New Zealand Journal of Crop and Horticultural Science, 2006, Vol. 34: 177–181
0014–0671/06/3402–0177

Use of near infrared spectra to identify cultivar in potato (*Solanum tuberosum*) crisps

N. YEE

W. T. BUSSELL
School of Natural Sciences
Unitec New Zealand
Private Bag 92025
Auckland, New Zealand
email: nyee@unitec.ac.nz

G. G. COGHILL
Department of Electrical Engineering
The University of Auckland
Private Bag 92019
Auckland, New Zealand

Abstract Near infrared spectra were collected of potato crisps from potato (*Solanum tuberosum*) cultivars 'Whitu' and 'Fianna'. Pattern recognition techniques were used to classify the spectra. Linear discriminant analysis performed as well as piecewise linear discriminant analysis in identifying the potato tuber variety used to produce the potato crisps. The success rate in separating the spectra into respective classes using discriminant analysis is 93%. This suggests that it is possible to use near infrared analysis for the purpose of identifying different cultivars in single batches of potato crisps.

Keywords Whitu; Fianna; piecewise linear discriminant analysis; classification

INTRODUCTION

The production of high quality potato (*Solanum tuberosum* L.) crisps by frying thinly sliced sections of tuber in oil requires production process settings that are cultivar specific. Mixing cultivars in batches of tubers to be processed, a common occurrence in New Zealand because cultivars overlap in maturity, has been demonstrated to cause product quality variations (Kita 2002).

Separating crisps from different cultivars on the processing line is a time consuming task unless automated. There are at present no successful automated inspection systems to separate crisps of different cultivars on a processing line. The possibility of using calorimetric techniques (Withers 1998) for separating cultivars is limited because only crisp colour is detected. The use of chemical properties of crisps is not possible because crisps are destroyed.

Near infrared analysis is a widely used tool for identification of different materials and chemical concentration determination. It has been used to determine nitrogen content of potato leaves (Young et al. 1997), tuber dry matter content (Dull et al. 1988), and disease identification in potatoes (Porteous et al. 1981). Near infrared analysis combined with pattern recognition techniques has been used to identify different materials in applications such as sorting of plastics for refuse recycling (van den Broek et al. 1997) and identification of nitrogen containing explosive materials in bomb detection (Lewis et al. 1997). However, to our knowledge, near infrared analysis has not been used to identify cultivar in a batch of potato crisps.

This paper describes the results of using near infrared analysis to identify potato tuber variety used in processed potato crisps. The statistical pattern recognition techniques of piecewise linear discriminant analysis (PLDA) and linear discriminant analysis (LDA) have been selected for the identification task. The intention is to use near infrared analysis in an automatic potato tuber variety identification system.

MATERIALS AND METHODS

Tubers of 'Fianna' and 'Whitu', whose time of availability for processing overlaps, were investigated. They were grown at Opiki, south-west

H05064; Online publication date 27 April 2006
Received 13 June 2005; accepted 8 March 2006

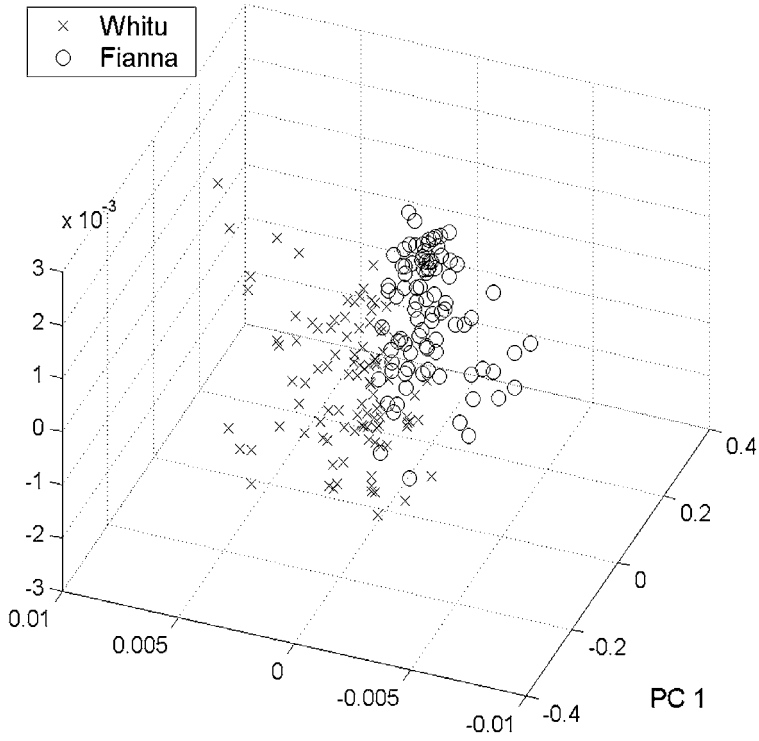

Fig. 1 Plot of the principal components one, two, and three.

of Palmerston North, New Zealand, and had been dug for no longer than a week when processed.

The tubers were processed in the standard conditions (Withers 1998; Yee 2005) at the processing factory in Auckland. They were cut into flat slices of 1.5 mm and fried in Canola oil in a fryer. Frying time was set at 180 s and oil temperature was 180°C.

A total of 192 samples of fried crisps in 20 g denominations were collected from the factory processing line. The first set contained 102 samples of 'Whitu', the second 90 samples of 'Fianna'. The samples were then inspected using a NIRSystems 6500 spectrometer. The spectra corresponded to the region between 400 and 2496 nm of the near infrared spectrum, representing 525 spectral wavelengths at 4 nm separation. The measurements were taken in reflectance mode and the mathematical treatment of the spectral variables was log(1/R), where R is the reflected spectral intensity at a specified wavelength.

The 192 spectra obtained were divided into two sets, a training set of 160 spectra (85 'Whitu' and 75 'Fianna') and a prediction set of 32 spectra (17 'Whitu' and 15 'Fianna'). The spectra were corrected for scatter using the piecewise multiple scatter correction algorithm of Issakson et al. (1993), with a pre-determined optimal piecewise multiple scatter correction window size of 260 nm (Yee & Klette 1999). Principal components analysis was performed on the 525 dimensional data set to reduce the dimensionality. The first three components explained 99% of the variance in the data (see Fig. 1).

Identification of cultivar in a potato crisp sample was achieved using linear discriminant analysis and piecewise linear discriminant analysis, optimised by simplex pattern recognition (Tou & Gonzalez 1974). Mathematical details of discriminant computation and simplex optimisation are given in Appendix 1.

Recognition values are taken as the percentage correctly classified in the cultivar class relative to all the spectra in the respective cultivar class. In training the classifier, the target group chosen was from 'Fianna', because processing this cultivar when the process setting is tuned for 'Whitu' results in product quality variations among the crisps processed from 'Fianna' tubers.

Recognition statistics were computed on the two data sets. (1) The first recognition statistic refers to the data set of 160 spectra used in training the classifier. The training set was used solely for assessing the prediction abilities of the classifier on data used to train the classifier. The recognition statistic is

Table 1 Pattern recognition results in separation of the two varieties of potatoes (*Solanum tuberosum*). (PLDA, piecewise linear discriminant analysis; LDA, linear discriminant analysis.)

Technique	No. training set	No. correctly classified	Recognition (%)	No. independent set	No. correctly classified	Recognition (%)
PLDA Whitu	85	82	96	17	16	94
PLDA Fianna	75	75	100	15	14	93
LDA Whitu	85	80	94	17	16	94
LDA Fianna	75	70	93	15	14	93

a percentage value of tubers correctly classified (from crisp spectra) relative to total number of tubers present for each respective class in the training data set. (2) The second recognition statistic refers to an independent set of 32 spectra not used in training the classifier. This independent set was used solely for assessing the prediction abilities of the classifier on data not used to train the classifier. The recognition statistic is a percentage value computed from the number of tubers correctly classified (from crisp spectra) relative to total number of tubers present for each respective class in the independent data set.

RESULTS AND DISCUSSION

The classification statistics achieved in computing the two classifiers are presented in Table 1. The PLDA achieved a recognition value of 100% in identifying 'Fianna' in the training set. The reverse error was a recognition value of 96% in identification of 'Whitu'. By comparison a single LDA achieved only a recognition value of 93% in identifying 'Fianna' and a reverse error recognition value of 94% for identification of 'Whitu' in the training set. This does not necessarily suggest that PLDA is more efficient than LDA in identification of 'Fianna' in new data, because the more complex PLDA rule will be more influenced by peculiarities of the training data which will not be present in new data. The overall prediction percentages are excellent for the independent set (Table 1, column 7), however both techniques perform the identification of different tuber varieties with equal success rates. Based on this, there is insufficient information to determine which techniques' performance was superior.

These results demonstrate that spectroscopic information, present in potato crisp spectra, allow the identification of potato tuber variety. The statistical procedures required to identify the two varieties from near infrared spectral data have several steps but the underlying measurements are simple to obtain, and the resulting models are of a high accuracy. The measurement using near infrared analysis is quicker and cheaper than other direct chemistry based methods and has been applied to samples that were directly off the production line; based on this, the next task is to scale up and automate the procedure for automated inspection on the factory floor.

PLDA and LDA classifiers have been used in this work, however artificial neural networks, genetic algorithms and other unsupervised pattern recognition techniques may improve the recognition percentage even further. Thus the future direction of this research will focus on investigating these techniques to improve recognition ability and scaling up the procedure for 100% automated inspection.

Since starch and glucose levels in potato tubers are affected by temperature variations and length of time in storage, and the amount of glucose has an effect on crisp quality (Gamble & Rice 1988), near infrared analysis is reliable only on crisps that have been produced from tubers stored at the same temperature for the same length of time.

ACKNOWLEDGMENTS

We thank Norm Stephenson (Conveyors and Engineering Ltd) and John K. Sinclair (Impac Machinery Ltd) for technical support, Krispa Foods Ltd and Tegal Laboratories for the use of their facilities and analytical equipment. The senior author also thanks Technology New Zealand for funding under contract IML801.

REFERENCES

Brissey G, Spencer R, Wilkens C 1979. High speed algorithm for simplex optimisation. Analytical Chemistry 51: 2295–2297.

Dull G, Birth G, Leffler R 1988. Use of near infrared analysis for the nondestructive measurement of dry matter in potatoes. American Potato Journal 66: 215–225.

Gamble M, Rice P 1988. The effect of slice thickness on potato crisp yield and composition. Journal of Food Engineering 8: 31–46.

Isaksson T, Kowalski B 1993. Piece-wise multiplicative scatter correction (MSC) and linearity in NIR spectroscopy. Applied Spectroscopy 42: 1273–1284.

Kaltenbach T, Small G 1991. Development and optimization of piecewise linear discriminants for the automated detection of chemical species. Analytical Chemistry 61: 936–944.

Kita A 2002. The influence of potato chemical composition on crisp texture. Food Chemistry 76: 173–179.

Lewis I, Daniel D, Griffiths P 1997. Interpretation of raman spectra of nitro-containing explosive materials. Part I: Group frequency and structural class membership. Applied Spectroscopy 51: 1854–1867.

Nollet L 1996. Handbook of food analysis. New York, Marcel Dekker. Pp. 72–73.

Porteous R, Muir A, Wastie R 1981. The identification of diseases and defects in potato tubers from measurements of optical spectral reflectance. Journal of Agricultural Engineering Research 26: 151–160.

Ritter L, Lowry S, Wilkens G, Isenhour T 1975. Simplex pattern recognition. Analytical Chemistry 14: 1951–1956.

Shaffer R 1995. Optimization methods for analysis of infrared spectral and interferogram data. Unpublished PhD thesis, Ohio University, Ohio, United States.

Tou J, Gonzalez R 1974. Pattern recognition principles. Reading, Massachusetts, Addison Wesley. Pp. 119–123.

Van den Broek W, Wienke D, Melssen W, Buydens L 1997. Optimal wavelength range selection by a genetic algorithm for discrimination purposes in spectroscopic infrared imaging. Applied Spectroscopy 51: 1210–1217.

Withers B 1998. Possible methods to reduce browning of potato crisps. Unpublished report, Department of Food Technology, Massey University, Palmerston North, New Zealand.

Wold H 1966. Multivariate Analysis. New York, Academic Press. Pp. 391–410.

Yee N 2005. Optimization of near infrared multi-spectral data. Unpublished PhD thesis, University of Auckland, New Zealand.

Yee N, Klette R 1999. Multiple scatter correction using an evolutionary algorithm. Auckland University, Auckland, New Zealand. CITR Technical Report No. 56.

Young M, MacKerron D, Davies H 1995. Factors influencing the calibration of near infrared reflectometry applied to the assessment of total nitrogen in potato. II. Operator, moisture and maturity class. Journal of Near Infrared Spectroscopy 3: 167–174.

Appendix 1 Classifier design.

In computation of a linear discriminant, a quantity is calculated for each member in the data set, the quantity is large for one class and small for the other class. The classifier then uses a decision boundary that depends on the mean, and can be chosen to reflect prior probabilities and the relative severity of wrong decisions. The piecewise linear discriminant analysis method for approximating non-linear separating surfaces uses multiple linear discriminants serially grouped to form a piecewise approximation of a nonlinear surface. In piecewise linear discriminant analysis (PLDA), linear discriminants are calculated sequentially, with each discriminant separating a portion of the patterns in the data space. The PLDA classifier is often termed a committee classifier, since the classification of unknown patterns requires the entire set of linear discriminants.

Figure 2A shows a graphical representation of such a classifier in which each linear discriminant has the same class on its pure-class side. Fig. 2B pictorially illustrates the problems associated with independent placement of individual linear discriminants.

Mathematically the procedure of classification requires the linear discriminant to be calculated such that:

$$\mathbf{w}^T \mathbf{X}_1 > 0 \quad (1)$$

$$\mathbf{w}^T \mathbf{X}_2 \leq 0 \quad (2)$$

where $\mathbf{X}_1$ represents the data points from class 1, $\mathbf{X}_2$ represents data points from class 2, and $\mathbf{w}$ is the discriminant weight vector. The data products in Equations 1 and 2 are termed discriminant scores. To offset the separating surface from the origin, the $\mathbf{X}_1$ and $\mathbf{X}_2$ vectors are augmented with a constant element.

In the PLDA method a Bayesian classification algorithm is used to calculate the initial discriminant weight vectors for the data set, given by:

$$\mathrm{w}_i = \mathrm{X}^T \mathrm{C}^{-1} \mathrm{m}_i - \frac{1}{2} \mathrm{m}_i^T \mathrm{C}^{-1} \mathrm{m}_i \quad (3)$$

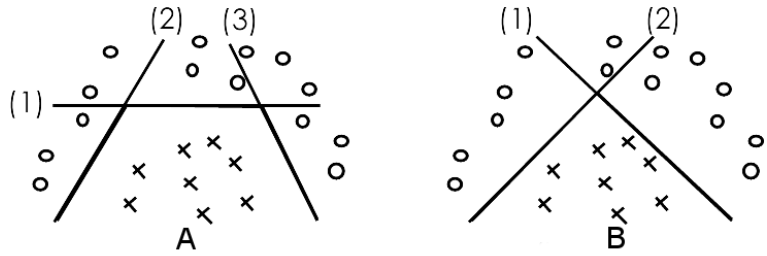

Fig. 2 **A**, Depiction of the three piecewise linear discriminants calculated sequentially and independently to form a classifier. **B**, Classifier after concurrent reoptimisation.

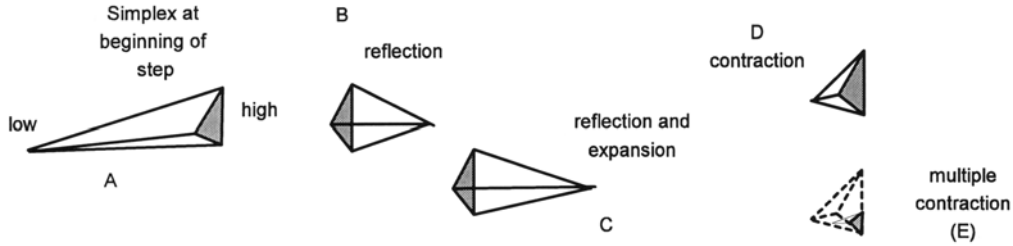

Fig. 3 **A**, Start of simplex; **B**, reflection from low point; **C**, reflection and expansion away from low point; **D**, contraction in one dimension from low point; and **E**, contraction along all dimensions towards high point.

where $\mathbf{w}_i$ is the discriminant weight function, **X** is the data set matrix, **C** is the covariance matrix, and $\mathbf{m}_i$ is the mean vector.

The Bayesian classifier is a statistical method based on incorporating prior beliefs as probabilities. Bayesian classifiers assume that the data belong to two multivariate normal distributions (Tou & Gonzalez 1974) which implies that the discriminant boundary is a plane. Assumptions of a multivariate normal distribution may be invalid, and consequently, the Bayesian discriminant is used only as a starting approximation for the final weight vector.

The implementation of PLDA classification uses a popular second form of optimisation known as simplex pattern recognition (Ritter et al. 1975). The computational procedure used can be found in Brissey et al. (1979).

The simplex method requires function evaluations of a geometric figure consisting in N dimensions, of $N+1$ points (or vertices) and all their interconnecting line segments, polygonal faces etc. For the application of linear discriminant optimisation, we are interested in simplexes that are non-degenerate, i.e., that enclose a finite inner N-dimensional volume. If any point of a non-degenerate simplex is taken as the origin, then the N other points define vector directions that span the N-dimensional vector space.

The simplex is started with $N+1$ points, defining the initial simplex, given by:

$$\mathbf{w}_i=\mathbf{w}_o+\lambda\, e_i \tag{4}$$

where $\mathbf{w}_o$ is the original linear discriminant weight vector, $\mathbf{w}_i$ is the updated linear discriminant weight vector, e_is are n unit vectors and l is the spanning constant.

The simplex uses a procedure of steps, moving the points of the simplex from where the function is lowest in the data space to higher points in the data space. The steps comprise of reflections (steps which conserve the volume of the data space and maintain nondegeneracy), expansions (which enlarge the data space), and contractions (which reduce the data space). These steps are described in Fig. 3. At each step, each point in the simplex is evaluated based on the response function. The response function chosen in this application followed that of other researchers investigating simplex optimisations (Shaffer 1995) in spectroscopic applications, given as:

$$e_i = N_s\, e^{(Ns - Nt)}\, M/\sigma \tag{5}$$

where e_i is the response function, M is the mean discriminant score, s is the standard deviation of discriminant scores with mean taken as zero, N_s is the number of class 1 patterns separated, and N_t is the total number of patterns placed on the pure side of the discriminant.

In the present work, the committee classifier calculated consisted of five linear discriminant weight vectors. Each vector was initiated using the Bayes calculation and the simplex algorithm was performed with 1000 iterations. The best vector was saved, and the procedure was then repeated until five weight vectors were calculated.