Apple NIR Spectral Classification Method

Min Li [a], Jin Cao, Linju Lu

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

[a] cassiei@163.com

Keywords: Apple, Near Infrared Spectroscopy, Principal Component Analysis, Fisher Decision Analysis, K- Nearest Neighbor Classification

Abstract. This paper proposed an apple near infrared spectral classification method. Red Fuji apples from Shandong and Shaanxi province , “Huaniu” apples from Gansu province were used as experimental materials. NIR data of three kinds of apples after preprocessing by wavelet soft threshold, was removed the noise and redundancy. Then the method of Principal Component Analysis (PCA) was used to reduce the dimension, and the Fisher Decision Analysis (FDA) was used for further feature extraction. Finally the K-Nearest Neighbor (KNN) classification was run, and K = 4. Through the experimental comparison, the method can achieve good feature extraction and classification of apples. The correct identification rate reaches above 96%. This method can realize different kinds of apples nondestructively, rapidly and accurately, which provides a new idea for near infrared spectral analysis technology.

Introduction

China is a big country of apple production. Apple quality influenced by the variety, origin, cultivation methods and climate, exists difference levels. The apple market exists the uneven in quality and cultivar identification difficult phenomenon, which seriously impact on China's export of apple. Currently on the market for Apple classification is mostly manual classification by human perception, time-consuming and laborious, but also can produce damage to apple. It is therefore an urgent need to develop a fast, accurate and non-destructive apple classification technology.

Near infrared spectroscopy analysis technique is non-destructive, fast and efficient detection, low cost. It has been developing quickly in the field of agricultural and food detection. For example, it has been applied to blueberry nutrient composition detection, apple bruise detection, apricot fruit quality detection and peach maturity detection etc. Some scholars use near infrared spectroscopy to identify the varieties of yam[1], pepper [2], and rhodiola[3]. However, the correct recognition rate is not high, such as Sun Suqin, Tang Junming and others to identify yam, the correct recognition rate is only for 70%[1].

Near infrared spectroscopy for fruit feature extraction methods are mainly partial least squares method(PLS), wavelet transform, PCA analysis, Fisher discriminant analysis(FDA). Near infrared spectroscopy for fruit classification methods are mainly KNN, support vector machine, and artificial neural network classification algorithm. PCA is an unsupervised learning method and a common method for feature extraction and dimensionality reduction of data. FDA is also called linear discriminant analysis. It is an effective feature extraction method. KNN is a classification method based on statistics. The algorithm carries on the classification according to the samples to be tested in the feature space K nearest neighbors in a sample of the majority of the classes of the samples, which has the characteristics of intuition, unsupervised learning, without prior knowledge of statistics and so on. It is an important non parametric classification method. This paper discusses three classification methods of three different sources of apples. The first method is PCA+KNN; The second method is FDA+KNN; The third methods uses PCA+FDA+KNN. Through the experiment, the third method is the best. It’s correct recognition rate is the highest of 96%, when K=4.

Spectral Data Acquisition

Experiment selects three kinds of uniform size, no damage apples from Shandong super red Fuji (represented here by "hfsd"), Shanxi producing ordinary red Fuji (represented here by "hfsx") and Gansu production “huaniu” apples (represented here by "hn"). Experiment chooses 60 components of each kind as experimental samples. The samples has been kept in the laboratory at room temperature for 20~25 ℃ for 12 hours to be measured.

Spectral collector uses Antaris II near infrared spectroscopy analyzer of Thermo Fisher company. The instrument has the maximum performance of near infrared laboratory requirements. It can save a lot of testing cost.

While spectral acquisition, the near infrared spectroscopy need to be preheat for 1 hour. Experiment uses reflection integral ball collecting mode to collect the near infrared spectroscopy of these three kinds of apple. Spectra wave number is 4000~10000cm-1. Scanning interval is 3.856cm-1. In order to reduce the error, near infrared spectroscopy analyzer scan each sample 3 times track along equatorial, taking the average as the final test data. Spectra of each sample is a 1557 dimensional data. Three kinds of Apple eventually get the 180 spectrum data of 1557 dimensions. The original spectrum is shown as Fig.1.

Spectral Data Preprocessing

Near infrared spectrum data contains inevitable noise signal from the high frequency random noise, baseline drift, light scattering, and non-uniform samples[4]. The noise signal may interfere with spectrum, effect correction model, thus reducing the accuracy of sample analysis. On the other hand, the near infrared spectrum data is compared commonly big, needs much storage space, and the modeling time is relatively long [4]. So in order to classify the samples by near infrared spectroscopy, spectral preprocessing is needed. Preprocessing mainly accomplishes two tasks: one is the noise reduction; the other is the data compression. In this paper, the wavelet soft threshold method was used for spectral data pretreatment. It not only can achieve effective denoising and data compression, but also can keep the details of spectral data as much as possible. Pretreatment effect of wavelet soft threshold is better than that of SavitZky-Golay smoothing method and MSC method.

The Classification Method

PCA is a kind statistical method of comprehensive multiple indicators into some indexes. It project a multidimensional spectral data space to a low dimensional data space along the direction of maximum covariance. So that reaches the aim of data dimension reduction. Vectors between the different principal component are orthogonal to each other. Through the reasonable choice of the principal component vectors, we can avoid redundant data modeling, and have little loss of spectral information [5].

FDA is one of the basic method of statistical pattern recognition. The basic idea is to project the original high-dimensional samples to the optimal discriminant vector space, in order to achieve the classification of the information and compress feature space dimension. After the projection, samples had the maximum distance among different classes and minimum distance with a same class in the new subspace. The key of this method is to solve the optimal discriminant vectors. Therefore, spectral data needs PCA first for dimensionality reduction. And then FDA is run to dimensionality reduction data for feature extraction.

KNN is a non-parametric classification algorithm, which has been widely applied in the field of data mining and pattern recognition, etc.. It’s basic idea is as follows: X is a sample set to be classified, which is divided into training and test sets; finding out K samples from the training set that is the closest or most similar to test set, then according to the K samples to determine the class of samples of test set.

The experimental analysis

The first 40 samples of three kinds of apples were chosen to constitute training set, while the left 20 samples of three kinds of apple constituting test set. So that the training set consists of 120 samples, and the test set consists of 60 samples. Three methods of cluster analysis were used to spectral data after pretreatment by wavelet soft threshold: (1) PCA was firstly used to spectral data to reduce dimension, then KNN was used to classify, namely (PCA+KNN). After the test, while K=3, the correct clustering rate reached the maximum 78.33%. (2) FDA was firstly used to spectral data ,then KNN was used to classify, namely (FDA+KNN). After the test, while K=3 , the correct clustering rate obtain the maximum 75%. (3) PCA was firstly used to spectral data , then FDA was used, finally the KNN was used, namely (PCA+FDA+KNN). While K=4, the correct clustering rate reached the maximum 96%.

The classification results of the third method is shown in figure 2. The correct identification number and the correct recognition rate are shown in table 1. Obviously, the third method for Apple's correct recognition is advantage, the correct recognition rate of which reaches more than 96%. It is far higher than that of the other two methods.

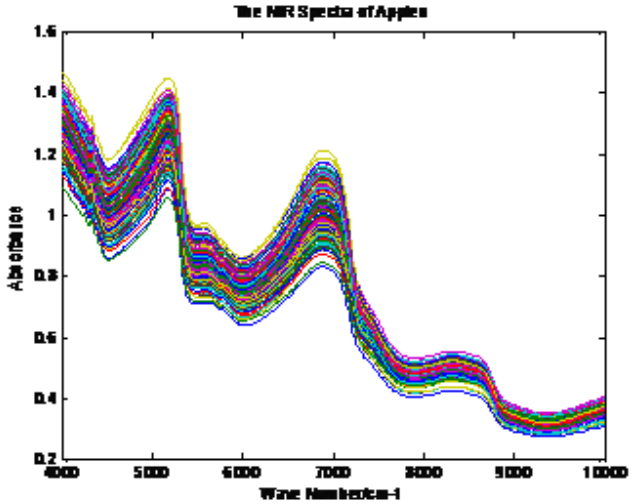

Fig.1. The NIR of three kinds of apples

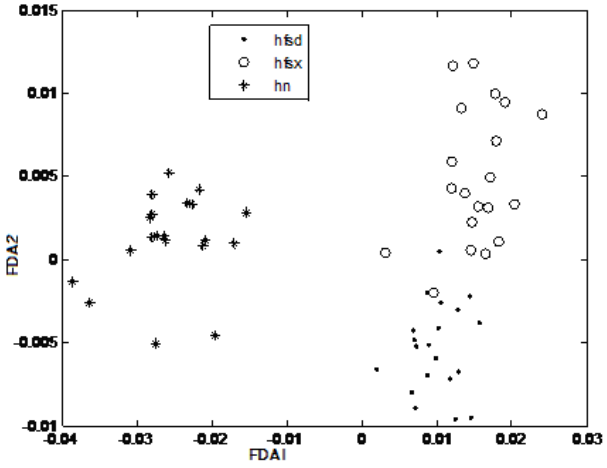

Fig.2.PCA+FDA+KNN classification results

Table 1 Comparison of Three Algorithms

Algorithms	Number of Correct Classifications	Correct Recognition Rate
PCA+KNN（K=3）	47	78.3333
FDA+KNN (K=3)	45	75
PCA+FDA+KNN (K=4)	58	96.667

Conclusions

Nondestructive, fast and accurate classification and identification method was proposed in this paper. It can realize different kinds of apples, and provide a new idea for near infrared spectral analysis.

Acknowledgements

This work was financially supported by the natural science foundation of Sichuan Province Education Office (12ZA070).

References

[1] S.Q.SUN,J.M.TANG, Z.M.YUAN, et al:Spectroscopy and Spectral Analysis, 2003, 23(2):258-261. (In Chinese)

[2] W. YU,K.L. YONG: Food Science, 2003, 24(11): 105-108. (In Chinese)

[3] S.H.WANG,Q.F. YIN, Y.L.FAN,et al:Spectroscopy and Spectral Analysis, 2004, 24(8): 957-960. (In Chinese)

[4] R.Q. GAO，S.F.FAN,et al: Spectroscopy and Spectral Analysis，2004.24(12):1563-1565. (In Chinese)

[5] Q.S.CHEN，J.W.ZHAO, H.D.ZHANG，X.Y.WNG:Acta Optica Sinca,2006.6(26):933-937.(In Chinese)

[6] X.B.BU, B.WU, H.W.JIA:Computer Engineering and Application,2013.49(2):170-173,193.(In Chinese)

[7]Y.B.SANG, Research of Classification Algorithm Based on K Nearest Neighbor(MS.,Chongqing University,China 2009),P.32.