Journal of Food and Drug Analysis, Vol. 19, No. 1, 2011, Pages 9-17 藥物食品分析 第十九卷 第一期

A Robust Identification Model for Herbal Medicine Using Near Infrared Spectroscopy and Artificial Neural Network

CI-WEN YANG[1,3], SUMING CHEN[1,2]*, FU OUYANG[1], I-CHANG YANG[1,2] AND CHAO-YIN TSAI[1,2]

[1.] Department of Bio-Industrial Mechatronics Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei,10617 Taiwan, R.O.C.

[2.] Bioenergy Research Center, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan, R.O.C.

[3.] Taipei Zoo, No.30, Sec.2, Xinguang Road, Taipei, 11656, Taiwan, R.O.C.

(Received: June 28, 2010; Accepted: January 4, 2011)

ABSTRACT

A robust identification model for herbal medicine was developed by combining near-infrared spectroscopy (NIR) and artificial neural network (ANN) to discriminate raw materials of herbal medicine, which are often similar in appearance and practically impossible to identify by visual inspection alone. The identification by chemical methods is usually higher in cost and lower in efficiency. Compared with other modern inspection methods, NIR is an alternative, which is non-destructive, rapid, and easy to operate. In this study, we employed ANN to analyze the absorption spectra of herbal medicines and successfully built an identification model, which is able to identify 30 different herbal medicines. The best identification model can reach a correct identification rate (CIR) of 99.67% when applied to a training set of 600 samples, and 100% CIR when applied to a test set of 300 samples.

Key words: near infrared (NIR) spectroscopy, artificial neural network (ANN), herbal medicine

INTRODUCTION

Identification of herbal medicines plays an important role in their processing and usage. Lots of plants can be used as raw materials for herbal medicines, but some of them, even not related species[(1)], are similar in appearance, and their storage as powder for transportation makes their appearance a useless indicator for identification. The promotion of herbal medicines in recent years, has made it necessary and urgent to analyze them by qualitative and quantitative inspection methods.

Common inspection methods of herbal medicines can be roughly divided into sensory and analytical inspection. The former includes the morphological inspection and histological technique. Although the morphological inspection is direct and simple, its accuracy heavily depends on the inspectors, who are somewhat subjective. The histological analysis can reveal the characteristics of the structure and the arrangement of tissue and cells, but it is not capable of identifying related species, which may share similar histological characteristics. Analytical inspection methods include thin layer chromatography (TLC)[(2)], gas chromatography (GC), high performance liquid chromatography (HPLC)[(3)], capillary electrophoresis (CE)[(4)], etc. These methods are accurate for chemical constituents analysis, but the complicated sample preparation procedures and long inspection time make them impossible for online inspection. Besides, they are destructive methods, thus inevitably damaging and consuming samples.

Near infrared (NIR) spectroscopy is also an analytical inspection method, but its advantages, such as non-destructive inspection, rapid measurement, simple operation, and less or no sample preparation, make it an ideal alternative for the aforementioned methods. It has been extensively adopted by bio-related industries such as agriculture[(5)], pharmacy[(6)], etc., and there were some applications for herbal medicines as well. Most of them were focused on identifying related species[(7)], but those applications only dealt with a small number of medicines[(1)]. There was some success using artificial neural network (ANN) to analyze NIR spectra for the identification of bio-materials, for example, the classification of damaged soybean seeds[(8)], discrimination of varieties of

* Author for correspondence. Tel:+886-2-3366-5350; Fax:+886-2-2362-7620; E-mail: schen@ntu.edu.tw

green tea[9], and the discrimination of Chinese bayberry varieties[10]. In this study, the spectra of 900 samples from 30 different medicines was measured on a NIR spectrometer, and then pre-processed by the standard normal variate (SNV) transformation before the spectra analysis. Two thirds of the spectra were selected, by the Kennard & Stone (KS) algorithm, as the training set and the remaining third was utilized as the test set. The spectra of the training set were compressed by principal component analysis (PCA), and then sent into the networks to start training and to reach the best identification model. Finally, the test set was used to validate the performance of the identification models.

The purpose of this study was to provide herbal medicine practitioners and researchers with a reliable identification method, which takes advantage of the non–destructive and rapid NIR spectroscopy inspection and ANN's high level of predictability to improve the efficiency of herbal medicines identification.

MATERIALS AND METHODS

I. *Materials and Apparatus*

A total of 900 samples from 30 different herbal medicine powders were provided by Sun Ten Pharmaceutical Co., Ltd. Each medicine had 30 different samples, which were loaded in 20 mL vials. For the purpose of analysis, the data sets were designed as follows: set A, set B, set C, set D, and set E. The first four data sets contained no samples that were the same, and the medicines in each set were randomly selected.

(I) Set A contained 150 samples of 5 medicines: *Citrus Undeveloped Exocarpium, Amomi Semen, Curcumae Radix, Achyranthis Radix,* and *Cyperus rotundus.*

(II) Set B contained 150 samples of 5 medicines: *Atractylodis Rhizoma, Pinelliae Tuber, Zingiberis Siccatum Rhizoma, Ephedrae Herba,* and *Evodiae Fructus.*

(III) Set C contained 150 samples of 5 medicines: *Perillae Folium, Saposhinkoviae Radix, Cinnamomi Ramulus, Bupleuri Radix,* and *Puerariae Radix.*

(IV) Set D contained 300 samples of 10 medicines: *Magnoliae Flos, Rhei Rhizoma, Polyporus, Clematidis Radix, Nelumbinis Folum, Anglicae Sinensis Radix, Ligustici Rhizoma, Platycodi Radix, Citrus Sinensis Exocarpium,* and *Saussureae Radix.*

(V) Set E contained 900 samples of 30 medicines, and it was composed of *Scutellariae Radix, Paeoniae Lactiflorae Radix, Hoelen, Salivae Miltiorrhizae Radix, Paeoniae Veitchii Radix,* set A, set B, set C, and set D in sequence.

The spectra of the above data sets were measured on a FOSS NIRSystems instrument Model 6500 NIR reflectance spectrometer configured with a rapid content analyzer (RCA) module. The spectrometer is equipped with a tungsten halogen lamp as a light source, and samples were scanned at 2 nm intervals in the range of 400 to 2498 nm, which encompassed visible and NIR wavelengths. Silicon detectors were used below 1100 nm and succeeded by lead sulfide detectors. Each spectrum had 1050 data points that were acquired using the software Vision 3.0 (FOSS NIRSystems). The mean absorption spectra of each medicine are shown in Figure 1 in which every spectrum represents the average of 30 samples of a medicine. The spectra were analyzed and the identification model was built using MATLAB 6.51 as the platform to run its Statistics Toolbox 4.1, Neural Network Toolbox 4.1, and other extended programs developed by our research team.

II. *Methods*

First, the original spectra were transformed by SNV transformation, followed by selection of two thirds of the samples from each set, by Kennard & Stone (KS) algorithm, as the training set while the remaining as the test set. After PCA, the number of variables of the training set was reduced from 1050 to a smaller number. ANN was then applied to build the identification model using the training set. The PCA-transformed test set was then fed into the network to verify the performance of the identification model by computing the correct identification rate (CIR) (Eq.1). The steps involved in the analysis will be illustrated in detail in the following sections.

$$CIR = \frac{number\ of\ samples\ correctly\ identified}{number\ of\ all\ samples} \times 100\% \quad (1)$$

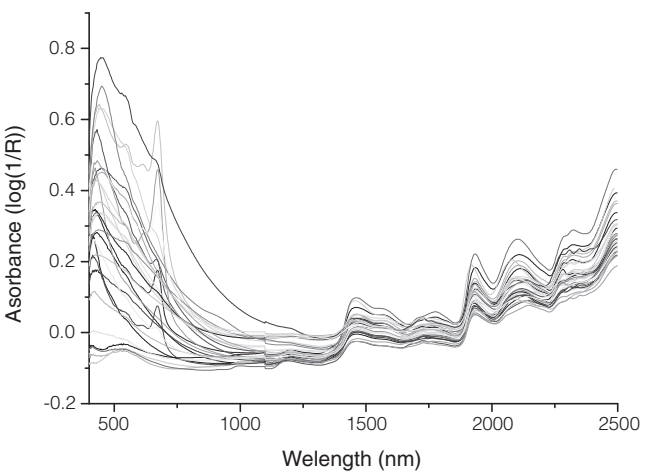

Figure 1. The mean absorption spectra of 30 medicines: each spectrum represents a medicine, and was calculated from all of its 30 samples.

Journal of Food and Drug Analysis, Vol. 19, No. 1, 2011

(I) *Pretreatment: SNV*

The absorption value of the NIR spectra are influenced by particle size in powder samples, which is an unfavorable factor during analyzing the content of certain chemical components in the samples from their spectra. SNV is a strategy which can deal with the problem[11], and the calculation of SNV is followed by Eq.2 :

$$y_{SNV} = \frac{(y - y_{mean})}{\sqrt{\frac{\sum (y - y_{mean})^2}{n-1}}} \quad (2)$$

Where y is a spectrum of any sample, n is the number of data points in this spectrum, y_{mean} is the mean of the n log 1/R values of y, y_{SNV} is the SNV transformed spectrum. For example, Figure 2 shows the original spectra of set A, and Figure 3 shows the SNV-transformed spectra of set A. It is obvious that after the SNV transformation, the spectra for each medicine exhibit a tighter distribution pattern.

(II) *Training Set Selection: The KS Algorithm*

This algorithm was first proposed by Kennard and Stone for experimental designs[12]. In our study, the algorithm was employed to select the most representative samples for the training set. The algorithm starts by finding the two most representative samples, and then searching for the third, fourth, etc. according to the following equation (Eq.3) until the number of the selected samples meets the goal:

$$d(j) = \min_{k=1}^{sn}[\sum ({}^{m}C(j,:) - {}^{m}T(k,:))^2]^{0.5} \quad (3)$$

The subset ${}^{m}T$ contains the selected samples, and the parameter sn is the number of the samples in the subset ${}^{m}T$. First, the algorithm selects a pair of spectra that are the farthest apart from each other in the subset ${}^{m}C$, which contains all the samples, and moves them into ${}^{m}T$. The spectra left in ${}^{m}C$ will be computed using the above equation and the spectrum with the largest d value will be moved into ${}^{m}T$, and a new iteration begins. These samples in ${}^{m}T$ best represent the distribution of the whole data set.

(III) *Variable Reduction: Principle Component Analysis*

PCA is aimed to transform the original data space into an orthogonal one. After PCA, most of the variance will be carried by a small number of variables. The variable that carries the most variance is referred to as PC 1, the variable that carries the second most variance is referred to as PC 2, and so forth. By keeping the PCs with more variance and ignoring the PCs with less variance, the number of variables is thus reduced.[13]

(IV) *Identification Model: Artificial Neural Network*

The multilayer perceptron (MLP) was utilized to establish the artificial neural network in this study. The network was composed of an input layer, a hidden layer, and an output layer, and we used a back-propagation algorithm, the Levenberg-Marquardt algorithm, to train the network. The number of nodes in the input layer was conformed to the number of PCs we kept after PCA. The number of nodes in the output layer was decided by the number of medicines to be identified. For instance, set A was composed of 5 medicines and its network had 5 nodes in its output layer. Set D was composed of 10 medicines and its network had 10 nodes in its output layer. The number of nodes in the hidden layer is abbreviated as *nh*, and its influence on the performance of the

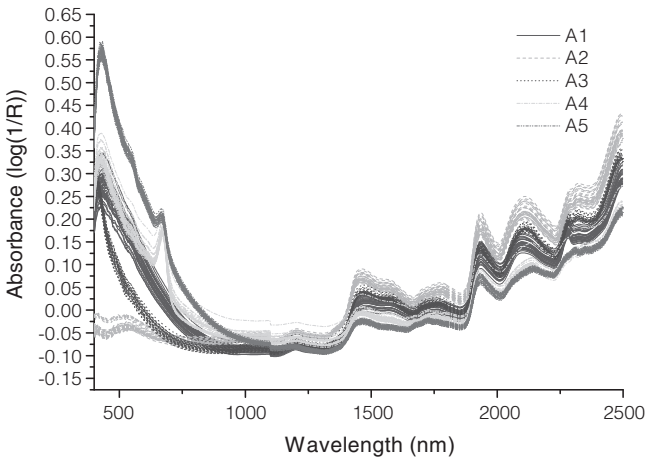

Figure 2. The original absorption spectra of the 150 samples in set A, in which a spectrum represents a sample and each medicine has 30 samples. A1 is *Citrus Undeveloped Exocarpium*, A2 is *Amomi Semen*, A3 is *Curcumae Radix*, A4 is *Achyranthis Radix*, and A5 is *Cyperus rotundus*.

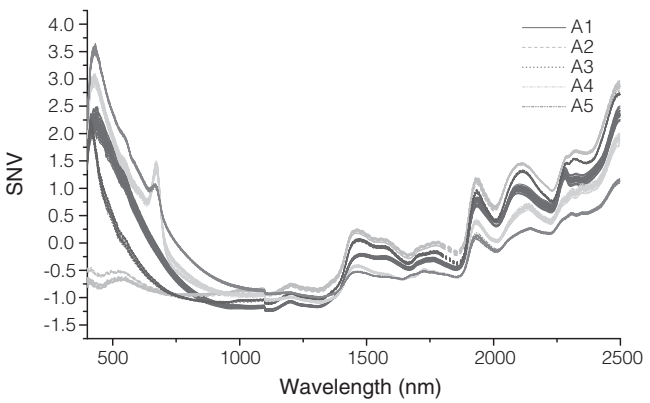

Figure 3. The SNV transformed spectra of the 150 samples in set A, in which a spectrum represents a sample and each medicine has 30 samples. A1 is *Citrus Undeveloped Exocarpium*, A2 is *Amomi Semen*, A3 is *Curcumae Radix*, A4 is *Achyranthis Radix*, and A5 is *Cyperus rotundus*.

identification model will be discussed in the next section.

The targets, which are the ideal output vectors of the networks, are designed as follows: if there are *k* different medicines, the targets of the output vectors should be in *k* dimensions. In each vector there is only one 1-value entry accompanied by *k*-1 0-value entries. For example, if a network is designed for identifying 3 medicines, the output vectors of the network should be in a 3-dimensional space, and there will be only 3 possible targets, which are [1, 0, 0], [0, 1, 0], and [0, 0, 1] according to three different medicines.

In Figure 4, the left gray rectangle is the input vector with "*ni*" input factors (3 × 1 matrix). The factor numbers of the hidden layer and the output layer are "*nh*" and "*nj*". The IW means the input weight matrix containing the weights of the connection between the input layer and the first layer, and b{1} is the bias of the first layer and so on. The LW{2,1} is the weight of the first layer output to the second layer. The perceptron in this research is a single layer network. The variable b{2} means the bias of the second layer.

The other important factors and parameters used for building networks are listed below: the transfer function of the hidden layer was the hyperbolic tangent sigmoid transfer function (Eq. 4), a_{hl} the output vector of the hidden layer and x_{hl} the input vector of the hidden layer. The transfer function of the output layer was the linear transfer function (Eq. 5), a_{ol} the output vector of the output layer and x_{ol} the input vector of the output layer. The goal of the training process is to minimize the mean squared error (MSE) (F value in Eq. 6), in which N is the number of nodes in the output layer, a and t are output and target vectors. The learning rate is fixed at 0.05, and the maximum training epoch is 1000.

$$a_{hl} = \frac{e^{x_{hl}} - e^{-x_{hl}}}{e^{x_{hl}} + e^{-x_{hl}}} \tag{4}$$

$$a_{ol} = x_{ol} \tag{5}$$

$$F = \frac{1}{N}\sum_{i=1}^{N}(t_i - a_i)^2 \tag{6}$$

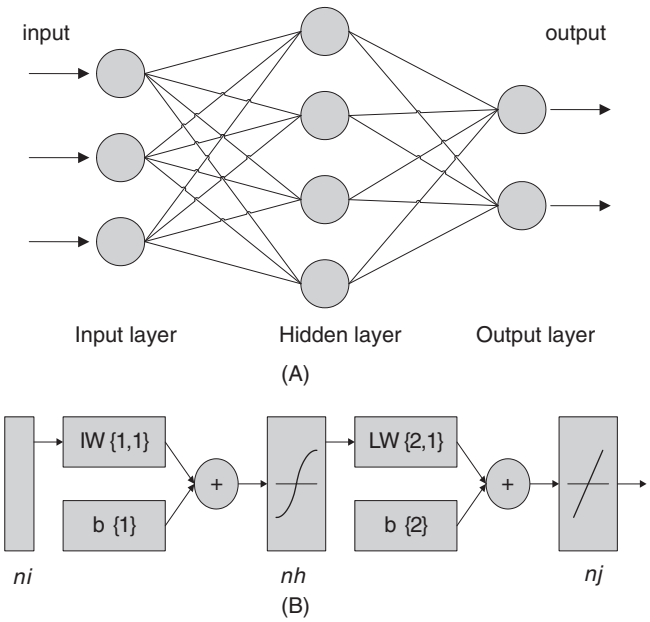

Figure 4. (A) Diagram of the neural network in this research including the input layer, the hidden layer, and the output layer. (B) The graphical mathematics model of the neural network.

RESULTS AND DISCUSSION

When applying PCA to different data sets, the amount of information contained in the same number of PCs is different. Table 1 shows that in sets A, B, and C, the accumulative variance contained in the first 3 PCs was near 99.0%. However, in sets D and E, which had more medicines, the accumulative variance was only about 96.6%. Although some difference did exist, the accumulative variance contained in the first 3 PCs of each set is quite large (above 95%), and the fourth PC contains only a little variance (below 2%), compared with the first three. For ease of comparing different parameters, like *nh*, we only adopt the first 3 PCs for analysis.

Table 1. The individual and accumulative variance of the first 5 principal components of set A (5 medicines), set B (5 medicines), set C (5 medicines), set D (10 medicines), and set E (30 medicines)

PC	Set A		Set B		Set C		Set D		Set E	
	indiv.	accumu.	indiv.	accumu.	indiv.	accumu.	indiv.	accumu.	indiv.	accumu.
1	90.8%	90.8%	92.3%	92.3%	87.5%	87.5%	89.6%	89.6%	88.2%	88.2%
2	4.8%	95.6%	5.4%	97.7%	9.1%	96.6%	4.1%	93.7%	6.2%	94.4%
3	3.3%	98.9%	1.3%	98.9%	2.3%	99.0%	2.9%	96.6%	2.2%	96.6%
4	0.8%	99.7%	1.0%	99.9%	0.8%	99.8%	2.0%	98.6%	1.6%	98.2%
5	0.3%	100.0%	0.0%	100.0%	0.1%	99.9%	0.7%	99.3%	0.8%	98.9%

This study explored the performance of the identification model while dealing with different medicines (set A, set B, set C) and different numbers of medicines per set (5: sets A, B, and C; 10: set D; 30: set E). During the analysis of all the data sets, the only adjustable variable was *nh*, but since the initial value of the weights and biases in the networks were randomly generated, this made the networks of a same *nh* different. To verify the networks' stability, we built 5 models for every *nh*.

I. *Results of Sets A, B, and C (5 Medicines)*

The data sets of 5 medicines were examined in this section, and there were no identical sample in sets A, B, or C. Since no generalized or reliable searching algorithm is available for the determination of *nh*, in most of the cases, the trial and error method is the most common solution. In this study, the initial value of *nh* was set equal to the number of output nodes, and then it was decreased gradually to determine the optimal value.

The results of the set A analysis is shown in Table 2. While *nh* = 5 and *nh* = 4, all of the 10 networks reached 100% CIRs. After that, the CIRs decreased with the *nh*. It can be seen that a higher *nh* did help improve the CIRs to some extent, but the contribution of *nh* stopped after a certain number, after which higher *nh* could only increase the computation time and the complexity of networks. Besides, the reason of the CIRs' uniform decrement is as follows: all samples of a medicine were misidentified with *nh* = 3, all samples of 2 medicines were misidentified with *nh* = 2; and all samples of 3 medicines were misidentified with *nh* = 1. In the 25 models built for set A (every *nh* has 5 networks), every medicine in set A could possibly be misidentified, and there was no medicine which was especially easy to be misidentified. For a network with *nh* = 3, where the CIR was 80%, all of the samples of *Cyperus rotundus* were misidentified as *Amomi Semen*. In the PCA score plot of set A (Figure 5), the distribution of all data points was clearly represented. Each cluster was formed by a certain medicine, and there was no overlap between any two of them. Misidentification in this 5-medicine case was caused by the model's clustering of *Cyperus rotundus* and *Amomi Semen* as one; thus identification of all *Cyperus rotundus* samples failed.

The results of sets B and C were exactly the same as set A. As the PCA score plot of their training sets (Figure

Table 2. The correct identification rates (CIR) of all the networks built for set A, which has 100 samples in its training set and 50 samples in the test set. Set B and set C have exactly the same results as set A has

nh	no.	CIR (%)	
		Training (100)	Test (50)
5	1~5	100.00	100.00
4	1~5	100.00	100.00
3	1~5	80.00	80.00
2	1~5	60.00	60.00
1	1~5	40.00	40.00

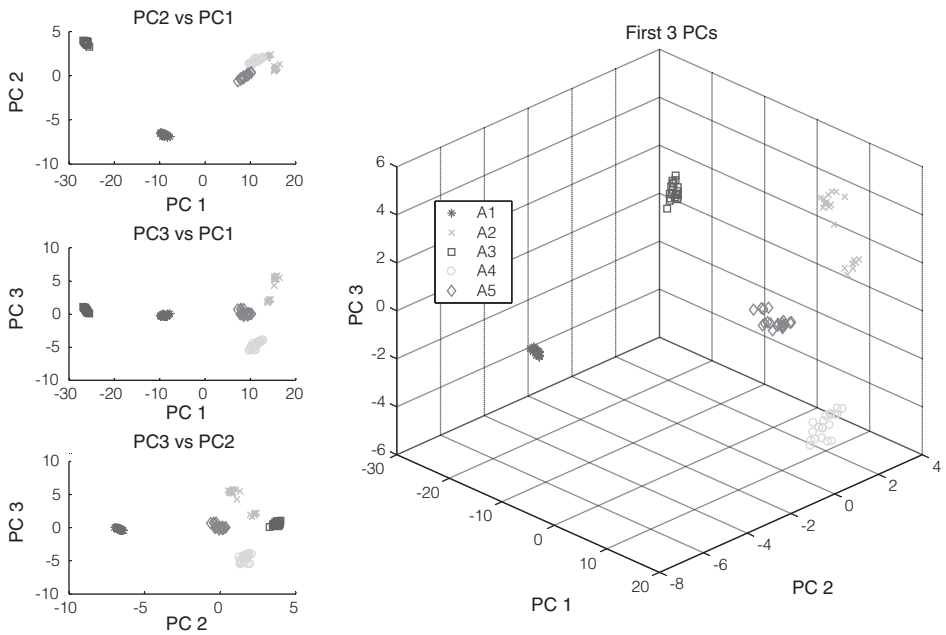

Figure 5. The PCA score plot of the 100 samples in the training set of set A, in which each medicine has 20 samples. A1 is *Citrus Undeveloped Exocarpium*, A2 is *Amomi Semen*, A3 is *Curcumae Radix*, A4 is *Achyranthis Radix*, and A5 is *Cyperus rotundus*.

6 and Figure 7) showed, there was no overlap which is why their results were the same as set A. This proved that in this study the performance of the networks was mainly related to *nh* and had nothing to do with the characteristics of the medicines. As a whole, the optimal network for 5 medicines is 3 × 4 × 5 (the number of nodes of each layer in sequence: input layer × hidden layer × output layer).

II. *Results of Set D (10 Medicines)*

The analysis process for set D is the same as the process for set A. First, we let *nh* = 10, which equaled the number of nodes in the output layer, then we pruned it gradually. As the results showed in Table 3, the CIRs were 100% when *nh* = 10 and 9; when *nh* = 8, it was possible

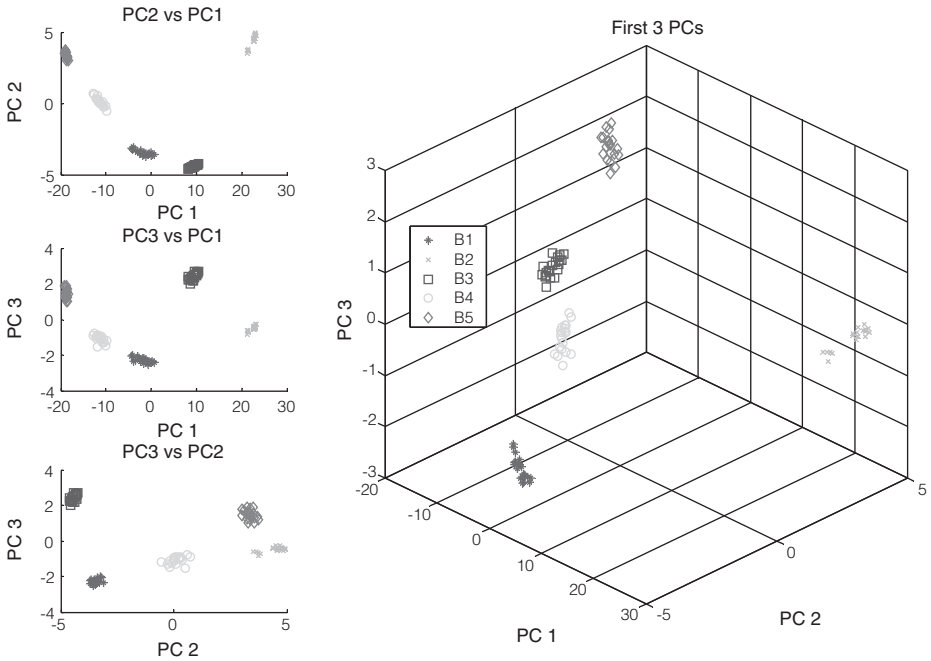

Figure 6. The PCA score plot of the 100 samples in the training set of set B, in which each medicine has 20 samples. B1 is *Atractylodis Rhizoma*, B2 is *Pinelliae Tuber*, B3 is *Zingiberis Siccatum Rhizoma*, B4 is *Ephedrae Herba*, and B5 is *Evodiae Fructus*.

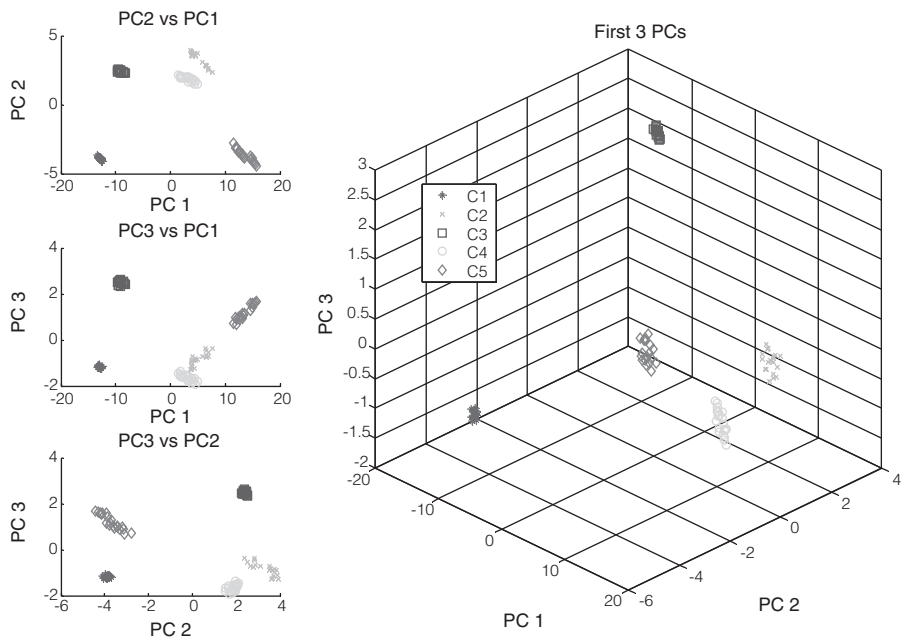

Figure 7. The PCA score plot of the 100 samples in the training set of set C, in which each medicine has 20 samples. C1 is the *Perillae Folium*, C2 is the *Saposhinkoviae Radix*, C3 is the *Cinnamomi Ramulus*, C4 is the *Bupleuri Radix*, and C5 is the *Puerariae Radix*.

to misidentify all samples in one medicine; when *nh* = 7, misidentification of 2 medicines was the worst; when *nh* = 6, there were no longer 100% CIRs and the worst model holds only a 70% CIR. In the PCA score plot for the training set of set D (Figure 8), there was no overlap between any two medicines as well, so we could infer that the reason of the misidentification of all samples in a medicine was the same as the 5-medicine case. The greatest difference between the former section and this one lies when *nh* = 7 or 8, when the networks were still able to reach 100% CIR, and the performance of the networks did not descend in a specific interval when *nh* decreased. This was related to the randomly generated nature of the initial values of the weights and biases. Since all of the networks were different at the beginning, their performance varied when the total nodes in the networks rose. For this reason, the optimal network structure should be 3 × 9 × 10 to reach a more stable and precise model.

III. *Results of Set E (30 Medicines)*

For set E, we let *nh* = 30 at the beginning, and then we pruned it gradually. Compared with Tables 2 and 3, the most obvious difference with the results shown in set E listed in Table 4 was that misidentification of all samples in a medicine no longer occurred. The training (600 samples) and test (300 samples) sets here were larger in size than the former cases. In the PCA score plot for the training set in set E (Figure 9), the distribution of the data points was very complicated. Most of the clusters were very close to each other or even overlapped. Therefore, it made sense that the data-fitting capability of the

Table 3. The correct identification rates (CIR) of all the networks built for set D, which has 200 samples in its training set and 100 samples in its test set

nh	no.	CIR (%)	
		Training (200)	Test (100)
10	1~5	100.00	100.00
9	1~5	100.00	100.00
8	1	100.00	100.00
	2	100.00	100.00
	3	90.00	90.00
	4	100.00	100.00
	5	90.00	90.00
7	1	80.00	80.00
	2	80.00	80.00
	3	80.00	80.00
	4	100.00	100.00
	5	90.00	90.00
6	1	80.00	80.00
	2	70.00	70.00
	3	70.00	70.00
	4	70.00	70.00
	5	80.00	80.00

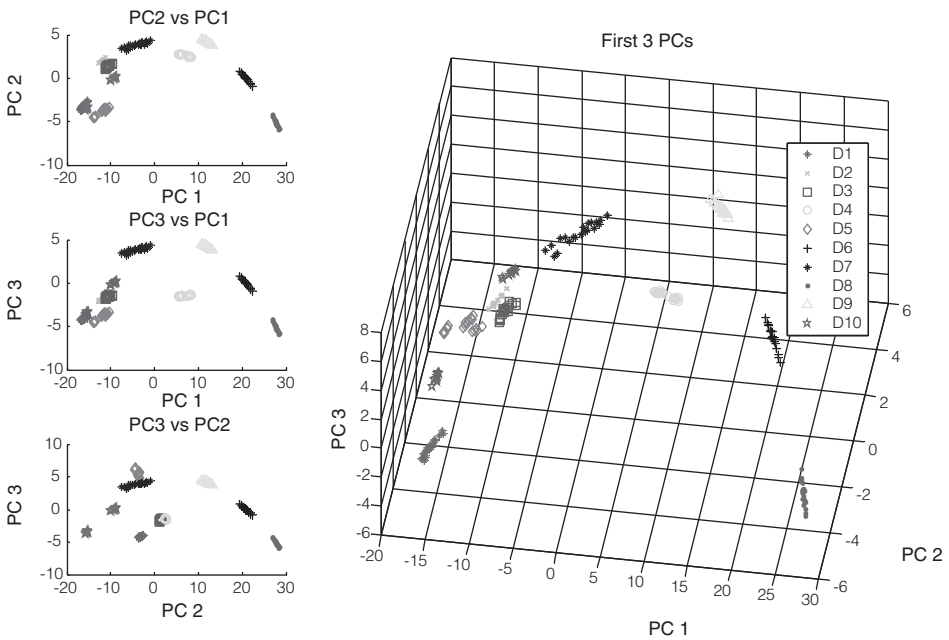

Figure 8. The PCA score plot of the 200 samples in the training set of set D, in which each medicine has 20 samples. D1 is *Magnoliae Flos*, D2 is *Rhei Rhizoma*, D3 is *Polyporus*, D4 is *Clematidis Radix*, D5 is *Nelumbinis Folum*, D6 is *Anglicae Sinensis Radix*, D7 is *Ligustici Rhizoma*, D8 is *Platycodi Radix*, D9 is *Citrus Sinensis Exocarpium*, and D10 is *Saussureae Radix*.

Table 4. The correct identification rates (CIR) of all the networks built for set E, which has 600 samples in its training set and 300 samples in its test set

nh	no.	CIR (%)		*nh*	no.	CIR (%)	
		Training (600)	Test (300)			Training (600)	Test (300)
32	1	98.17	98.67	29	1	96.17	95.67
	2	99.17	99.67		2	98.00	97.67
	3	98.67	98.33		3	93.17	93.33
	4	96.50	96.00		4	96.83	98.33
	5	96.50	99.67		5	99.17	98.33
31	1	98.67	98.00	28	1	96.83	97.00
	2	98.67	97.67		2	94.17	94.67
	3	99.00	98.33		3	94.67	94.33
	4	98.67	97.67		4	97.17	96.33
	5	96.00	96.00		5	93.17	93.33
30	1	98.33	98.00				
	2	99.83	99.67				
	3	97.33	96.00				
	4	99.67	100.00				
	5	99.67	98.33				

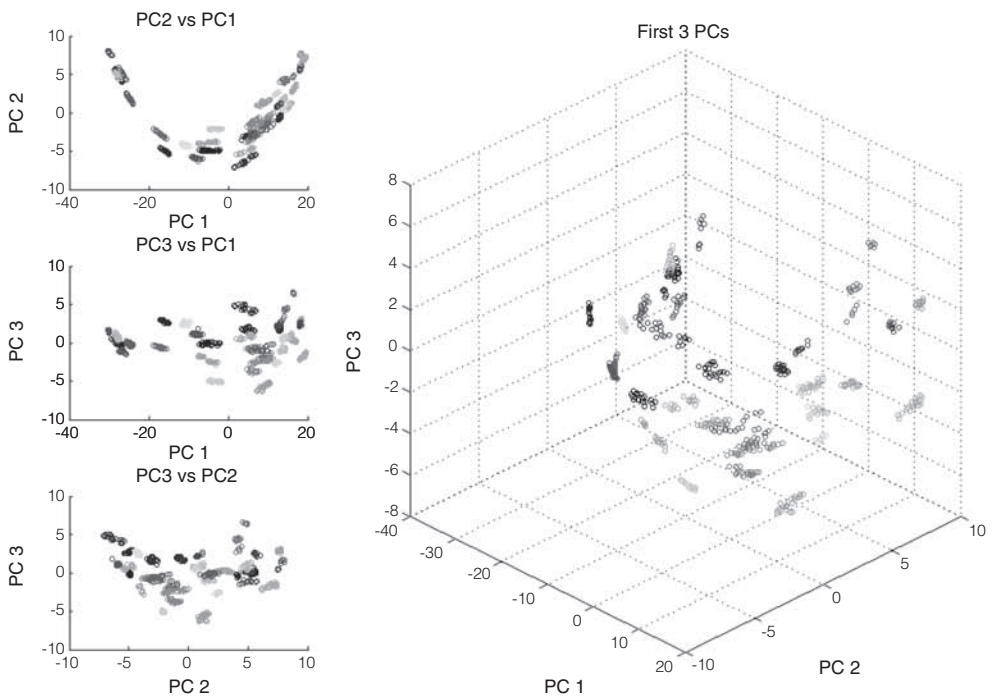

Figure 9. The PCA score plot of the 600 samples in the training set of set E, in which each medicine has 20 samples.

networks was not as perfect as in the cases of the small data sets, and the existence of partial misidentification rather than all the samples in a medicine was reasonable. For instance, in the no. 4 model built with *nh* = 29, there were 19 misidentified samples; 9 from the 2nd medicine, 3 from the 7th medicine, 1 from the 8th medicine, 1 from the 21st medicine and 5 from the 27th medicine. In the test set there were 5 misidentified samples; 1 from the 2nd medicine, 1 from the 7th medicine and 3 from the 27th medicine. The best network with the highest performance was the 4th one with *nh* = 30. The CIR of its training set was 99.67% (2 misidentification) while the CIR of the test set was 100%. After careful examinations, it could be concluded that those medicines with the most misidentified samples had more similarity in their spectra pattern/characteristics or in successive analytical assessments.

Since not every network of *nh* = 30 could reach a 100% CIR, *nh* was increased to 31 and 32 to search for better identification models. However, the result showed that there was no apparent improvement in the performance, as the average CIR remained about 98%. Consequently, the optimal network structure should be 3 × 30 × 30 which is a compromised balance of performance and the complexity of structure.

The identification model in Woo *et al.* (1999) was able to identify 3 medicines of non-related species with a 100% CIR[1]. In this study the best model could reach 100% CIRs when applied to 5, 10, and 30 medicines. The breakthrough in the number of medicines showed that ANN is a promising method which could be applied to the identification of more herbal medicines. Although an increase in the number of herbal medicines in the database would take more computing time and setting effort to reach the best identification models, the strategy of *nh* selection provided in this study could simplify the whole process. An acceptable *nh* could be found immediately by a small number of trials, and then the best *nh* could be determined.

CONCLUSIONS

This study successfully built an identification model for herbal medicines by NIR spectroscopy and ANN, and the model identified 30 medicines with a 100% CIR when applied to the test set. The study also provided a strategy for the determination of the number of nodes in the hidden layer, to shorten the expenditure of time in setting network structures. In the future, it is possible to increase the number of medicines in the herbal medicine database, and to practically apply NIR spectroscopy for quality assurance in the herbal medicine industry.

ACKNOWLEDGMENTS

The authors thank the support from Sun Ten Pharmaceutical Co., Ltd, which is the provider of all the herbal medicines, and FOSS NIRSystems, Inc, which is the supplier of the Model 6500 NIR reflectance spectrometer configured with a rapid content analyzer (RCA) module.

REFERENCES

1. Woo, Y., Kim, H. and Cho, J. 1999. Identification of herbal medicines using pattern recognition techniques with near-infrared reflectance spectra. Microchem J. 63: 61-70.
2. Simonvska, B., Vovk, I., Andrensek, S., Valentova, K. and Ulrichova, J. 2003. Investigation of phenolic acids in yacon (*Smallanthus sonchifolius*) leaves and tubers. J. Chromatogr. A 1016: 89-98.
3. Chuang, C. C., Su, C. H., Huang, W. Y. and Sheu, S. J. 2008. Classification of *Fangchi Radix* Samples by multivariate Analysis. J. Food Drug Anal. 16:48-56.
4. Yang, J. J., Long, H., Liu, H. W., Huang, A. J. and Sun, Y. L. 1998. Analysis of terandrine of fangchinoline in traditional Chinese medicines by capillary electrophoresis. J. Chromatogr. A 811: 274-279.
5. Wang, D., Dowell, F. E. and Lacey, R. E. 1999. Single wheat kernel color classification by using near-infrared reflectance spectra. Cereal Chem. 76: 30-33.
6. Candolfi, A., De Maesschalck, R., Massart, D. L., Hailey, P.A. and Harrington, A. C. E. 1999. Identification of pharmaceutical excipients using NIR spectroscopy and SIMCA. J. Pharm. Biomed. Anal. 19: 923-935.
7. Zhang, Z Y., Wang, Y. M., Fan, G. Q. and Harrington, P. D. B. 2007. A comparative study of multilayer perceptron neural networks for the identification of rhubarb samples. Phytochem. Anal. 18: 109-114.
8. Wang, D., Ram, M. S. and Dowell, F. 2002. Classification of damaged soybean seeds using near-infrared spectroscopy. Trans. ASAE 45: 1943-1948.
9. He, Y., Li, X. L. and Deng, X. F. 2007. Discrimination of varieties of tea using near infrared spectroscopy by principal component analysis and BP model. J. Food Process Eng. 79: 1238-1242.
10. Li, X. L., He, Y. and Fang, H. 2007. Non-destructive discrimination of Chinese bayberry varieties using Vis/NIR spectroscopy. J. Food Process Eng. 81: 357-363.
11. Barnes, R. J., Dhanoa, M. S. and Lister, S. J. 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra transformation and de-trending of near-infrared diffuse reflectance spectra. Appl. Spectrosc. 43: 772-777.
12. Kennard, R. W. and Stone, L. A. 1969. Computer aided design of experiments. Technometrics 11: 137-148.
13. Jachson, J. E. 1991. A user's guide to principal components. Wiley. New York, U.S.A.