Near infrared spectroscopy for evaluation of apples using the K-mean algorithm

Masahiro Muramatsu,[a] Yoshiyasu Takefuji[a] and Sumio Kawano[b]

[a]*Graduate School of Media and Governance, Keio University 5322 Endo, Fujisawa 252-0861, Japan*

[b]*National Food Research Institute, 2-1-2 Kannondai, Tsukuba 305-8642, Japan*

Introduction

Nondestructive quality evaluation schemes for agricultural products using near infrared (NIR) spectroscopy have been developed since Norris detected the difference in moisture content in grain in 1965.[1] Norris revealed that diffuse reflectance and transmittance spectra of agricultural products contain information about chemical structures because each of the structures has specific absorption properties.[2] This detection is profitable for both producers and consumers, since NIR can predict internal compositions of products. The producers can provide valuable products and consumers can receive high quality foods. Through Norris' contribution, NIR spectroscopy has been used practically and spread widely as an automatic on-line method for evaluating food quality.

A number of studies using NIR spectroscopy have been performed to determine compositions of different types of fruits and vegetables. Most research uses the regression approach and shows high correlations as follows. Birth *et al*. determined dry matter in onions[3] and Dull *et al*. determined soluble solids in cantaloupe melons.[4] Kawano *et al*. analysed sugar content of intact peaches and showed high correlation ($R = 0.97$) between NIR measurements and Brix values which indicate the sweetness of the fruit.[5] Temma and his colleagues revealed high correlations between NIR measurements and constituents of apples including Brix value ($R = 0.94$), sourness ($R = 0.83$) and firmness ($R = 0.75$).[6]

In this paper, a K-mean algorithm is applied to the analysis of intact apples for sugar content. The K-mean algorithm is a clustering algorithm proposed by MacQueen.[7] The clustering is a grouping of data with similar characteristics and is used for various data analyses including spectral computing.[8]

K-mean algorithm

The K-mean algorithm is a clustering method for grouping data with similar characteristics and the steps of the algorithm are as follows.[9]

1. Partition the input vectors into K clusters randomly and compute the central points of the clusters as an initial condition.
2. Assign each vector to the cluster with the nearest central point and update the central points and the clusters.
3. After several iterations, when the change of the cluster becomes small, the programme is terminated.

Near Infrared Spectroscopy: Proceedings of the 10th International Conference

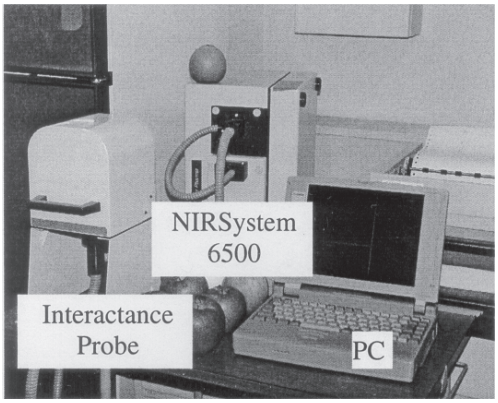

Figure 1. NIR instrument.

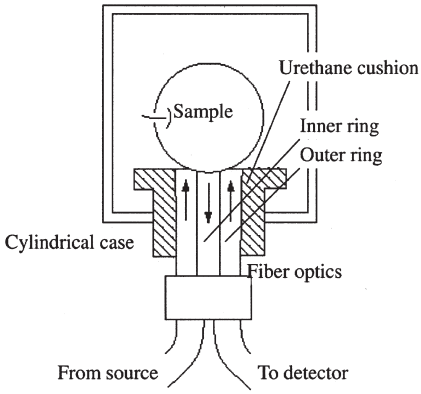

Figure 2. Interactance probe.

Experiments

Spectral acquisition

NIR spectra were acquired with an NIR instrument shown in Figure 1. The instrument has three components. One component is an "Interactance Probe". The fibre-optic probe used has an outer ring illuminator and inner ring receptor (Figure 2). The sample was located on the urethane cushion in a cylindrical case. Another component is an NIR spectrophotometer called NIRSystem 6500. A absorbance of the sample was measured by comparing near infrared energy reflected from the sample with that from the standard reference (8 cm-diameter teflon sphere). Teflon was used as a standard material because it has low absorption. Birth *et al*. also used a Teflon rod as a standard reference. The other component is a PC.

67 intact apples (variety Fuji), cultivated in Aomori prefecture in Japan, were used as the sample. The experiment was performed at 22°C. The spectral data were acquired every 2 nm from 400 to 1100 nm and data from 750 to 1050 nm were used in subseuqent calculations.

Chemical measurement

Reference chemical measurements of Brix values were performed after NIR spectra were acquired. The Brix values are determined using a commercially available refractometer.

Results and discussion

Figure 3 shows the original NIR spectra of the 67 apples. Since NIR spectra of foods are composed of broadbands arising from overlapping absorption, 2nd derivative spectra are used for calculation. With the single regression, the highest correlation coefficient is –0.654 when the wavelength of the spectrum is 904 nm (segment size is 8 nm and gap size is 0 nm). Therefore, the following calculations were performed using the conditions mentioned above. Table 1 is a quotation from *Near Infrared Spectroscopy of Food Analysis*[10] and shows absorbance wavelengths against specific chemical structures.

Table 1. NIR absorbance wavelengths.

Vibration	Wavelength (nm)
C–H str. third overtone	874, 900, 913, 938
O–H str. second overtone	990, 1000
2 × C–H str. + 3 × C–H def.	1015

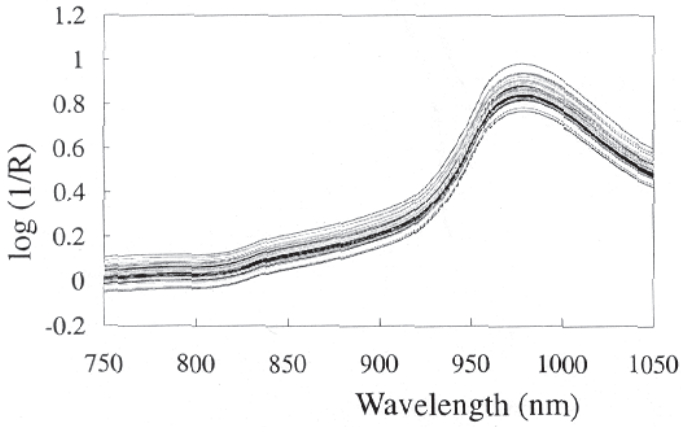

Figure 3. Original spectra of apples (*R*: reflectance).

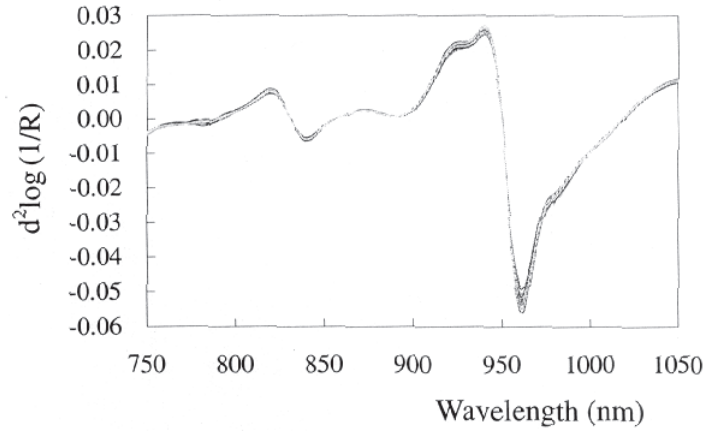

Figure 4. 2nd derivative data (*R*: reflectance).

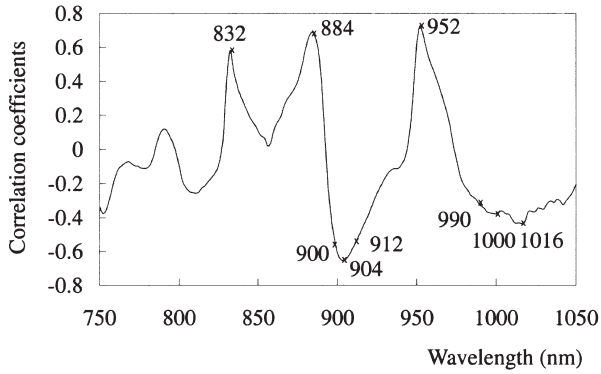

Figure 5. Correlation coefficients between d²log(1/*R*) and Brix value plotted against wavelength.

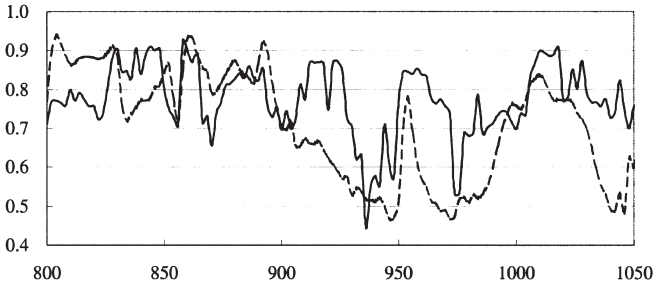

Figure 6. Correlation coefficient between actual and computed Brix value plotted against the second wavelength (the first wavelength: 904 nm.

Figure 4 shows the 2nd derivative spectra and Figure 5 shows correlation coefficients between the Brix values and 2nd derivative spectra which are called "correlation plots". In the case of the 2nd derivative spectra, the correlation plots should show negative peaks at the sugar bands. Therefore, the three wavelengths, 832, 884 and 952 nm, which show high positive correlation coefficients, are not regarded as the absorbance wavelengths.

The K-mean analysis was performed with a 57 sample calibration set and a 10 sample prediction set. The input is the spectra at the two wavelengths and the output is five Brix values which are the mean Brix values of the samples of the five clusters in the calibration set. The number of iterations is 100. In order to estimate the K-mean results, the regression results are measured (Figure 6). Figure 7 shows the plots of the samples in the calibration set. The circles indicate the central points of the cluster and the other marks indicate the samples which belong to the same cluster. The K-mean results show correlations between the actual and predicted Brix values as high as those produced by the regression results. In Figure 6, K-mean results are shown for the best values of correlation coefficients when initial conditions are changed. The highest correlation coefficient of K-mean is 0.939. However, the coefficient is calculated with the spectra at 804 and 904 nm. According to the NIR absorbance wavelengths shown in Table 1, 801 nm is not considered as an NIR absorbance. Then, the highest correlation coeffi-

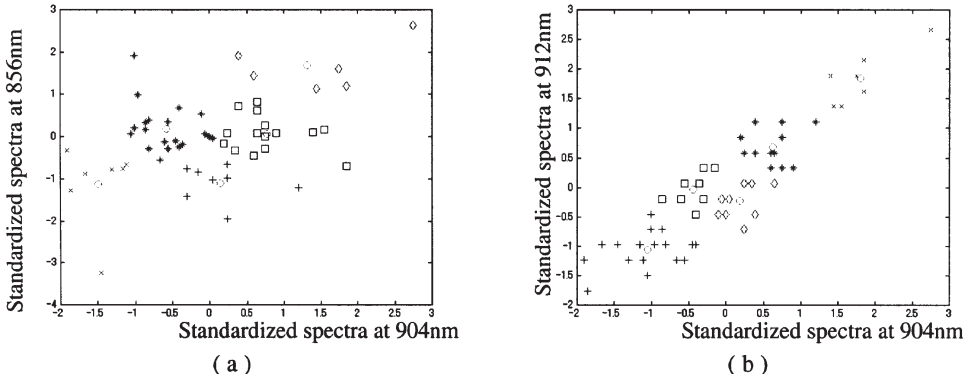

Figure 7. Plots of the sample . The clustering results with the spectra at (a) 856 and 904 nm and (b) 904 and 912 nm. The circle indicates the central point of the cluster and the other marks indicate the samples of each cluster.

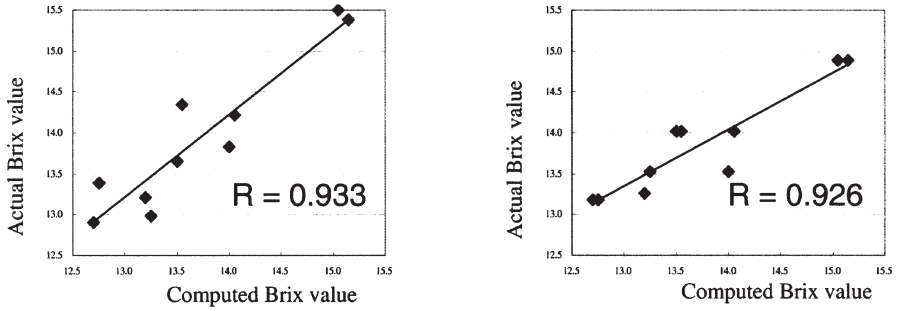

Figure 8. Plot of actual vs predicted Brix value. The regression result is calculated using (a) wavelengths 862 and 904 nm and (b) the K-mean result is calculated using wavelengths 865 and 904 nm.

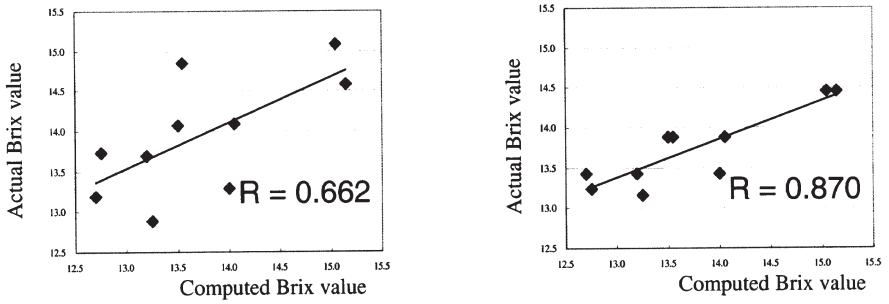

Figure 9. Plot of actual vs predicted Brix value. The results are calculated using wavelengths of 904 and 912 nm for both (a) the regression and (b) the K-mean procedures.

cient related to the absorbance is 0.926 at 865 and 904 nm [Figure 8(b)]. On the other hand, that of regression is 0.933 at 862 and 904 nm [Figure 8(a)].

The difference between K-mean and regression occurs when the spectra at 904 and over 906 nm are used for the analysis. When the spectra at 904 and 912 nm are chosen, the K-mean result shows a higher correlation than the regression (Figure 9). It is noted that 912 nm is considered as a NIR absorbance in Table 1.

Finally, it is confirmed that the K-mean analysis shows high correlations between the actual and computed Brix values corresponding to the selection of the used wavelengths.

Conclusion

In this paper, a new approach analysing near infrared spectroscopy of apples using K-mean algorithm is proposed. The result of K-mean analysis shows correlations between the actual and computed Brix values which are as high as those produced by regression analysis. Moreover, it is confirmed that the K-mean analysis shows high correlations between the actual and computed Brix values corresponding to the selection of the wavelengths used.

References

1. D.R. Massie and K.H. Norris, *Transaction of the ASAE* **8(4),** 598 (1965).
2. P. Chen, *Proc. Int. Conf. On Agric. Machinery Engineering*, **1,** 171 (1996).

3. G.S. Birth, G.G. Dull, W.T. Renfroe and S.J. Kays, *J. Amer. Soc. Hort. Sci*. **110,** 297 (1985).
4. G.G. Dull, G.S. Birth, D.A. Smittle and R.G. Leffler, *J. Food Sci*. **54,** 393 (1989).
5. S. Kawano, H. Watanabe, M. Iwamoto, *J. Japan Soc. Hort. Sci*. **61,** 445 (1992).
6. The Report of Aomori Advanced Industrial Technology Center, (1992).
7. J. MacQueen, *5th Berkeley Symp. Math. Statist. Prob.* **1,** 281 (1967).
8. H.C. Romesburg, *Cluster Analysis for Researchers.* Lifetime Learning Publications, California, USA (1984).
9. L. Kaufman, *Finding groups in data*. Wiley-Interscience, New York, USA (1990).
10. B.G. Osborne and T. Fearn, *Near Infrared Spectroscopy in Food Analysis*, Longman Scientific & Techinical, Harlow, UK (1986).