Study on Quick Identify of the Brand of Seabuckthorn Juice Based on PCA and SVM

Zhipeng Liu **Shujuan Zhang**
College of Engineering
Shan xi Agricultural University
Taigu, China
sxaulzp@126.com zsujuan@263.net

Abstract. To achieve the Seabuckthorn juice brand non-destructively, we put forward a fast identify method based on Visible_near infrared reflectance (NIR) spectroscopy . We use the Field Spec 3 spectroradiometer to collect 40 sample spectra data of the three kinds of Seabuckthorn juice separately. the sample data was preprocessed by Using of average smoothing method and multiplicative scatter correction (MSC) method. Then principal component analysis (PCA) was used to process the spectral data after pretreatment. The samples were divided into 90 model samples and 30 prediction samples, the sample of eight modeling data as input variables of the support vector machine (SVM) to build SVM model, and to identification juice brands. 30 unknown samples of the three brands were predicted for classification, and the results showed that the SVM model on the identification seabuckthorn juice brand has achieved a 99.9% accuracy. Therefore, near infrared spectroscopy coupled with principal component analysis and SVM can be quickly and accurately distinguish the brand of seabuckthorn juice.

KeyWords: Visible_NIR spectroscopy; Principal component analysis(PCA); Support Vector Machine(SVM); Seabuckthorn juice brand

1 Introduction

Seabuckthorn (Hippophae Rhamnoides Linn.),also known as vinegar Liu, is belong to deciduous shrubs or small trees[1].In China, distribution of genus is the largest and the species is the most.It is Distributed in Shanxi, Shaanxi,and Inner Mongolia, 19 provinces and autonomous regions,the total area is 1800 Ten thousand mu. Seabuckthorn have strong adaptability and vitality,and is resistant to Sand wind, drought, water wet, salt, poor soil, heat and cold. Since the mid-80s of the 20th century,China has become to focus on comprehensive development and processing and utilization about seabuckthorn,the number of research datum and works about Seabuckthor is increasing[2].In the recent years,With the implement of the grain project, seabuckthorn is cultivated large areas. Seabuckthorn fruit contains protein,

Research fund for the doctoral program of higher education.(20101403110003)

unsaturated fatty acids, vitamins for the whole,and is the valuable medicinal and edible plant resources.Currently, seabuckthorn juice is as a green drink,its internal

quality such as taste, sugar content, acidity and vitamin content has received widespread attention.The Brand range is large,and the differences in taste and quality are more significant.So a simple, rapid and nondestructive identification of juice brand and technology are essential.

Near infrared spectroscopy (NIRS) technology developed rapidly in the late 80's last century can achieve non-destructive testing techniques of physical test. Almost the main structure of all the organic composition can be found the signals in near-infrared spectroscopy, the table and obtained easily spectrum,unpollution testing,low-cost detection, so Near Infrared Spectroscopy (NIR) has the reputation of the Giants[3].In the recent years,the application of near infrared spectroscopy technology in many areas has obtained a great development,it has played a significant role in the advancement of scientific research in the field of production and technological progress.Nowdays, using near-infrared spectroscopy,domestic and foreign scholars have identified the brands such as soy sauce[4], bayberry juice[5], soybean[6], white vinegar[7], yogurt[8], milk powder[9], coffee[10], tea[11], melon[12], chilli[13], but no identification on seabuckthorn juice brand.

PCA is a data mining techniques in the multivariate statistic,and is one of the means of cluster analysis.The aim is to select a smaller number of new variables to replace the original more variables under the premise of not losing the main spectral information,it resolve the difficulties which is not be analyzed because of the overlapping bands,and is widely used as a spectrum analysis mathematical method[14]. SVM (Support Vector Machine) approach is based on statistical theory, Huber robust regression and Wolfe duality theory[15][16],its advantage is strong generalization performance,high precision fitting, global optimization. To address prediction and identification in the case of the limited sample,it provided a strong theoretical basis and effective solution. SVM method is based on structural risk minimization principle rather than empirical risk minimization principle,it avoid learning phenomenon in the artificial neural networ, has very good generalization performance, can find the global optimal solution,and has many advantages when solving the small sample, nonlinear, high dimension and local minima problems in identifying the performance of such models[17][18].

The ultimate goal of this study is to realize the identification on juice brand,using data mining method combining the visible - near infrared spectroscopy,principal component analysis and support vector machine.

2 Experiment Material

2.1 Experimental Device

The experimental equipment includes notebook computer, spectrometer, halogen light,and correction whiteboard. FieldSpec 3 spectrometer is produced by ASD （Analytical Spectral Device） in America, spectral sampling interval of 1nm, the sampling range of 350～2500nm, scanning frequency of 30 times, Probe field angle of 25 °, diffuse reflectance sampling of spectrum, 14.5V halogen light source

matching spectrometer. Spectral data is exported in the form of ASCII code for processing, analysis software includes ASD View Spec Pro V5.0，Unscramble V9.7 and DPS（Data Procession System For Practical Statistics）.

2.2 source of Experimental Samples

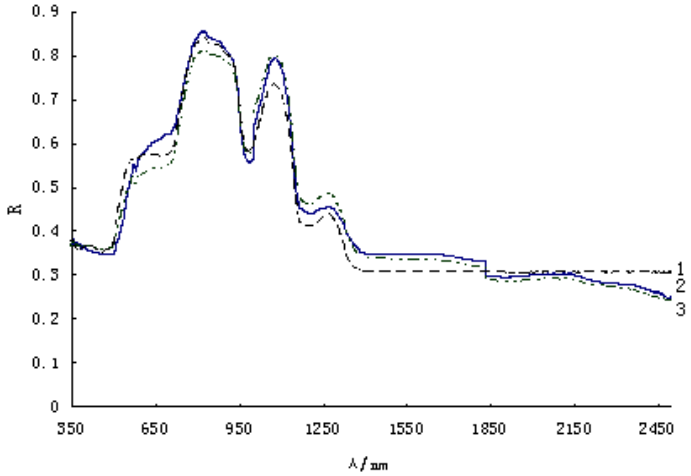
Three brands of juice with the same date with the batch production purchased from the market,totaling 120 samples, were randomly divided set of 90 samples for modeling and set of 30 samples for prediction, Xiapusaier sea buckthorn juice, sea buckthorn juice and Wei Shijie Huiyuan juice respectively 40 samples. The juice is filled into the high-1.4cm, 6.5cm diameter of dish, the liquid height 10mm. Then spectrometer sample probe is placed in the top of sea buckthorn juice, sea buckthorn juice from the surface of 50mm, and the surface into a 45 °angle fixed. Sheng-like container is away from light source center 350mm, and the surface into a 45 °angle of illumination.

2.3 Spectral data pretreatment

In order to remove the effects from the high-frequency random noise, baseline drift, uneven samples, light scattering,we apply smoothing method for spectrum pretreatment, smooth points to 9.All of the pretreatment process is carried out in the Unscramble V9.7 software for greatly filtering the high-frequency noise generated by various factors, further multiplicative scatter correction MSC(Multiplicative Scatter Correction) processing[11].

3 Results and Analyses

3.1 Spectrum curve analysis



Note:1-Xiapusaier juice spectral curve; 2-Huiyuan juice spectral curve; 3-Wei Shijie juice spectral curve

Fig.1. Visible_Near infrared reflectance spectroscopy of three different varieties of Seabuckthorn juice

The typical curve of near-infrared diffuse reflectance spectra of three varieties of sea buckthorn juic is shown in Figure 1. Horizontal axis in Figure 1 for the wavelength range of 350～2500nm, the vertical axis for the spectral reflectance. Fig.1 shows various species of sea buckthorn juice is different spectrum, with some characteristics and fingerprints. Through application of ASD ViewSpec Pro V5.0 software, spectral reflectance data of samples averaged is measured, and converted into ASCII code to export, then through Unscramble V9.7 software for principal component analysis.

3.2 main component qualitative cluster analysis

The purpose of PCA is data reduction, to eliminate the information overlapping part of the co-exist information,by converting lots of original spectral variables to change a smaller number of new variables into a linear combination of original variables, and to make new variables maximize the representation of data structure of the original variable[4][5]. Using Unscramble V9.7 software,we can obtain the PCA clustering of three juice 60 samples(Fig.2,A1:Huiyuan juice A2:Xiapusaier juice A3: Wei Shijie juice).

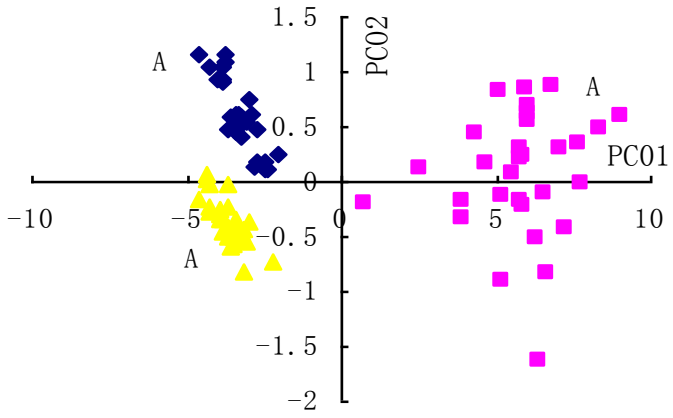


Fig.2. PCA scores plots (PC1 × PC2) for Seabuckthorn juice sample across the entire spectral region

Fig.2 shows PCA scores plots (PC1 × PC2) for seabuckthorn juice sample across the entire spectral region, Horizontal axis represents each sample of the first principal component scores, the vertical axis represents each sample value of the second principal component scores. "Xiapusaier juice", "Wei Shijie juice" and "Huiyuan juice" are clearly divided into three categories(Fig.2),to show the main components 1,2 well effect the three seabuckthorn juice.Fig.2 also shows 20 Huiyuan juice samples with good degree of polymerization is better distributed in the second and third quadrant and aggregate near the X-axis.The degree of polymerization of 20 Wei Shijie juice samples is better,and is mostly distributed in the aggregate in the first quadrant and near the Y axis. The degree of polymerization of Xiapusaier juice is inferior than the first two samples, but also is distributed in the fourth quadrant.The three seabuckthorn juice is little overlaped and can be distinguish. Analysis showed that the first two principal components has some clustering effect on the three kinds of juice, can qualitatively distinguish between different brands of juice, but can not give quantitative identification model.so we consider the establishment of brand identification model based on a number of principal components and support vector machine.

3.3 The principal component analysis results

Sample spectral band from 350～2500nm total 2150 points,when using the whole spectrum, calculation amount is large,because of weak spectral information of some samples of the region, and lacking of correlation with sample composition or its nature.Table 1 shows the accumulative reliabilities of eight principal components.The accumulative reliabilities shows the explaination extent of the principal component to the original variables.The accumulative reliabilities of the first eight principal components is to 99.99%,it shows that the eight principal components can explain 99.99% of the original wavelength.

Table 1 shows that the principal component analysis is a very effective data mining method, which has compressed more than 2000 wavelength variable into 8 new variable each other perpendicular, the eight new variables do not affect each other, but also are on behalf of the information contained in the most of original variables.

Table 1.Accumulative reliabilities of the first 8 PCs

No.of PC	PC01	PC02	PC03	PC04	PC05	PC06	PC07	PC08
Accumulative reliability /%	97.997	99.315	99.720	99.877	99.965	99.981	99.986	99.992

3.4 support vector machine model

The first thing is to resolve the kernel function selection,when applicating the identification model of support vector machine,since the effection of predictive power to sample data is different according to applying different kernel function in SVW algorithm. The most common is the radial basis function.In the paper,considering input and output is highly nonlinear in the most forecasts, we select the exponential radial basis function as the SVM kernel function within the product.

The common kernel functions are as follows:

1) Polynomial kernel function:

$$K(x, x_i) = (x \cdot x_i + 1)^d \tag{1}$$

2) Gaussian kernel function:

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

3) Sigmoid function:

$$K(x, x_i) = \tanh(v < x, x_i > + c) \tag{3}$$

We make the principal component score of 90 (30 per brand) samples as the training set of the support vector machine algorithm,while the remaining 30 (10 for each brand) sample principal component score as the prediction set of the support vector machine,and classifying juice brand with a class on the remaining class support vector machine method. A class on the remaining class support vector machine algorithm is as follows.

We suppose the training set of known samples (Dn) contains n samples,

$$D = \{(x_i, y_i) | i = 1,2,\cdots,n\} \in (X \times Y)^n, x_i \in X = R^n, y_i \in Y = \{1,\cdots,M\}, i = 1,\cdots,n \tag{4}$$

The choice of kernel function K(x,xi)determines the structure of feature space, Kernel function value is equal to the inner product of the vector X and the vector Xi in the respective feature space $\Phi(x)$ and $\Phi(x_i)$.

Radial basis function:

$$K(x, x_i) = \exp\left[-\frac{\|x - x_i\|^2}{2\sigma^2}\right] \quad (\sigma \text{ is the parameter}). \tag{5}$$

So we select the appropriate parameters σ according to each output indicator,then fit the sample by SVM, establish the appropriate fitting model SVM_i ($i = 1,2,\cdots,30$) for respectively fitting y_i ($i = 1,2,\cdots,30$) in table 2.

Table 2 .SVM model predictions for unknown samples

Note: Variety value 1-Huiyuan juice; 2-Xiapusaier juice; 3-Wei Shijie juice

Sample number	Real value	Predicted value	Bias	Sample number	Real value	Predicted value	Bias
1	1	0.99319	0.00681	16	2	1.95086	0.04914
2	1	1.00822	-0.00822	17	2	1.99182	0.00818
3	1	0.99265	0.00735	18	2	1.98151	0.01849
4	1	0.99086	0.00914	19	2	1.98210	0.01790
5	1	0.99323	0.00677	20	2	1.99062	0.00938
6	1	0.99323	0.00677	21	3	3.01081	-0.01081
7	1	1.00735	-0.00735	22	3	2.88968	0.11032
8	1	0.99332	0.00668	23	3	2.99442	0.00558
9	1	0.99134	0.00866	24	3	2.95287	0.04713
10	1	0.982303	0.017697	25	3	2.96318	0.03682
11	2	2.01765	-0.01765	26	3	2.97322	0.02678
12	2	1.97105	0.02895	27	3	2.99949	0.00051
13	2	1.98735	0.01265	28	3	2.95702	0.04298
14	2	2.00447	-0.00447	29	3	3.01189	-0.01189
15	2	1.96037	0.03963	30	3	3.00159	-0.00159

Table 2 shows the support vector machine model predicts unknown samples of sea buckthorn juice with below ±0.1 deviation and 100% recognition ratio, this study establishes the SVM near infrared spectroscopy classification model with more accurate identification, compared with the conventional model of discriminant analysis.

4. Conclusion

In this paper,on the base of the spectral data acquisition and analysis of the three kinds of juice on the diffuse reflectance,we establishes brand juice identification model, identifying three seabuckthorn juice brands by combining the principal component analysis with the support vector machine method.The experiments and calculations indicate that the model predicted good results,with below ±0.1 prediction deviation of the unknown juice sample and the identification rate of 100%, the near-infrared spectral identification model established in the paper is more accurate than the conventional model of discriminant analysis. It is suggested that the support vector machines has a good application performance in the near infrared spectroscopy juice brand identification.

References

[1]. Jvyun Ma. Progress in the chemical constituents of sea buckthorn[J]. Heilongjiang Traditional Chinese Medicine,2001,14(3)：208-209

[2]. Fan Li. Seabuckthorn Research[J]. Research Advances,2009,6(1):7

[3]. McClare W F Anal.Chem.A[J],1994,66(1):43

[4]. Xiaoxing Tong,Yidan Bao,Yong He. Application of near infrared spectroscopy technique for detection of the brand of soy sauce[J]. Spectroscopy and Spectral Analysis, 2008,28(3):597-601

[5]. Haiyan Cen, Yidan Bao,Yong He. Bayberry juice based on spectroscopy of rapid identification method for species[J]. Spectroscopy and Spectral Analysis, 2007,27(3):503-506

[6]. Turza S,Tot h A ,Varadi M.Near Inf rared Spect roscopy : Proceedings of the International Conference ,Chichester , U K: NIR Publi2 cations ,1998.183

[7]. Li Wang,Fei Liu. Application of visible - near infrared spectroscopy technology white vinegar brands and rapid detection of pH values[J]. Spectroscopy and Spectral Analysis, 2008,28(4)

[8]. Yong He,Shuijuan Feng,Xiaoli Li. Application of Near Infrared Spectroscopy Identification of species of yogurt[J]. Spectroscopy and Spectral Analysis, 2006，26(11)：2021

[9]. Min huang,Yong He,Haiyan cen. Application of visible - near infrared spectroscopy Fast Discrimination of varieties of infant milk powder[J]. Spectroscopy and Spectral Analysis, 2OO7，27(5)：916

[10]. Yanyan Wang,Yong He. Based on Visible - Near Infrared Identification of the coffee brand[j]. Spectroscopy and Spectral Analysis,2007,27 (4)

[11]. Xiaoli Li,Yong He,Zhengjun Qiu. A visible - near Infrared Spectroscopy new method for identification of tea varieties[j]. Spectroscopy and Spectral Analysis, 2007,27(2):279-282

[12]. Zsolt Seregely , Tamas Deak , Gyorgy Denes Biszt ray. Chemomet rics and Intelligent Laboratory Systems , 2004 , 72 : 195

[13]. Wei Yu,Kelang yong. Fluorescence characteristics of capsaicin[J]. Food Science, 2003,24(11) : 105

[14]. Li wang,Fei Liu. Application of visible - near infrared spectra for white vinegar brands and rapid detection of pH values[J]. Spectroscopy and Spectral Analysis, 2008,,28(40):813

[15]. Qingan Cui,Zhan He,Fuxin Cui. Support Vector Machine Prediction Model of Coke Quality[J]. Chemical Engineering, 2006, 1(1): 28- 31

[16]. Xiaofang Du,Jinlong Zhang. Sales of agricultural products based on support vector machine[J]. Management Science, 2005, 4(8): 129- 134

[17]. Naiyang Deng,Yingjie Tian. A new method of data mining towels a support vector machine. Beijing: Science Press,2004.6

[18]. guozheng Li,Meng wang,Huajun Zeng. Introduction to the shoe support machine. Beijing: Electronic Industry Press, 2004.3