method-ultra performance liquid chromatography tandem mass spectrometry [J]. Quality and Safety of Agricultural Products, 2019(01): 45-48.
[7] Wang Xiaoling, Wu Jing, Tan Mingqian. Low-field nuclear magnetic resonance combined with chemometric methods for rapid detection of adulterated walnut oil [J]. Journal of Analysis and Testing, 2015 (07): 789-794.
[8] Chen Yi, Gan Bei, Li Min, etc. Research progress in the application of NMR technology in the rapid detection of olive oil quality [J]. Journal of Food Safety and Quality Inspection, 2014(10): 3030-3034.
[9] Wang Xiaoling, Wu Jing, Tan Mingqian. Low-field nuclear magnetic resonance combined with chemometric methods for rapid detection of adulterated walnut oil [J]. Journal of Analysis and Testing, 2015 (07): 789-794.
[10] Li Wenlong, Zhai Haibin. Technical standardization research progress of NIR spectroscopy for important quantitative analysis [J]. Chinese Journal of Traditional Chinese Medicine, 2016, 41(19): 3511-3514.
[11] Ma Wei, Wang Xiu, Li Weiguo. Application of Near Infrared Spectroscopy (NIRs) Detection Technology in Facility Fruit and Vegetable Production [J]. Agricultural Engineering Technology, 2015(31): 60-63.
[12] Zhu Yutian, Li Jincai, Gao Sujun, et al. Research progress of near infrared spectroscopy in the field of rapid detection of edible oil [J]. China Oils and Fats, 2017(07): 140-143.
[13] Zeng Lulu, Tu Bin, Yin Cheng, etc. A qualitative and quantitative analysis of peanut oil adulteration based on Aurora NIR ensemble SVM [J]. Journal of the Chinese Cereals and Oils Association, 2016, 31(08): 126-130+137.
[14] JiaoJiao Yuan, Chengzhang Wang, Hongxia Chen. Study on the quantitative analysis of camellia oil adulterated with soybean oil by near infrared transmittance spectroscopy (NITS) [J]. Journal of the Chinese Cereals and Oils Association, 2012, 27(3): 110-114.
[15] Su Donglin, Zhang Juhua, Li Gaoyang, et al. Research progress in the application of near-infrared spectroscopy combined with chemometrics in the quality detection of tea seed blended oil [J]. Chinese Journal of Food Science, 2018, 18(7): 332-338.
[16] Su Donglin, Zhang Juhua, Li Gaoyang, et al. Research progress in the application of near-infrared spectroscopy combined with chemometrics in the quality detection of tea seed blended oil [J]. Chinese Journal of Food Science, 2018, 18(7): 332-338.
[17] Wang Y, Yezhou C. Identification and detection of adulterated Camellia oleifera Abel. Oils by near infrared transmittance spectroscopy[J]. International Journal of Food Properties, 2016, 19(2): 22.
[18] Li S, Zhu X, Zhang J, et al. Authentication of pure camellia oil by using near infrared spectroscopy and pattern recognition techniques[J]. Journal of Food Science, 2012, 77(4): 374-380.

[19] Liu Lingling, Wu Yanwen, Zhang Xu. Fourier transform infrared spectroscopy combined with pattern recognition method to quickly identify the authenticity of edible oil [J]. Acta Chimica Sinica, 2012, 70(8): 995-1000.

[20] Zhu Qisi, Zhu Liqiong, Zhong Guocai. Fourier transform infrared spectroscopy to quickly identify adulteration of camellia seed oil [J]. Food Science and Technology and Economy, 2015, 40(3): 37－39.