Near Infrared Spectroscopy Analysis Based on Support Vector Machine

Min LI[1,a], Linju LU[1,b], Jin CAO[1,c]

[1]School of Physics & Electrical Engineer of Leshan Normal University, Leshan 614000,China

[a]cassie_li@163.com, [b]171139040@qq.com, [c]372923834@qq.com

Keywords: Near Infrared Spectroscopy, Multiplicative Scatter Correction, Linear Discriminant Analysis, Support Vector Machine.

Abstract. This paper put forward a kind of qualitative identification method of tea authenticity based on near infrared spectroscopy(NIR). Authentic Zhuyeqing tea and fake Zhuyeqing tea were the research objects. Multiplicative Scatter Correction (MSC) was used to NIR data of 2 kinds of Zhuyeqing tea as a pre-processing. Principal Component Analysis (PCA) was then used to spectral data for dimensionality reduction and redundant removal. Next Linear Discriminant Analysis (LDA) was used for further feature extraction. Finally Support Vector Machine (SVM) was run for identification. Rradial basis was chosen as the support vector kernel function. On the condition of $C = 100, \delta = 0.5$, modeling recognition effect is the best of 97%. Experiments show that the algorithm can effectively identify 2 kinds of Zhuyeqing tea.

1.Introduction

Leshan city is located in the Sichuan province of China West. It has produced much high-quality Green Tea. It has a long history of tea production. Now, it has formed a lot of famous brands. Zhuyeqing tea is one of them. Its mellow and sweet taste renowned chinese and foreign consumers. Accordingly, the price of it is higher than that of others. But because of the lackness of effective tea identification method, there is much fake Zhuyeqing tea on chinese tea market, which harms the interests of consumers and also damages the market reputation of Leshan Zhuyeqing tea seriously. In order to protect the famous tea brands of Leshan city and seize the tea market, an effective tea identification method is urgently needed.

The traditional identification methods of tea are mainly two types: chemical methods and sensory evaluation methods[1]. Although the chemical methods can correctly identify the varieties of tea, but the process is complex and the price is expensive. Sensory identification methods are influenced by the factors of external environment and human interference. Their correct identification rates are low. Near infrared spectroscopy analysis is a new analysis and research method. It has the advantages of fast analysis, more output and without destroying the samples. Its online analysis advantage is suitable for products. It has got the wide application in food, medicine, agriculture, petrochemical and other fields[2]. This research takes Leshan authentic Zhuyeqing tea and fake Zhuyeqing tea as the objects, trying to combine near infrared spectroscopy analysis with pattern recognition methods, putting forward an effective method for identification authenticity of Zhuyeqing tea.

2.Tea Spectrum Acquisition

The experimental instrument is the FTIR-7600 fourier transform infrared spectrometer. Wave number is at the range of 7800~350cm-1. Resolution is 4cm-1. Scanning times are 32. Data point interval is 1.928cm-1. Experimental materials are authentic Zhuyeqing from Zhuyeqing company of Leshan city Mount Emei and fake Zhuyeqing from market of Leshan city. Two kinds of tea were crushed, and then filtered by 40 mesh sieve. Tea powder and potassium bromide were mixed by 1:100. Every 1g mixture is as an sample. Every sample was scanned 3 times, taking the average value as sample data. The collection environmental temperature is 25.2oC and relative temperature is 49%. Voltage is 220V. Each kind of tea collected 32 samples. Thus a total of 64 samples were received.

Each sample data is 1868 dimensional. Wave number range is 4001.569~401.1211cm-1. Samples are shown in table 1.

Table.1 The experimental samples

The Types of Tea	Total Samples	The Training Set	The Testing Set
authentic Zhuyeqing	32	15	17
fake Zhuyeqing	32	15	17

3.The Principle of Support Vector Machine

3.1 linear separable problem

Support vector machine theory originates from two kinds of linear separable data processing. As X is the input space and Y is the output space, pattern set $X = \{x_i\} \in R^n$ is usually composed of two types, namely $Y = \{-1,1\}$. The training set composed of n samples is as bellow:

$S = [(x^1, y^1),...,(x^n, y^n)] \subseteq (X \times Y)^n$

According to the principle of structural risk minimization, an objective function is construct to find a separating hyperplane meeting the requirements , and make the distance between training set points and separating hyper plane as far as possible. As shown in Fig.1[6], the solid circle and hollow circular respectively represent two types of samples. H is the optimal classification hyper plane. H_1 and H_2 are across various samples. They are the nearest from the classification hyperplane and parallel to the classification hyperplane. The classification hyperplane is denoted as Eq. (1). After normalized, the current separating sample set (x_i, y_i) should satisfy Eq(2). At this time, the training samples of H_1, H_2 become support vectors.

$$\omega \cdot x_i + b = 0 \tag{1}$$

$$y_i(< \omega \cdot x > +b) - 1 \geq 0 \tag{2}$$

In above, ω is the super normal vector of the classification hyperplane. Margin is equal to $2/\|\omega\|$ as the regional distance.

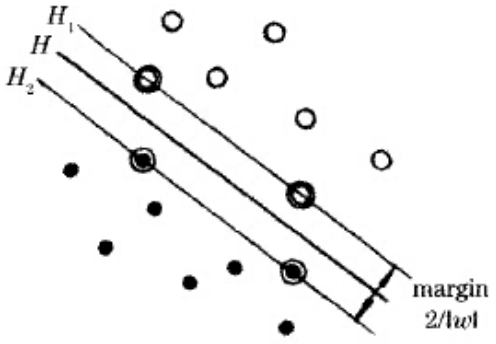

Fig1. Optimal plane of linearly separable samples

3.2 nonlinear separable problems

When the problem is nonlinear separable, nonlinear transform can be realized by using the kernel function, which changes a nonlinear separable problem into a linear one in a high dimensional space. and seeks the optimal classification plane in this high dimensional space. Of course, classification results are different while using different kernel function. At present , the commonly used kernel functions are as bellow[4]:

1) polynomial kernel function:

$$K(x_i, x) = [(x_i \cdot x) + 1]^{\delta} \tag{3}$$

2) the radial basis kernel function:

$$K(x_i, x) = \exp\left[-\frac{\|x - x_i\|}{2\delta^2}\right] \quad (4)$$

3) Sigmoid kernel function:

$$K(x_i, x) = \tanh[\delta(x_i \cdot x) + \gamma] \quad (5)$$

3.3 the Determination of kernel function and parameters

The priority problem is the selection of kernel function, while using support vector machine to establish an identify model. The different choice of kernel functions can make great influence to the properties of the model. On the condition of the absence of a priori knowledge guide, the radial basis kernel function can often get better fitting results. Radial basis can map the nonlinear sample data into a high dimensional feature space, dealing with nonlinear sample data. The value of radial basis ($0 < K \le 1$) is more simple than that of the polynomial value ($0 < K$ or $1 < k < \infty$). Sigmoid kernel function calculation speed is very slow. This experiment uses radial basis as the support vector kernel function. The main parameters to be determined are the penalty coefficient C and the width of kernel function δ .There is no mature method for these parameters choices. They are usually determined by repeated experiments. Through experiments, on the condition of $C = 100, \delta = 0.5$, modeling recognition effect is the best.

4. The Experimental Analysis

The original spectrum of two kinds of tea is shown in fig.2. Mmultiplicative Scatter Correction(MSC) pretreatment was introduced. After MSC pretreatment, the difference of tea NIR will be greatly reduced, and at the same time, random variation will get the maximum deduction. That laid a good foundation for subsequent classification. The spectra after pretreatment is shown in Fig.3.

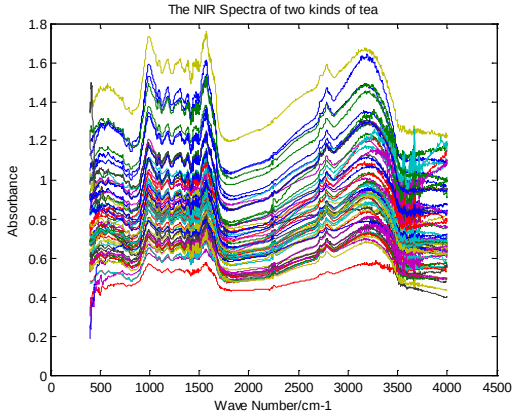

Fig.2 the NIR Spectrun of 2 kinds of Zhuyeqing

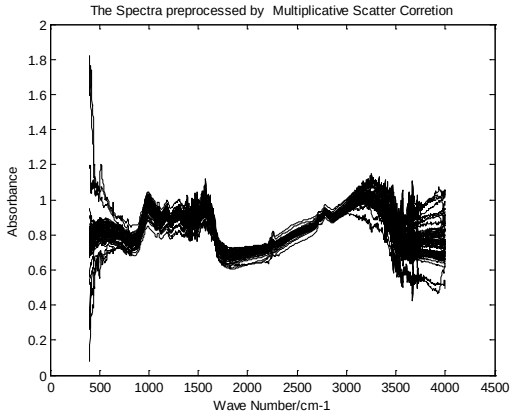

Fig.3 the spectrum after MSC preprocessing

After pretreatment, the principal component analysis(PCA) was introduced to reduce dimension of sample data. The data after PCA also needs further analysis, such as discriminant analysis and cluster analysis[5]. The total 64 sample data of 2 kinds of tea processed by PCA is reduced to 20 dimensional, obtaining 20 principal components. The first two principal components are shown in figure 4. Where "Tea1" is authentic Zhuyeqing, "Tea2" is fake Zhuyeqing. Tea NIR spectra after PCA reducing dimension, then the Linear Discriminant Analysis (LDA) was used for effective feature extraction. At last ,SVM was introduced to identify real or fake Zhuyeqing tea. Rradial basis was chosen as the support vector kernel function. On the condition of $C = 100, \delta = 0.5$, modeling recognition effect is the best. The correct recognition rate is 97%. The identification result of two kinds of tea is shown in figure 5.

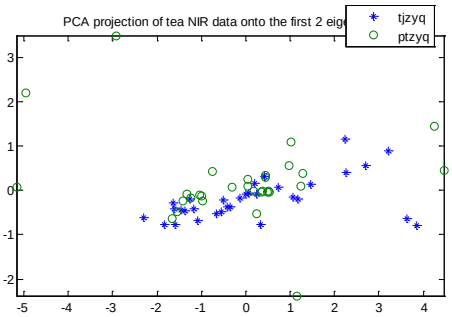

Fig. 4 the first two principal components of two types of tea

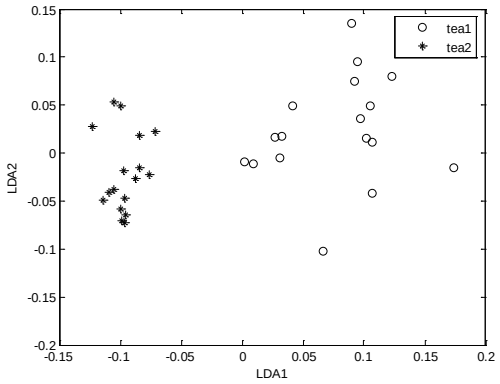

Fig.5 identification results of two kinds of tea

5. Conclusions

Taking authentic Zhuyeqing tea produced in Leshan and fake Zhuyeqing as experimental materials, this paper combined with near infrared spectroscopy and pattern recognition methods to identify two kinds of tea. Step 1: spectral data was pretreated with MSC; Step 2: PCA was run to remove redundancy;Step3: LDA was introduced for feature extraction; Step 5: SVM identification was run. Through the experimental analysis, this method can identify the authenticity and effective for Zhuyeqing tea. The correct recognition rate is 97. The study provides a new idea for the study on qualitative identification of authenticity of tea.

Acknowledgements

This work was supported by Science and Technology Bureau of Leshan city（No.14NZD017）.

References

[1] XIE Li juan, YING Yi bin , YING Tie-jin, et al. Spectros copy and Spectral Analysis ,2008, 28(5):1062.
[2] Williams P C. Cereal Chemistry,1975,52:561-576.
[3] ZHAO Jiewen, CHEN Quansheng, HUANG Xingyi, et al. Journal of Pharmaceutical and Biomedical
Analysis, 2006, 41(4): 1198-1204.
[4] Chen Quansheng, Zhao Jiewen, Zhang Haidong,et al， Acta Optica Sinica,2006, 26(6):933-937.
[5] LI Qiong-fei, YANG Zeng-ling, HAN Lu-jia. Journal of Infrared Millimeter Waves ,2007, 26(6) : 414-418.
[6] CHEN Quansheng, ZHAO Jiewen, FANG C H, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy,2007, 66(3): 568-574.