Rapid classification and blends analysis of tobacco mixtures using near infrared and artificial neural networks

Su-Chin Lo
United States Tobacco Company, R&D, Nashville, TN 37203, USA.

Introduction

Commercial tobacco blends are a complex mixture of different tobaccos and additives which are typically formulated to meet a specific composition. This complexity carries over to their substantial variation and the characterization and quantitative analysis of their properties are difficult. Therefore, an important challenge in the tobacco industry is the intake quality of raw material and the analysis of tobacco blends for process verification. One simple question has to be asked: Can we estimate which tobacco types have been mixed and to what extent are they present in terms of rapid analysis with minimal sample preparation?

One approach based on near infrared (NIR) spectroscopy and discriminant analysis to predict the tobacco blend identity has been proposed recently.[1] The differentiation of tobacco groups is clearly improved by the use of discriminant partial least square (PLS) techniques. For years, near infrared analysis has been developed as a quick and inexpensive analytical method and can provide useful information on the composition of agriculture products.[2,3] NIR spectroscopy provides a total chemical profile of a sample and may strengthen the discriminant power for further data analysis. However, only a few studies have been reported on NIR mixture classification using advanced multivariate data analysis.[4,5]

In the first study of this report, NIR spectroscopy coupled with chemometric methods were investigated to differentiate individual types of tobacco mixtures. One of the algorithms used here for mixture analysis is the artificial neural network (ANN), which is a robust, flexible and general method of modelling data. The use of ANN has primarily been focused on the classification of single categories but identification of separate components from mixture spectra has seldom been addressed. The results obtained by ANN were compared by other popular chemometric techniques including the k-nearest neighbours (KNN), soft independent modelling of class analogy (SIMCA) and Discriminant PLS (DPLS). The next question was whether a simple NIR diffuse reflectance measurement coupled with multivariate calibration would provide a fast and efficient way in the estimation of the blending ratios within a reasonable error. Among the calibration methods, the linear multivariate model from regular PLS regression was used but non-linear methods like ANN were also evaluated.

Theory

Spectral classification or discriminant analysis is a statistical technique for classifying individuals or objects into mutually exclusive and exhaustive groups on the basis of a set of spectral data. This process is a typical application of supervised pattern recognition methods. The goal of classification is to calculate class models and class boundaries based on a training set with samples

of known classes. An unknown sample is usually assigned to *one* class with the largest class density at that point. The two most popular classification methods used in chemometrics are k-nearest neighbours (KNN) and soft independent modelling of class analogy (SIMCA). Both KNN and SIMCA are similarity based techniques but their approaches are different. Classification with KNN is based on Eucliden distances among sample vectors while SIMCA develops latent variable models for each category in the training set.[6,7] However, these methods may not provide detailed class information if the unknown is a mixture.

The discriminant PLS (DPLS) is a generalized form of discriminant analysis based on the PLS modelling with binary *y*-variables. In fact, the DPLS algorithm is designed to correlate spectral (*X*) variations with class (*Y*) variations, i.e. to maximize the covariance between *X* and *Y* variables.[8] Therefore, the DPLS components can be thought of as the rotation of the principal component analysis (PCA) which maximizes the orthogonality of predicted class contributions. In order to determine individual components in the mixture, the *Y* matrix was designed to describe the presence/absence of the specific classes of a sample. A value of 1.0 was assigned to the spectral profile with the same type of tobacco, while a value 0.0 was assigned to the other types.

Artificial neural networks have been applied to a variety of chemical problems including non-linear multivariate regression and pattern recognition. The main goal of ANN classification is the *generalization*, i.e. being able to induce a concept that accurately classifies an unknown example.[9,10] Among the different architectures of ANN, the multilayer feed-forward neural network forms a dynamic system containing highly interconnected and interacting units in which a non-linear computing model can be developed. The most popular training method in multilayer networks is the back-propagation algorithm which was based on the gradient descent in error.

Likes the DPLS, the output of ANN is also a binary value representation by the use of presence/absence (1 or 0) in the multi-*Y* blocks of the training step. However, the output values from each class may fall between 0 and 1 simultaneously, if the unknown contains multiple classes. In this case, the output with multiple values can be thought as a "probability of class membership" for representing the possible composition in the mixture. One can choose a threshold value (0.1 to 0.2) for the determination of the presence of classes in the mixture. If the index of one class was below the arbitrary limit, the class was declared to be a either minor component or non-classified in a mixture.

Experimental

Tobacco samples

Data set 1 for mixture classification

A total of 127 pulverized tobacco samples containing six different tobacco leaves corresponding to type A1, A2, B1, B2, C1 and C2 were used in the training set for the classification purpose (Table 1). Two synthesized mixture spectra were prepared by adding two or three randomly selected spectra together from the training set. Three mixtures containing two or three tobacco types were also prepared by mixing equal amounts of tobacco and they were all selected from same crop year. Another three binary and ternary mixtures were also prepared by mixing tobacco components from another crop year.

Data set 2 for blends analysis

To quantify the composition of tobacco components, blended tobacco samples consisting of leaf type A1, B1 and C1 were mixed by weight percentage based on the Simplex mixture experimental design.[11] The concentrations of a mixture composed of three components are

Table 1. The design of output elements of mixture classification.

Tobacco types	Number of sample in training set	Class assignment in KNN and SIMCA	Class assignment in DPLS and ANN
A1	37	1	1 0 0 0 0 0
A2	16	2	0 1 0 0 0 0
B1	20	3	0 0 1 0 0 0
B2	20	4	0 0 0 1 0 0
C1	20	5	0 0 0 0 1 0
C2	14	6	0 0 0 0 0 1

constrained in that the relative concentrations of the components have the sum of 100%. There were 22 training samples prepared in the laboratory ranging from 0 to 100% (by weight) and 18 independent samples served as a validation set with appropriate percentage by weight.

Spectroscopy and statistical methodology

NIR diffuse reflectance spectra (log 1/*R*) were recorded on a NIRSystems (Silver Spring, MD) Model 6500 scanning spectrophotometer. The scanning range is between 1100 and 2500 nm with 2 nm data intervals. All reflectance spectra were transferred to GRAMS386 (Galactic Industries) for further data processing including the second derivative of the log(1/*R*) using the Savtisky–Golay 11-point quadratic smoothing algorithm. For the classification problem, auto-scaling was performed for each method (PCA, KNN, SIMCA, DPLS and ANN). However, only mean-centering was performed in the blend analysis including the PLS and ANN methods.

DPLS and ANN were constructed using multi-class *Y* matrix (A1, A2, B1, B2, C1 and C2 in Table 1) as an output. Cross-validation was used as an optimization process to select a suitable number of factors in the DPLS model. In ANN, a three-layer modified back-propagation learning (extended delta-bar delta rule) method was employed for improving the training speed. A sigmoid function was used as transfer function for the hidden units as well as the output units and the learning rate and momentum term were set up to 0.3 and 0.5 respectively. The topology of ANN in the classification was set up (20-30-6) which used the first 20 principal components from raw spectral data as inputs, 30 units for hidden layers and six classes as an output layer. The performance of a prediction scheme of each method was based on the Matthews' correlation coefficient (C_M):

$$C_M = \frac{pn - uo}{\sqrt{(n+u)(n+o)(p+u)(p+o)}}$$

where p is the number of true positive predictions, n is the number of true negative predictions, u is the number of under-predicted cases (class-like pattern was classified as non-class pattern) and o is the number of over-predicted cases (non-class-like pattern was classified as class pattern). The advantage of the correlation coefficient is that it gives a quantitative measurement of the

prediction. A correlation $C_M = 1$ indicates perfect identification of a mixture and $C_M = 0$ is expected for a prediction no better than a random one.[12]

In the blend analysis, the multivariate calibration models were built by PLS-1 regression and the non-linear ANN mapping method. Three-component tobacco mixtures (types A1, B1 and C1) were evaluated. Model optimization was carried out by cross-validation in PLS regression. The prediction error versus the number of factors in the PLS calibration model was used to evaluate the predicted composition. However, in order to resolve the over-training problem, the early stopping rule in ANN was based on the monitoring of the errors from both the training set and validation set simultaneously. The topology of ANN was 12-5-3. In the case of function approximation, a sigmoid function was used as transfer function for the hidden units but a linear function was applied for the output units. A modified Levenberg–Marquardt back-propagation learning algorithm was implemented and the initial learning rate and momentum term were set up to 0.8 and 0.7 respectively.[13] Overall, in both PLS regression and ANN, the prediction error is expressed as the root mean square of the difference between the known (y_i) and predicted ($\hat{y}_i$) values (*RMSEP*) for the samples in the training set and the validation set.

Results and discussion

Observations of the raw spectra of tobacco samples

The substantial chemical variation depends on the variety of tobacco leaves, on its circumstances of cultivation, seed, soil, growing location, ageing and curing conditions. Besides the principal chemical components such as carbohydrates (sugars and polysaccharides) and proteins, there are many constituents occurring and playing an important role in the organoleptic properties of tobacco. These chemicals include organic acids, polyphenols, volatile oils, waxes, enzymes, alkaloids, amides, amines and inorganic salts.

PCA was first performed on the overall spectral collection. The first, second and third principal factors explain 46%, 25% and 9% of the total variation in NIR measurements, respectively. Figure 1 plots the score of the first and third principal components (PCs) of a training set containing 127 spectra of six different tobacco types. It shows a satisfactory separation from each group. The A1

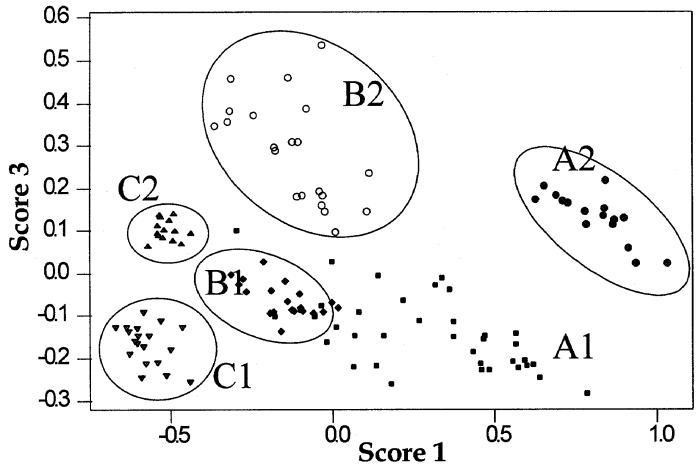

Figure 1. Projection of the scores on the first and third components for tobacco samples.

type tobacco was distributed widely along the axes of first principal components, indicating its wide variation. On the other hand, type A2, B1, B2, C1 and C2 tobaccos showed a clear separation with each other and also fell into a narrow and well identified region. In an overall evaluation, based on our knowledge of tobacco samples, the first PC clearly related to the variation of the "curing process" while the third PC may describe the variations of "growing location" among those tobaccos. These results demonstrated an ability to discriminate between tobacco types and were reasonable because the NIR spectral variations come from the intrinsic chemical constituents from tobacco.

Classification of tobacco mixtures

The results of the Matthews' correlation coefficient for each method are listed in Table 2. Only one or two components were partially identified on six mixtures by the KNN and SIMCA methods. The single output from KNN indicates that only one major component in eight blend mixtures can be identified without any mis-identification cases. However, the limitation of KNN in mixture analysis is that it is not possible to detect whether a sample is a pure component or a mixture. Among those outcomes, synthesized ternary mixture S1 was the only successful case in a multiple-class identification by SIMCA. As shown in Table 2, based on the Matthews' correlation coefficient, the overall prediction performance of KNN ($C_M = 0.51$) was slightly better than SIMCA ($C_M = 0.42$) since the latter one has more mis-classification rates in this study.

The values obtained from each output serve as an index of "class probability" which only indicates the presence or absence information in the mixture. For our ANN model, the presence/absence profile was set to 0.9/0.1 to cope with the range of non-linear sigmoid function. To make a firm decision about the number of main components in the samples, the "positive presence" threshold was assigned to 0.15. In this evaluation, seven out of eight mixtures were identified correctly using the ANN method. Only one mixture from another crop year was over-identified, i.e. one non-existing component was picked up by ANN. The complexity of chemical composition from other sources of raw material may seriously handicap interpreting the results obtained. However, the identification of eight mixtures from DPLS gave more "ambiguous identification" results. Only five out of eight mixtures could be distinguished completely.

With regard to the overall performance of mixture analysis, the ANN method was able to effectively classify most mixture samples, with higher Matthews' coefficient ($C_M = 0.96$) and good generalizing capabilities as compared with other methods. This encouraging result suggests that ANN should be a robust technique in mixture identification, as well as it being applied to the classification and non-linear function mapping. When coupled with rapid NIR measurement, both DPLS and ANN methods provide a simple concept in the identification of components from a mixture. Traditional classification algorithms, such as KNN and SIMCA, could not confirm the sample composition or lack of quantitative results. The use of DPLS is straightforward and there

Table 2. The overall prediction coefficients.

Method	Matthews' correlation coefficient
KNN	0.51
SIMCA	0.42
DPLS	0.80
ANN	0.96

Table 3. Root-mean-square errors for tobacco blends analysis.

	Training data set (44 spectra) RMS errors (%w/w)		Validation data set (36 spectra) RMS errors (%w/w)	
	PLS model	ANN model	PLS model	ANN model
A1	2.19	1.19	3.10	2.24
B1	3.65	2.03	4.40	3.27
C1	3.55	1.61	2.79	2.01

is only one parameter (number of latent variables) that needs to be selected. The drawback for DPLS is its theory and implementation are not well described in the statistical literature.

Quantitative analysis of blending ratio

Table 3 gives the weight percentage RMS errors on the predictions produced by PLS and ANN on both training and validation sets of tertiary mixtures. In PLS models, the RMS errors of the training set were from 2.19 to 3.65%, while RMS errors from the validation set ranged from 2.79 to 4.40%. The prediction results obtained by using a sigmoid/linear ANN method based on 12 principal components gave a better performance. The average RMS errors from the ANN model in the training and validation sets were 1.61 and 2.50%, respectively. These results reveal that the ANN method provided at least 44% improvement in the training set and about 27% improvement in the validation set than the PLS calibration. It is clear that the non-linear spectral effects were resolved by the ANN method. The reason may be due to the relatively non-homogeneous character of the tobaccos, though the wide distribution of texture or particle size tended to cause non-linear spectral responses.

Summary

In this paper we report the result of our investigations of mixture analysis based on the NIR and advanced multivariate data analysis. The qualitative and quantitative analyses of tobacco mixtures with near infrared spectroscopic measurements were demonstrated. Our results clearly indicate that simple ANN or DPLS methods with suitable design enable the classification and quantitation of a spectral mixture within reasonable error ranges. The following conclusions are drawn: (i) ANN provides better discriminant power and generalization in the identification of individual components from mixtures and (ii) The tobacco blending ratio can be estimated within 2.5% (w/w) average error using non-linear ANN calibration technique. Thus, the method developed here may be a potential means to analyze tobacco mixtures under blending operations for final quality verification.

References

1. L.M. Dominguez and S.K. Seymour, in *Making Light Work: Advances in Near Infrared Spectroscopy*, Ed by I. Murray and I.A. Cowe. VCH, Weinheim, pp. 179 (1992).
2. *Near Infrared Technology in the Agricultural and Food Industries*, Ed by P. Williams and K. Norris. Amer. Assoc. Cereal Chem., Inc., St Paul, MN (1987).
3. W.F. McClure, *Anal. Chem.* **66,** 43A (1994).
4. S. Lo and C.W. Brown, *Appl. Spectrosc.* **46,** 790 (1992).
5. H. Hobert and K. Meyer, *Fresenius J. Anal. Chem.* **344,** 178 (1992).

6. T.M. Cover and P.E. Hart, *IEEE Trans. Inform. Theory* **IT-13,** 21 (1967).
7. S. Wold, *Pattern Recognition* **8,** 127 (1976).
8. R. Vong, P. Geladi, S. Wold and K. Esbensen, *J. Chemometrics* **2,** 281 (1988).
9. T.B. Blank and S.D. Brown, *Anal. Chem.* **65,** 3081 (1993).
10. J.A. Burns and G.W. Whitesides, *Chem. Rev.* **93,** 2583 (1993).
11. J.A. Cornell, *Experiments with Mixtures: Designs, Models and the Analysis of Mixture Data,* 2nd Edn. John Wiley & Sons, New York (1990).
12. B.W. Matthews, *Biochim. Biophys. Acta* **405,** 442 (1975).
13. M.T. Hagan and M.B. Menhaj, *IEEE Trans. Neural Networks* **5,** 989 (1994).