Identification of Lettuce Storage Time based on Spectral Preprocessing Technology and PCA+SVM

Sun Jun[1,2]*, Jin Xiaming[1], Mao Hanping[2] and Wu Xiaohong[1]

[1]School of Electrical and Information Engineering of Jiangsu University,
Zhenjiang 212013,P.R. China
[2]Laboratory Venlo of Modern Agricultural Equipment, Jiangsu University,
Zhenjiang 212013,P.R. China.

(Received: 27 September 2013; accepted: 04 November 2013)

Based on differences in near infrared spectroscopy of lettuce leaf in different storage time, the non-destructive identification method of lettuce storage time was researched. Because there are various noises in lettuce leaf spectroscopy coming from external environmental factors, chemical composition and photoelectric detection circuit, several pretreatment methods, such as smoothing, MSC (Multiplication scatter correction), SNV (Standard Normalized Variable), BC (Baseline Correction), VN (Vector Normalize), FD (First Derivative) and SD (Second Derivative) were studied and used to pre-treat lettuce leaf spectroscopy respectively. PCA (Principal Component Analysis) method was used to reduce dimension and extract feature of spectroscopy pretreated. Finally, the identification model was built based on SVM(Support Vector Machine).Test results showed that, the model based on lettuce leaf spectroscopy pretreated by SD is best because its lettuce storage time accuracy was 100%,which was better than those of other models. The result also showed that, the spectrum noise factors affected the modeling accuracy, so it is important to select an appropriate pretreatment method, and SD+PCA+SVM is an effective identification model of lettuce storage time based on Near Infrared Spectroscopy.

Key words: Near Infrared Spectrum; Lettuce; Storage Time.

Lettuce is a kind of common nutrious vegetable eaten by people, which is rich in protein, carbohydrate, vitamin, as well as some mineral substance, and it can be functioned as medicine to clear heat, diminish inflammation, analgesia and hypnosis, lower cholesterol levels, treat neurasthenia. However, it's better to keep the vegetable fresher, so we shouldn't keep it too long, because tincture, shape, and taste will change with time going, and addled and degenerative vegetable not only can't meet people's need of taste and nutrition, but also does harm to people's health. The decrease of the agricultural products freshness will cause the physiological property change in blade surface and internal organization, thus making changes in the spectral reflection characteristic of crop leaves. Based on this theory, some foreign scholars use spectrum technology to do research on the storage time and freshness of agricultural product[1-5]. At home, Liming Wu measured the FTIR spectrum of bee milk in different temperature and storage time, and evaluated bee milk freshness quickly based on computer aided analysis technology[6]. Qiaohua Wang collected respectively spectral reflectance curve of shell eggs and whole liquid eggs, and researched the relationship between the freshness and the spectral reflection characteristics[7]. Xiaohong Wu distinguished the storage time of the pork sample using Adaboost+OLDA[8]. Shibang Ma studied the method of distinguishing the beef freshness, using the spectrum technology and SVM(Support Vector Machine)[9]. It can be seen, the spectrum technology has a large potential in diagnosing the internal

* To whom all correspondence should be addressed.
Tel.: +86 13775544650; Fax:+86 051188780088;
E-mail: Sun2000jun@ujs.edu.cn

quality of the agricultural product, but there is still no any report about the storage time and the freshness of the lettuce and any other vegetable.

Because the absorption band of NIR(Near Infrared Reflection) region is relatively broad and overlaying severely, spectral signal detected by the spectrum instrument include a variety of noise, besides the information of the measured sample. The original NIR has to be preprocessed, in order to lessen noise, cut down or remove effects coming from various kinds of untargeted elements, and enhance its resolution and response rate[10].

In this article, some methods, such as smoothing, multiplicative scatter correction, Standard normal variable correction combined with trend method (SNV), vector normalization, first derivative, second derivative spectra pretreatment, will be separately used to preprocess the original spectrum from the lettuce leaves. Later, combined with PCA feature extraction and SVM algorithm, an optimal modeling method will be sought to distinguish lettuce storage time.

MATERIALS AND METHODS

Sample preparation

The experiment on lettuce cultivation was conducted in the glass house of Jiangsu University. In mature period, 30 strains of lettuce samples were picked, and 2 slices of full blade at the same part on each lettuce were gathered in the lab, 60 slices in total, and label tags were pasted to each leaf stalk end. The indoor constant temperature in the lab is 27°C. Separately, 1h,12h,24h,36h,48h after being picked, the lettuce leaves samples were taken out from the refrigerator freezer, after being lain for 15 minutes indoor, the NIR spectra of lettuce leaves sample were collected. Later, the leaves samples were immediately put into the sealed plastic bag in the refrigerator freezer whose temperature is set as 7°C.

Spectral acquisition

AntarisII Near-IR Analyzers was chosen as NIR instrument, which is FT-NIR of made in Thermo Fisher of US, using InGaAs detector, whose spectral region is 10000-4000cm^{-1}, scan times is 32, wave number interval is 7.712cm^{-1}, resolution ratio is 16cm^{-1} [11].In the process when the NIR data is collected, the indoor temperature is kept on 27°C or so, and the humidity keeps constant.

The NIR data were collected from the left, right and middle point in each lettuce leave, avoiding the main vein. The average of the three spectra data is regarded as the origin spectrum data of the sample. When collecting the sample data, we collected 60 sample spectroscopic data in a storage time, so there were 300 sample spectroscopic data at 5 different storage time. The curve graph drew in terms of the original spectroscopic data of all samples was shown in the Fig.1.

Spectrum Pretreatment

Smoothing

In this article, moving average smoothing method whose window include odd number (2w+1) of wavelength points is used[12]. The length of this window is set as 9 data points. The figure of spectrogram after smoothing process is shown as Fig. 2.

Multiplicative scatter correction

MSC think that the relationship between the reflection absorbance values of each sample under the arbitrary wavelength point and the corresponding absorbance of average spectrum is similarly linear. Linear intercept and the slope can be obtained by linear regression of spectral set, and can be used to correct each of the spectrum[13]. The spectrogram after MSC is shown in Fig. 3. It can be seen from Fig. 3, the spectral radiation has been corrected preferably, and an ideal spectrum has been gained.

Standard normal variable correction combined with detrended method

SNV is used to correct the spectral differences caused by the samples' scattering[13]. The spectrogram was processed in the standard normal variable correction combined with detrended method and shown in Fig. 4. It can be seen from Fig. 4 that, the curve of spectrum has become more smooth, compact, and it is effective to remove the effect of noise.

Baseline correction

Spectrum is divided into some equal windows according to wavelength based on baseline correction. Eigenvalues of each window points are connected with straight lines. Curve is fitted as an approximate linear relationship, and the points not on the fitting curve are considered as invalid points. The spectral data after correction is the difference that original spectral data minus baseline data. The spectrogram after the baseline correction is shown in Fig. 5. It can be seen from Fig. 5 that, baseline shift of the spectrum, to a certain extent, has been improved, whose quality has been also improved.

Vector Normalization

Average spectrum intensity values are computed firstly, subtracted from the spectrum, and then the quadratic sum of the intensity values are calculated lastly. The spectrum is divided by its root[13]. The spectrogram processed by the vector normalization is shown in Fig. 6. It can be seen from Fig. 6 that, the curve of spectrum has become smoother, and the spectral quality has been improved.

First derivative

It is regarded as a common spectral pretreatment method, used to remove the baseline shift in spectrum. The spectrogram processed by the first derivative is shown as Fig. 7. We can see overlapping peaks of absorption peaks in the spectrum was exaggerated, increasing the spectrum information.

Second derivative

It is used to eliminate baseline drift in spectrum, which can distinguish the overlapping peak and increase the resolution and sensitivity[14]. Smooth point of the second derivative is set as 9. The spectrogram processed by the second derivative is shown in Fig. 8. We can see that, the spectral peaks give more prominence, and the spectral difference is magnified, together with, the spectral characteristics is enhanced while eliminating the particle size.

Feature extraction

Near-infrared spectroscopic data is a set of multivariate data, including redundant information. Multivariate Data Analysis (MVDA) made it possible to extract useful information, remove duplication of information and reduce data dimensions from the original spectral data. PCA (Principal Component Analysis), spectral data dimensionality reduction algorithm, is separately used in this paper for dimensionality reduction and feature extraction of spectral data after every pretreatment. The main objective of PCA (Principal component analysis) is to use fewer variables to explain most of the variation in the original data, and many highly relevant variables are changed into those which are independent or unrelated to each other. Usually the number of elected variables is less than that of the original ones, which are several new variables that can explain most of the variation of in the data. The PCA can convert multivariate data sets to a new set of data, and the principal components of the new variables are orthogonal to each other.

Seven types of spectral pretreatment results in section 2 were respectively analyzed, using principal component analysis PCA algorithm, the results of which are shown in table 1.

When the number of principal components of second derivative is 29, the accumulated contribution rate exceeds 99%. So the number of principal components is set as 29. From table 1, it can be seen that, under the premise that the requested cumulative contribution rate is 99%, the number of principal component of SD is 29, but it is OK to extract 5 principal components in the remaining pretreatment methods.

SVM algorithm

Support Vector Machine (SVM) is a new tool for resolving machine learning methods based on optimization[15]. Its main principle is as follows. Sample set is assumed as (xi,yi),i =1,2,…,N, where N is the number of training samples, xi is sample characteristics (xi∈ Rn) , yi is the sample category (yi∈ {+1,-1}).If the training set is non-linear, the training set data need to be mapped into a high dimensional linear feature space based on non-linear function. And an optimal separating hyperplane is constructed in the feature space, shown as formula (1).

$$f\left(x\right) = \omega \cdot \phi\left(x\right) + b = 0 \qquad ...(1)$$

Where, φ(x) is a nonlinear function, ωis a normal vector of hyperplane, and b is offset of the hyperplane. And ω and b together determine the location of optimal separating surface.

RESULTS AND DISCUSSIONS

Supporting vector machine is a new tool to solve the problem of machine learning using optimization method[15]. SVM classification algorithm is used in this paper to set up model to identify lettuce storage time, with polynomial as kernel function, the penalty factor c and kernel parameters[3] which has significant impact on model take default values.

Table 1. Spectral principal component analysis result after preprocessing

Pre-treatment method	Contribution rate					Cumulative contribution rate PC accumulation
	PC1	PC2	PC3	PC4	PC5	
Raw	96.5571%	2.026%	1.3392%	0.0436%	0.023%	99.9889%
Smoothing	96.5655%	2.0224%	1.3351%	0.0432%	0.023%	99.9892%
MSC	84.9819%	10.1512%	3.2252%	0.7606%	0.4717%	99.5906%
SNV+Detrending	77.8362%	17.8382%	1.6026%	1.2288%	0.8032%	99.309%
BC	66.0807%	22.138%	8.677%	1.8541%	0.5167%	99.2665%
VN	81.4691%	16.6739%	1.1996%	0.4094%	0.1424%	99.8944%
FD	87.6779%	9.9452%	1.2089	0.4184%	0.2291%	99.4795%
SD	45.8466%	25.3014%	9.7617%	1.2044%	0.7267%	82.8408%

Table 2. PCA + SVM classification results of the samples treated by different pretreatment methods

Pre-treatment method	Principal component (PC n)	Calibration accuracy	Test accuracy
Raw	5	98%	63%
Smoothing	5	99%	68%
MSC	5	100%	72%
SNV+detrending	5	100%	89%
Baseline correction	5	100%	94%
Vector normalize	5	100%	75%
First Derivative	5	100%	99%
Second derivative	29	100%	100%

In the experiment of collecting sample, a total of 60 samples of lettuce leaves were collected. Spectral data were gathered for 5 times, 1h, 12h, 24h, 36h and 48h after the acquisition respectively. Spectrum of every lettuce leaf was gathered, that is to say, there were 60 spectra samples in each time class, amounting to 300 samples in total in 5 time classes. 40 samples were selected randomly from each class, a total of 200 as a training data set, the remaining 100 as a test data set.

Original spectra were pretreated respectively using the 7 kinds of spectrum pretreatment methods, and features were extracted using PCA respectively. 5 principal component features were adopted using the method of smooth, diversity scattering calibration, standard normal variable correction combined with detrend, baseline calibration, vector normalization, and first

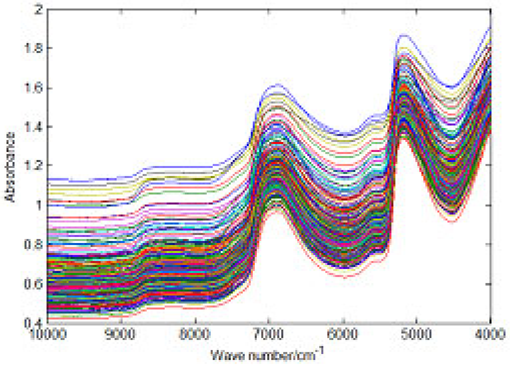

Fig. 1. Original curve of spectrum of all sample

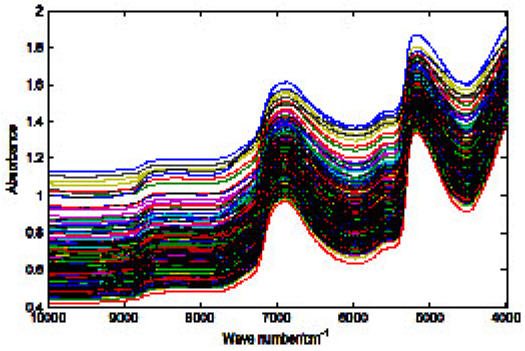

Fig. 2. Spectra after smoothing pre-treatment

derivative.29 principal component features were adopted when using second-order derivative. Accumulated contribution rate of the principal components reached 99%. SVM was used to set up models based on the samples handled by different pretreatment methods. The results of test experiment were shown in Table 2.

From table 2 it can be seen that various classification models have high calibration accuracy close to 100%. From the forecast results, SD+PCA+SVM test accuracy is up to 100%, superior to that of other combination models, the robustness of which is strongest.

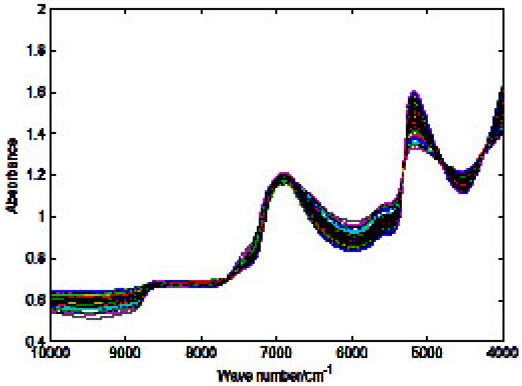

Fig. 3. Spectra after MSC pre-treatment

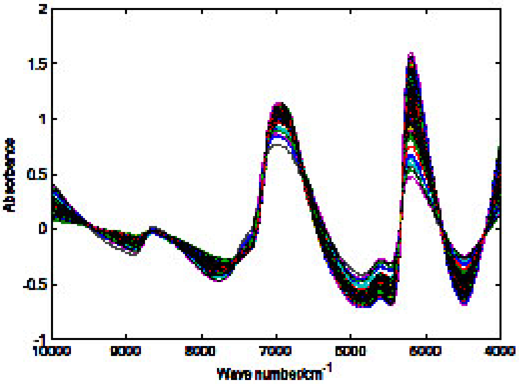

Fig. 4. Spectra after SNV+detrending

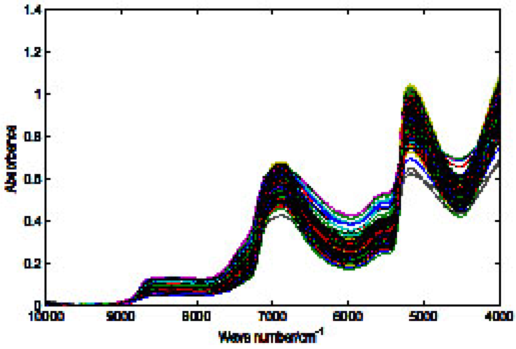

Fig. 5. Spectra after BC pre-treatment

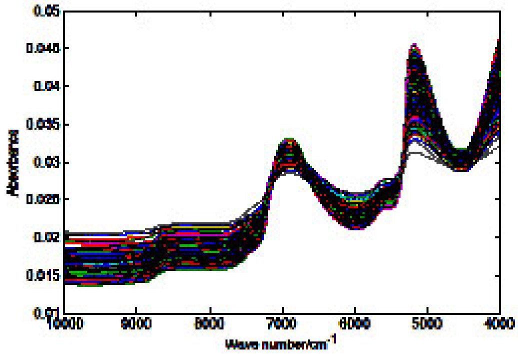

Fig. 6. Spectra after VN pre-treatment

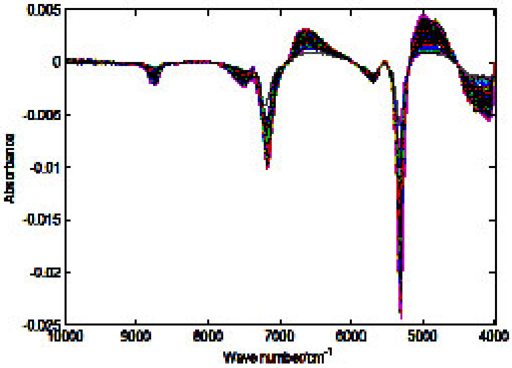

Fig. 7. Spectra after FD pre-treatment

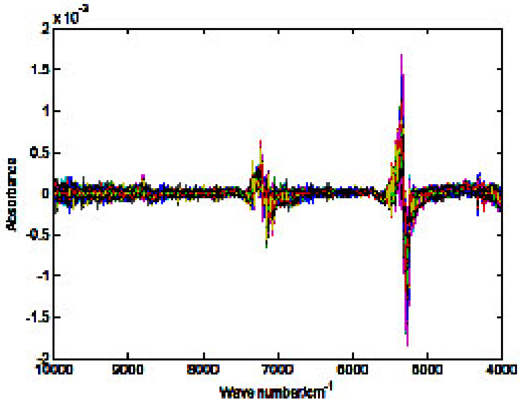

Fig. 8. Spectra after SD pre-treatment

CONCLUSION

Several pretreatment methods were used to pre-treat lettuce leaves spectrum, such as smoothing, MSC, SNV, BC, VN, FD,SD. Overall, these spectrum pretreatments can effectively remove noise, making spectral curves smoother.

Dimension of the spectroscopic data processed by various pretreatment methods was reduced based on principal component analysis method, and SVM was used in classification identification modeling finally. Test results showed that calibration accuracy rate is 100% in the SD preprocessing model, and identifying accuracy rate is 100%, superior to the ones of other models based on spectra samples pre-treated in several pre-treatment methods. It is illustrated further that, the noises in the spectra are the factors that affect the accuracy of modeling. The selection of the appropriate spectrum pretreatment method is very important. Results also show that SD+PCA+ SVM is an effective method based on spectroscopic technique to identify lettuce storage time.

ACKNOWLEDGMENTS

This work is supported by National natural science funds projects (31101082), A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions(PAPD) and the Open Fund Project of KeyLaboratory of Agri-informatics , Ministry of Agriculture (2013007).

REFERENCES

1. Prevolnik, M., Candek-Potokar, M., Novic, M., et al. An attempt to predict pork drip loss from pH and colour measurements or near infrared spectra using artificial neural networks., *Meat Science*, 2009; **83**: 405–411.
2. Del Moral, F.G., Guillen, A., del Moral, L. G., et al. Duroc and Iberian pork neural network classification by visible and near infrared reflectance spectroscopy., *Journal of Food Engineering*, 2009; **90**: 540–547.
3. Geesink, G.H., Schreutelkamp, F.H., Frankhuizen, R., et al. Prediction of pork quality attributes from near infrared reflectance spectra. *Meat Science*., 2003; **65**: 661–668.
4. Flatten, A., Bryhni, E.A., Kohler, A., et al. Determination of C22:5 and C22:6 marine fatty acids in pork fat with Fourier transform mid-infrared spectroscopy. *Meat Science*., 2005; **69**: 433-440.
5. Takashi K., Agnar H.S., Karsten H. VIS/NIR spectroscopy for non-destructive freshness assessment of Atlantic salmon fillets. *Journal of Food Engineering*.,2013; **116**: 758-764.
6. Wu, lM., Zhou Q., Zhao J. et al. Research on overall assessment of royal jelly freshness by FTIR spectroscopy. *Spectroscopy and Spectral Analysis*., 2009; **29**(12): 3236-3240.
7. Wang, Q.H., Zhou, P., Xiong, L.R., et al. Spectral reflectance characteristic and freshness of chicken egg. *Journal of Huazhong Agricultural University.,* 2008; **27**(1): 140-143
8. Wu, X.H., Tang K., Sun J. Discrimination of pork storage time using near infrared spectroscopy and Adaboost+OLDA. Spectroscopy and Spectral Analysis., 2012; **32**(12): 3288-3241.
9. Ma, S.B., Xu, Y., Peng, Y.K., et al. Assessment of beef freshness based on spectral technology with support vector machine. *Journal of Food Safety and Quality*., 2012; **3**(6):603-607.
10. Xia, J.F., Li, P.W., Li, X.Y., et al. Effect of different pretreatment method of nondestructive measure Vitamin C content of umbilical orange with near-infrared spectroscopy. *Transactions of the Chinese Society for Agricultural Machinery*., 2007; **38**(6):107-111.
11. Shi, J.Y., Zou, X.B., Zhao, J.W. NIR spectra in non-invasive measurement of cucumber leaf chlorophylls content. *Transactions of the Chinese Society for Agricultural Machinery*., 2011; **42**(5): 178-182.
12. Wang, W.M., Dong, D.M., Zheng, W.G. et al. Pretreatment method of Near-Infrared diffuse reflection spectra used for sugar content prediction of pears. *Spectroscopy and Spectral Analysis*., 2013; **33**(2):359-362.
13. Zhang, X.C., Wu, J.Z., Xu, Y. Near infrared spectroscopy and its application in modern agriculture. *Electronic Industry Press*, 2012; pp 104.
14. Ni, Z., Hu, C.Q., Feng, F. Progress and effect of spectral data pretreatment in NIR analytical technique. *Chin J Pharm Anal.,* 2008; **28**(5): 824-829
15. Ding, S., Qin, Q.Q., Chen, L. *et al.* Hyperspectral Classification with Swarm Intelligence Optimization Algorithms. *SENSOR letters*, 2012;**10**(8):1759-1767.