Journal of the Taiwan Institute of Chemical Engineers 43 (2012) 566–572

Contents lists available at SciVerse ScienceDirect

Journal of the Taiwan Institute of Chemical Engineers

journal homepage: www.elsevier.com/locate/jtice

Adaptive neuron-fuzzy inference system combined with principal components analysis for determination of compound thiamphenicol powder on near-infrared spectroscopy

Nan Qu [a], Mingchao Zhu [b], Yulin Ren [c], Sen Dou [a,*]

[a] College of Resources and Environmental Sciences, Jilin Agricultural University, Changchun 130118, China
[b] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
[c] College of Chemistry, Jilin University, Changchun 130021, China

ARTICLE INFO

Article history:
Received 4 September 2011
Received in revised form 9 January 2012
Accepted 29 January 2012
Available online 3 March 2012

Keywords:
Near-infrared spectroscopy
Principal component analysis
Adaptive neuron-fuzzy inference system
Compound thiamphenicol powder

ABSTRACT

The adaptive neuron-fuzzy inference system (ANFIS) is an effective modeling tool developed recently. It has gained much interest in solving classification and function approximation. In this paper, a new application based on ANFIS was presented for nondestructive determination of thiamphenicol powder drug with near-infrared (NIR) spectroscopy. The principal component analysis (PCA) technique was applied to extract relevant features from a number of spectral data in order to reduce the input variables of the ANFIS. The generated scores of the principal components (PCs) subsequently were used as the input variables of the ANFIS instead of the spectra data and constituted the principal component analysis-adaptive neuron-fuzzy inference system (PCA-ANFIS) model. A hybrid-learning algorithm which combined the least squares method and the gradient descent method was applied to optimize the parameters of PCA-ANFIS. Various optimum PCA-ANFIS models based on the conventional spectra and pretreated spectra (standard normal variate (SNV), multiplicative scatter correction (MSC) and the first-derivative) were established and compared. Experiment results indicated that the PCA-ANFIS model obtained from data sets achieved satisfactory accuracy, and the PCA-ANFIS approach with MSC pretreated spectra was found that it provided the best results. In order to present the advantages of PCA-ANFIS, the principal component regression (PCR) was also used, which was compared with PCA-ANFIS. Experiment results showed that the proposed PCA-ANFIS was more efficient than PCR.

1. Introduction

Reflectance near-infrared spectroscopy is a rapid, cost-effective and nondestructive technique. It has been applied widely in agricultural, textile, petrochemical and pharmaceutical industries [1–4], especially the application of NIR spectroscopy for the analysis of pharmaceutical samples has been significantly increased during the last decade [5–7]. Multivariate calibration methods are usually used to extract relevant information from NIR spectral data to predict analyte concentrations or properties of complex samples. Principal component regression (PCR) and partial least squares regression (PLS) as two of the multivariate calibration methods are most frequently used [8,9]. However the both methods possess some deficiencies such as modeling of data sets containing strong nonlinear relationships [10], whereas there always exists nonlinear mapping between the spectral data and concentration of the component [11–13].

In recent years, artificial neural networks (ANNs) have been widely applied as a chemometrics method. Their ability to handle nonlinearities makes them a valuable contribution to the discipline. It has been demonstrated that it is possible to obtain excellent results for dealing with multivariate calibration problems with ANNs [14,15]. However, ANNs still faces certain drawbacks when dealing with highly dimensional systems within the input space, *i.e.*, the "curse of dimensionality" [16–19]. Principal component analysis is regarded as the main extraction method and can reduce redundant variables from the original input data [20]. Through PCA analysis, the original multiple variables can be represented by several principal components and lead to the reduction of input dimension in the ANN model.

Since Zadeh proposed the fuzzy logic theorem to describe complicated systems; fuzzy inference system (FIS) has became very popular and been successfully used in various scientific areas. One of the successful fuzzy applications is to model complex nonlinear systems by a set of fuzzy rules. One important property of fuzzy modeling approaches is that FIS can approximate virtually any nonlinear functions to arbitrary accuracy provided that enough rules are given [21]. The major benefit of FIS is that its

* Corresponding author.
E-mail address: qunan.jlau@gmail.com (S. Dou).

1876-1070/$ – see front matter © 2012 Taiwan Institute of Chemical Engineers. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.jtice.2012.01.014

knowledge representation is explicit, using simple IF–THEN relations. But this model uses the human-determined membership functions that are fixed. Therefore, they are rarely optimal in terms of reproducing the desired outputs. In the last decade, the integration of ANN and FIS has given birth to new research into neuron-fuzzy systems. Neuron-fuzzy systems have the potential to capture the benefits of both these fields in a single framework. A popular framework of neuron-fuzzy is the adaptive neuron-fuzzy inference system (ANFIS), which shows the significant results in nonlinear modeling. ANFIS has attracted the growing attention and interest in robot control, pattern recognition, data analysis, decision making, nonlinear noise cancellation, *etc.* [22–26].

The purpose of this work is to develop a nondestructive quantitative analysis method of pharmaceutical sample based on PCA-ANFIS and NIR spectroscopy. The proposed method can be divided into two stages. In the first stage, PCA is applied to compress a large number of data to much smaller principal components and generate principal component scores that subsequently are used as the input variables of the ANFIS instead of original data. In the second stage, ANFIS model is built to create the fuzzy inference system and then to determine the concentrations of pharmaceutical samples. Various PCA-ANFIS models based on conventional spectra and pretreated spectra (SNV, MSC and the first-derivative) have been successfully achieved respectively. Among all the pretreated spectra of PCA-ANFIS models, only the PCA-ANFIS model on MSC spectra has the smallest root-mean-square-error (RMSE) and the highest correlation coefficient (R); thus the application of the MSC spectra obtains the best model and satisfactory results.

2. Theoretical

2.1. Principal component analysis

PCA is a classical statistical method which has been extensively applied in almost every discipline, chemistry, biology, engineering, meteorology, *etc*. There are some excellent descriptions of the algorithm of PCA [27,28] and this paper will provide only a general overview. The most important application of PCA is to reduce the number of original variables and represent a multi-dimensional data table in a low-dimensional space with minimal loss of information of the original data set. PCA decomposes an X matrix into two smaller matrices, one of scores (T) and the other of loadings (P) as follows:

$$X = TP^{\mathrm{T}} \tag{1}$$

The method generates a new set of variables (loadings) which is called PCs. Each PC is a linear combination of the original variables. All the PCs are orthogonal to each other, so there is no redundant information. The PCs are extracted so that the first PC accounts for the maximum variance of the original multivariate data set, and the second PC explains the maximum variances of the residual data set. Then, the third one will describe the most important variability of the next residual data set and so on. Generally, only a handful of PCs are enough to account for the most variance of the original data set.

2.2. Adaptive neuron-fuzzy inference system

The ANFIS is a multilayer feed-forward network which used neural network learning algorithms and fuzzy reasoning to map an input space to an output space [29]. It implements a first-order Sugeno fuzzy model. For simplicity, we assume that the fuzzy inference system under consideration has three inputs x, y and z, and one output f. For a first-order Sugeno fuzzy model, a typical rule set with two fuzzy if–then rules can be expressed as

$$\begin{aligned}\text{Rule } i: \quad & \text{if } x \text{ is } F_i^x \text{ and } y \text{ is } F_i^y \text{ and } z \text{ is } F_i^z \\ & \text{then } f_i = p_i x + q_i y + r_i z + s_i \quad \text{for } i = 1, 2\end{aligned}$$

where F_i^x, F_i^y and F_i^z are the fuzzy sets for the inputs x, y and z, respectively; p_i, q_i, r_i and s_i are linear parameters in the then-part of fuzzy if–then rules, and are called consequent parameters. The architecture of ANFIS with five layers is shown in Fig. 1, and a brief introduction of the model is as follows.

Layer 1: Each node of this layer generates membership grades of an input variable. The node output is $O_{1,i}$ given by

$$O_{1,i} = \mu_{F_i^x}(x) \quad \text{for } i = 1, 2 \tag{2}$$

$$O_{1,i} = \mu_{F_{i-2}^y}(y) \quad \text{for } i = 3, 4 \tag{3}$$

$$O_{1,i} = \mu_{F_{i-4}^z}(z) \quad \text{for } i = 5, 6 \tag{4}$$

where x, y and z are the crisp inputs to node i, and F_i^x, F_{i-2}^y and F_{i-4}^z are the linguistic labels characterized by appropriate membership

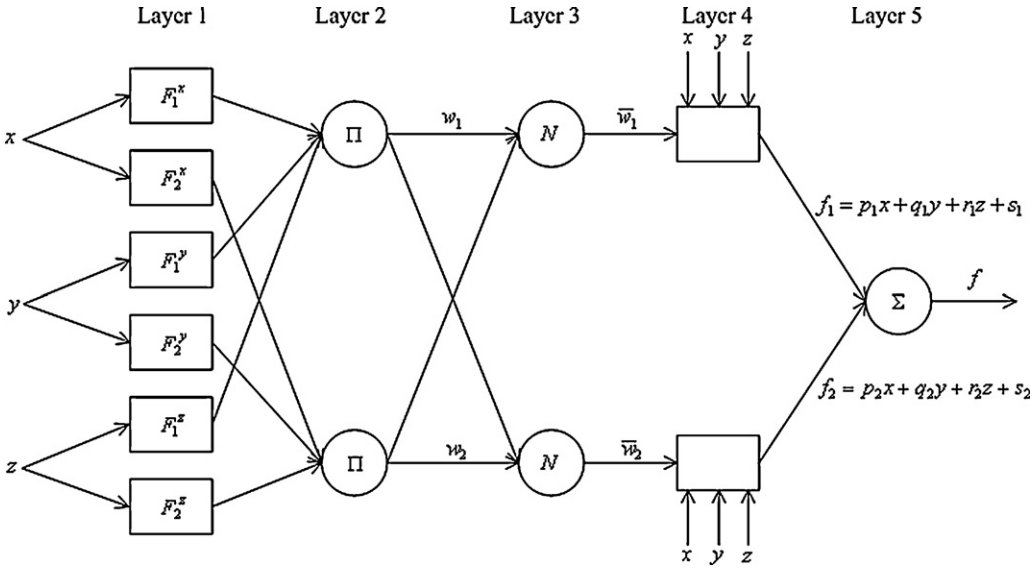

Fig. 1. ANFIS structure for three input Sugeno fuzzy model with two rules.

functions $\mu_{F_i^x}$, $\mu_{F_{i-2}^y}$ and $\mu_{F_{i-4}^z}$, respectively. The Gaussian membership functions are increasingly popular for specifying fuzzy sets as they are nonlinear, smooth and their derivatives are continuous, it is given by

$$\mu_{F_i^x}(x) = \exp\left(-\frac{(x-c_i^x)^2}{2(\delta_i^x)^2}\right) \quad \text{for } i = 1,2 \tag{5}$$

$$\mu_{F_i^y}(y) = \exp\left(-\frac{(y-c_i^y)^2}{2(\delta_i^y)^2}\right) \quad \text{for } i = 1,2 \tag{6}$$

$$\mu_{F_i^z}(z) = \exp\left(-\frac{(y-c_i^z)^2}{2(\delta_i^z)^2}\right) \quad \text{for } i = 1,2 \tag{7}$$

where δ_i^x, δ_i^y, δ_i^z, c_i^x, c_i^y and c_i^z are the parameters of the membership functions in the premise part of fuzzy if–then rules that change the shape of the membership functions with minimum and maximum equal to 0 and 1, respectively. Parameters in this layer are referred to as premise parameters.

Layer 2: In this layer, the nodes are fixed nodes. They are labeled with Π, which multiplies the incoming signals and sends the product out. The outputs of this layer can be represented as

$$O_{2,i} = w_i = \mu_{F_i^x}(x)\mu_{F_i^y}(y)\mu_{F_i^z}(z) \quad \text{for } i = 1,2 \tag{8}$$

Each node output represents the firing strength of a rule.

Layer 3: In this layer, the nodes are also fixed nodes, they are labeled with *N*. The *i*th node calculates the ratio of the *i*th rule's firing strength to the sum of all rules' firing strengths. The outputs of this layer are given by

$$O_{3,i} = \bar{w}_i = \frac{w_i}{\sum_{j=1}^{2} w_j} \quad \text{for } i = 1,2 \tag{9}$$

which are called normalized firing strengths.

Layer 4: In this layer, the nodes are adaptive nodes. The output of each node in this layer is computed as

$$O_{4,i} = \bar{w}_i f_i = \bar{w}_i(p_i x + q_i y + r_i z + s_i) \quad \text{for } i = 1,2 \tag{10}$$

where $\bar{w}_i$ is the output of layer 3 and f_i is a linear function of input variables.

Layer 5: In this layer, there is only one single fixed node labeled with Σ. This node computes the weighted average of the output signals of the previous layer as

$$O_{5,i} = \sum_{i=1}^{2} \bar{w}_i f_i = \frac{\sum_{i=1}^{2} w_i f_i}{\sum_{i=1}^{2} w_i} \tag{11}$$

It can be seen that there are two adaptive layers in this ANFIS architecture, namely the layer 1 and the layer 4. Layer 1 has modifiable premise parameters related to the input membership function. Layer 4 also has modifiable consequent parameters pertaining to the first-order polynomial.

The task of the learning algorithm for this architecture is to tune all the modifiable parameters to make the ANFIS output match the training data. There are a number of proposals on how to define these parameters in the literature. The most popular one is the hybrid learning algorithm which combines the least squares method and the gradient descent method [30]. The hybrid algorithm is composed of a forward pass and a backward pass. In the forward pass, the functional signals go forward till layer 4 and the consequent parameters are identified by the least squares estimate. In the backward pass, the error rates propagate backward and the premise parameters are updated by the gradient descent.

2.3. Subtractive clustering

Consider a collection of n data points $\{x_1, x_2, \ldots, x_n\}$ in M dimensional space. Assume that the data points have been normalized in each dimension so that their coordinate ranges in each dimension are equal. Consider each data point as a potential cluster center and define a measure of the potential of data point x_i as

$$P_i = \sum_{j=1}^{n} e^{-\alpha\|x_i - x_j\|^2} \tag{12}$$

where

$$\alpha = \frac{4}{r_a} \tag{13}$$

r_a is a positive constant called *radius*.

After the potential of every data point has been computed, select the data point with the highest potential as the first cluster center. Let x_1^* be the location of the first cluster center and P_1^* be its potential value. Then revise the potential of each data point x_i by the formula

$$P_i \Leftarrow P_i - P_1^* e^{-\beta\|x_i - x_1^*\|^2} \tag{14}$$

where

$$\beta = \frac{4}{r_b^2} \tag{15}$$

To avoid obtaining closely spaced cluster centers, set r_b to be somewhat greater than r_a. A good choice is $r_b = 1.5r_a$. When the potential of all data points has been revised, select the data point with the highest remaining potential as the second cluster center. Then further reduce the potential of each data point according to their distance to the second cluster center. In general, after the *k*th cluster center has been obtained, revise the potential of each data point by the formula

$$P_i \Leftarrow P_i - P_k^* e^{-\beta\|x_i - x_k^*\|^2} \tag{16}$$

where x_k^* is the location of the *k*th cluster center and P_k^* is its potential value. The process of acquiring new cluster center and revising potentials repeats until the remaining potential of all data points falls below some fraction of the first cluster center P_1^*.

Consider a set of c cluster centers $\{x_1^*, x_2^*, \ldots, x_c^*\}$ in an M dimensional space. Let the first N dimensions correspond to input variables and the last M–N dimensions correspond to output variables. Decompose each vector x_i^* into two component vectors y_i^* and z_i^*, where y_i^* contains the first N elements of x_i^* (*i.e.*, the coordinates of the cluster center in input space) and z_i^* contains the last M–N elements (*i.e.*, the coordinates of the cluster center in output space).

Consider each cluster center x_i^* as a fuzzy rule that describes the system behavior. Given an input vector y, the degree of fulfillment of rule i is defined as

$$\mu_i = e^{-\alpha\|y_i - y_i^*\|^2} \tag{17}$$

Compute the output vector z via

$$z = \frac{\sum_{i=1}^{c} \mu_i z_i^*}{\sum_{i=1}^{c} \mu_i} \tag{18}$$

$$z_i^* = G_i y + h_i \tag{19}$$

where G_i is an $(M–N) \times N$ constant matrix and h_i is a constant column vector with M–N elements. The equivalent if–then rules

then become the Takagi-Sugeno type, where the consequent of each rule is a linear equation in the input variables [31].

3. Experimental

3.1. Apparatus and software

The NIR spectra were measured with a Shimadzu®3101 spectrophotometer (Tokyo, Japan) with ISR-3101 integrating sphere to collect the sample spectra. The apparatus was controlled by UVPC Personal Spectroscopy Software, which is a commercial available NIR spectral analysis software package. The microcomputer with Inter CORE dual core 2.66 GHz CPU, 2GB DDR3 RAM and Windows XP operating system was used for data processing. The PCA-ANFIS was implemented by using the MATLAB software package (MATLAB version R2006a with fuzzy logic toolbox 2.2.3 and statistics toolbox 5.2). The PCR was performed by using TQ 6.6.1 (Thermo Nicolet, Madison, WI, USA) software package.

3.2. Sample preparation

All the raw material powders include thiamphenicol as active component and starch as excipient. The average concentration of thiamphenicol was 70.78% (g/g), and the concentration range of all the samples was 59.27–83.55% (g/g). All of the standard referenced concentrations were measured according to the Chinese Pharmacopoeia method [32], as the reference method: samples were homogenized, and the amounts of thiamphenicol powder samples were accurately weighed and grinded, afterwards they were dissolved in 30 ml of ethanol and added 20 ml of potassium hydroxide solution (50%). Then the sample solution was heated to reflux for 4 h, cooled and diluted in 100 ml water. The dilute nitric acid was used to neutralize, and the more 7.5 ml of dilute nitric acid was added. Finally, the concentration of thiamphenicol was determined by potentiometric titration method with a silver glass electrode and silver nitrate solution (0.1 mol/l) as the titrant. The result of titration was corrected by blank test. Every 1 ml of (0.1 mol/l) silver nitrate titrant was equivalent to 17.81 mg of $C_{12}H_{15}Cl_2NO_5S$.

3.3. Collection of NIR spectra

The scan wavelength range was from 1100 to 2500 nm (9091 to 4000 cm^{-1}) and the entrance slit of the NIR spectrophotometer used was 12 nm. Each recorded spectrum was the average of 10 scans and contained absorbance data at wavelength intervals of 1 nm between 1100 and 2500 nm, giving 1401 available wavelengths.

The conventional NIR spectra that form the training set are shown in Fig. 2.

3.4. Evaluation of the performance of the models

The 60 powder samples were randomly divided into two separate data sets, *i.e.*, the training set included 45 samples and the test set included 15 samples. The leave-one-out cross-validation method was used for the model selection criterion. The "leaving-one-out" method is leaving one sample out and using the rest of the samples to build the model. Then the model is used to predict the sample being left out. This step is repeated for every sample in the training set samples. Root-mean-square-error (RMSE) of cross-validation is obtained by leave-one-out cross-validation via the set of training samples, which gives an estimate of the models' performance. The test set samples are used as an independent set to calculate the final prediction error. The predictive abilities of training set and test set of the different models (conventional, SNV, MSC and the first-derivative spectra) were compared in terms of the RMSE. The RMSE given by

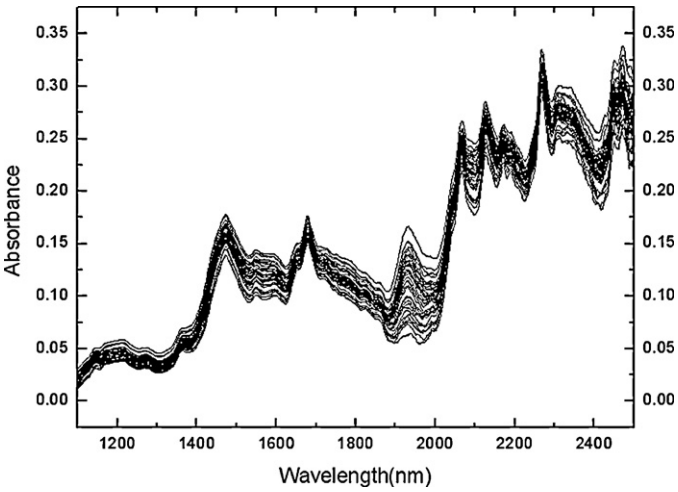

Fig. 2. NIR absorbance data for all thiamphenicol samples in training set.

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(c_i^{\mathrm{p}} - c_i^{\mathrm{r}})^2}{n}} \tag{20}$$

In Eq. (20) n is the number of samples included in the set being processed, c_i^{p} is the predicted value for the sample and c_i^{r} is the value of the standard referenced method for sample.

4. Results and discussion

4.1. Training and optimization of PCA-ANFIS using MSC spectra

4.1.1. Input variable analysis using the PCA method

For the MSC spectra, PCA was performed firstly to generate the scores and the loadings from the training set. For PCA to work properly, we subtracted the mean from each of the data dimensions. The scores were used as the input variables of the ANFIS instead of original data, and the loadings were applied to calculate the test set scores. The PCA results were shown in Fig. 3. As can be seen from this figure, the PCA process of the original data gave two important principal components because the total variance percentage of these two components is 95.28%. Thus, information in the original data would lose little based on these

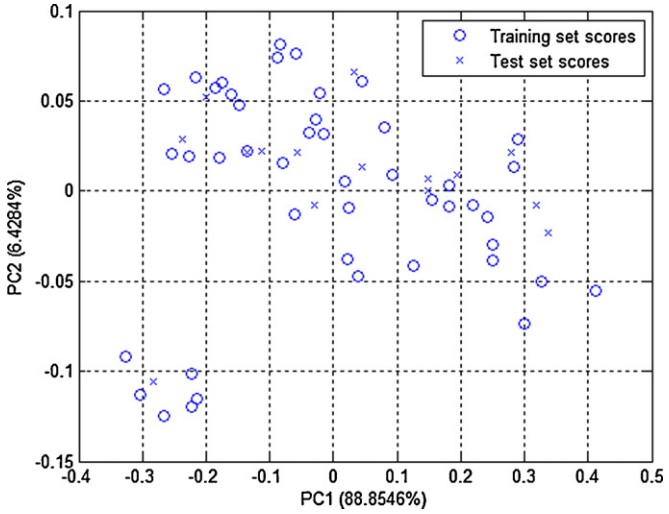

Fig. 3. The variance explained by the corresponding principal component.

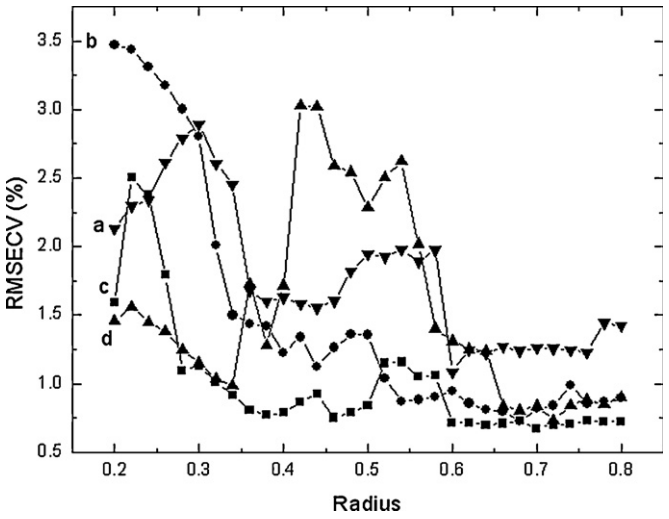

Fig. 4. The effect of *radius* value to the RMSE value of PCA-ANFIS models (a) conventional spectra, (b) SNV spectra, (c) MSC spectra and (d) the first-derivative spectra.

two components. Therefore, PCA was used for reducing the size of the sample matrix (1401 variables) by retaining the first 2 PC scores as descriptors.

4.1.2. The calculated results of PCA-ANFIS model

The PCA-ANFIS model was utilized to differentiate the spectral characteristics as well as to quantify the concentration of each of the pharmaceutical samples. The hybrid learning algorithm was used to build the PCA-ANFIS model according to input–output data patterns. Using training data, including the first 2 PC scores obtained by PCA as input variables to PCA-ANFIS and concentrations of the standard referenced method as the output variables, a FIS along with a subtractive clustering was developed for concentration prediction. The subtractive clustering can automatically determine the number of clusters. It assumes that each data point is a potential cluster center and calculates a measure of the likelihood that each data point would define the cluster center, based on the density of surrounding data points. In this algorithm, an important parameter is *radius*. The *radius* is a scalar between 0 and 1 that specifies a cluster center's range of influence, assuming that the data fall within a unit hypercube. Specifying small cluster *radius* will usually yield many small clusters in the data, resulting in many fuzzy rules and vice versa. Satisfactory values for the *radius* of each cluster are usually between 0.2 and 0.8.

At the beginning of the PCA-ANFIS model training, varying the level of parameter *radius* (from 0.2 to 0.8) will generate a response of their corresponding RMSE of cross-validation. And the selection of optimum parameter *radius* was made with the curve of root-mean-square-error of cross-validation as shown in Fig. 4. The minimum RMSE value was expected to occur when the optimal parameter of PCA-ANFIS model is retained. The *radius* of 0.70 was used for each cluster after the parameter had been evaluated, which led to generation of 4 fuzzy rules. This FIS was then used as an initial FIS for PCA-ANFIS modeling. PCA-ANFIS finds the best function mapping the input variables to the output variable. Fig. 5 shows the final membership functions of input variables. The final four fuzzy if–then rules take the following form:

$$\text{Rule } k: \quad \text{if } \text{score}_1 \text{ is } F_k^1 \text{ and } \text{score}_2 \text{ is } F_k^2 \quad \text{then } f_k = p_k\text{score}_1 + q_k\text{score}_2 + r_k$$

where $k = 1, \ldots, 4$, score_1 and score_2 are first two PC scores, F_k^1 and F_k^2 are the fuzzy sets with membership functions $\mu_{F_k^1}$ and $\mu_{F_k^2}$, respectively. p_k, q_k and r_k are consequent parameters pertaining to the first-order polynomial f_k, their values are shown in Table 1.

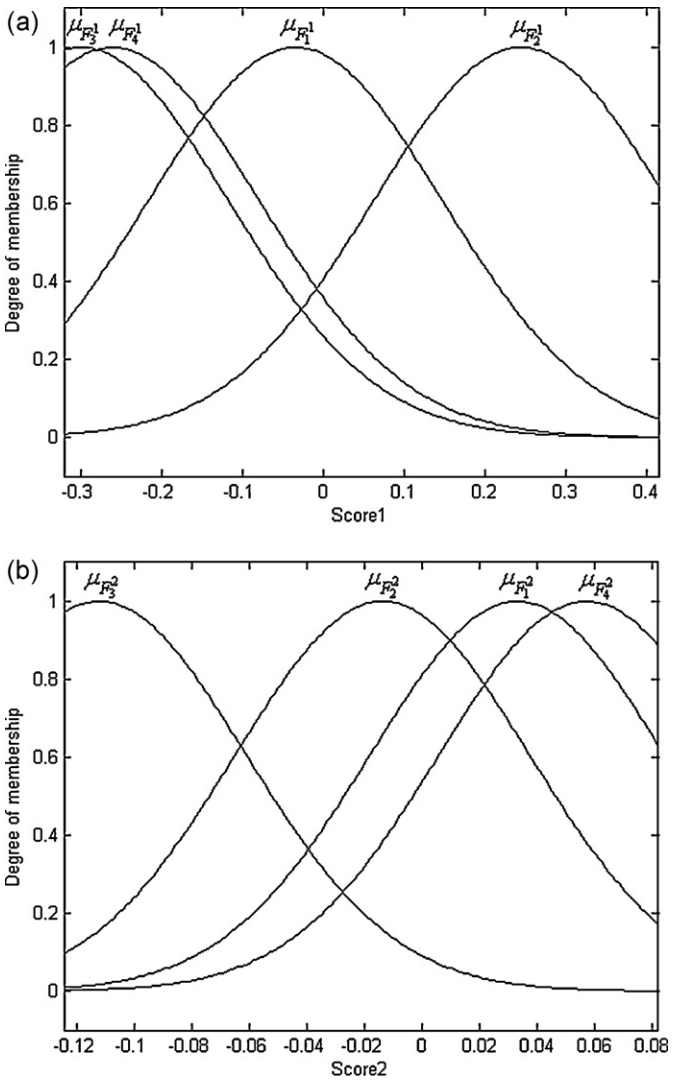

Fig. 5. Final membership functions for (a) score_1 and (b) score_2.

4.2. PCA-ANFIS using other spectra and evaluation

The PCA was performed on the conventional spectra, SNV and first-derivative corrected spectra, too. The selected numbers of PC of conventional spectra, SNV and the first-derivative spectra were two, three and three, respectively (Table 2). The PCA-ANFIS models of the conventional spectra, SNV and the first-derivative spectra were established. The established process was similar to the MSC model. Using different input parameters, the different PCA-ANFIS models will be generated. The RMSE was used to evaluate the fitness of the models. The optimal PCA-ANFIS topology parameter settings are listed in Table 2. The generated fuzzy rules of conventional spectra, SNV and the first-derivative spectra are three, two and four, respectively.

When the adjustable parameter *radius* of PCA-ANFIS model was optimized, the optimal PCA-ANFIS had a high ability to predict the

Table 1
Consequent parameters for MSC spectra.

Rule i	p_i	q_k	r_i
1	−0.5268	1.4320	0.7972
2	0.1150	0.0410	0.8017
3	0.2119	0.3263	0.7416
4	−0.7065	1.1310	0.1743

Table 2
Optimized parameters used for construction of PCA-ANFIS models.

Parameter	Conventional	SNV	MSC	First-derivative
PCs	2	3	2	3
Amount of variance explained	94.44%	96.18%	95.28%	95.51%
radius	0.60	0.68	0.70	0.72

concentrations of thiamphenicol samples. The relationship of the standard reference concentration with each of the predicted ones of the training set is clearly summarized in Table 3. The goodness of fit of PCA-ANFIS models based on conventional and pretreated spectra was compared in terms of the RMSE and R. RMSE measures the average deviation between the standard reference concentration and the predicted concentration. Small RMSE value of a model indicates a better fit of data for that model. The R values indicate a linear relationship between the standard reference concentration and the predicted concentration by the model and the higher R value, the better is the adequacy of the model to describe the data. As shown in Table 3, the smallest RMSE and higher R occur in the quantitative prediction by means of optimal PCA-ANFIS set up by MSC spectra, which shows MSC spectra have the priority over conventional spectra, SNV spectra and the first-derivative spectra. Therefore, the PCA-ANFIS model established by MSC spectra is the best.

To verify the reliability of the constructed models, the trained PCA-ANFIS models were then used to predict concentrations of thiamphenicol samples using test set data not used in the training procedure. These predicted concentrations are compared with standard reference concentrations to check the PCA-ANFIS model performance. Concentrations of the test set samples were accurately predicted with RMSE of 0.9796% for conventional spectra, 0.5101% for MSC spectra, 0.6896% for SNV spectra and 0.7428% for the first-derivative spectra, respectively. The thiamphenicol concentration correlation plots for the optimal PCA-ANFIS model of MSC spectra are shown in Fig. 6 for the training and test sets. It can be seen that the predicted concentrations are highly correlated with the standard reference concentrations, thus the model of MSC spectra can give better performance.

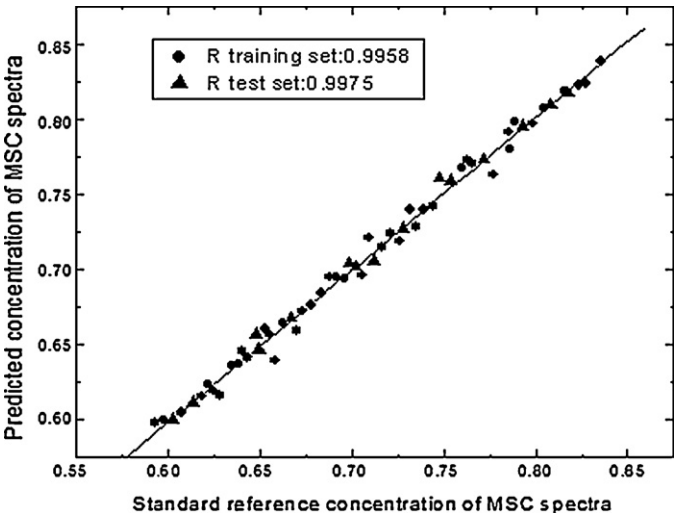

Fig. 6. The predicted concentrations of MSC pre-processed spectra are plotted against the standard reference concentrations (PCA-ANFIS).

4.3. Determination by PCR calibration models

In order to stand out advantages of PCA-ANFIS models, the PCR models were used to make a prediction of the concentration of compound thiamphenicol powder. PCR has been demonstrated as a useful technique to quantitative analysis by NIR spectra with increased matrix complexity. These models generated by PCR used the same data sets and the same number of PCs as in the PCA-ANFIS models. We established the PCR models of conventional spectra, SNV, MSC and first-derivative spectra to compare them with PCA-ANFIS models. The RMSE value on both the training set and the test set was used for the model selection criterion. The thiamphenicol concentration correlation plots for the optimal PCR model of MSC spectra are shown in Fig. 7 for the training and test sets. A summary of the comparison of PCA-ANFIS models with PCR models is given in Table 3. As can be seen, the PCA-ANFIS models achieve better prediction performance than PCR models. The results indicate that the PCA-ANFIS models have more advantages than the PCR models.

4.4. Four-fold cross-validation

To validate the performance of calibration models further, the four-fold cross-validation is applied. These models generated by ANFIS and PCR used the same data sets. The 60 samples were randomly divided into four subsets, each of which contained 15 samples. The four calibration models were obtained, each time one of the four subsets was assigned to a test set and the other three subsets were allocated to a training set. The overall performance

Table 3
Statistical parameters of relationship between the standard reference concentrations and predicted ones of the best PCA-ANFIS models.

Model	Spectra	Set	R	RMSE (%)	Four-fold cross-validation	
					R	RMSE (%)
PCA-ANFIS	Conventional	Training set	0.9911	0.9145	0.9900	0.9005
		Test set	0.9920	0.9796	0.9911	1.0087
	SNV	Training set	0.9949	0.6961	0.9949	0.6872
		Test set	0.9974	0.6896	0.9949	0.6359
	MSC	Training set	0.9958	0.6426	0.9952	0.6700
		Test set	0.9975	0.5101	0.9961	0.5422
	First-derivative	Training set	0.9945	0.7217	0.9939	0.7409
		Test set	0.9938	0.7428	0.9936	0.8015
PCR	Conventional	Training set	0.9698	1.6738	0.9799	1.3412
		Test set	0.9775	1.4682	0.9696	1.6085
	SNV	Training set	0.9799	1.3708	0.9892	0.9682
		Test set	0.9828	1.2432	0.9856	1.0055
	MSC	Training set	0.9909	0.9237	0.9908	0.9279
		Test set	0.9960	0.7231	0.9885	0.9130
	First-derivative	Training set	0.9719	1.6235	0.9718	1.5462
		Test set	0.9865	1.2295	0.9777	1.5993

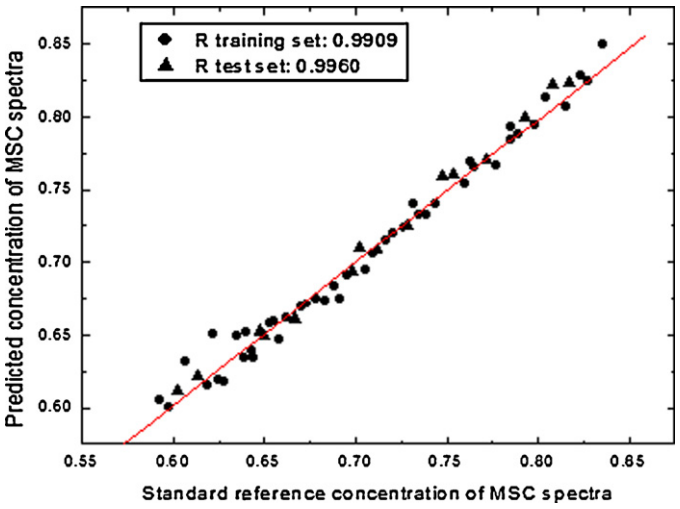

Fig. 7. The predicted concentrations of MSC pre-processed spectra are plotted against the standard reference concentrations (PCR).

was calculated as the average performance of the four models over the corresponding test partitions of the data. The performance of the models was assessed using RMSE and the correlation coefficient R. The average results achieved via four-fold cross-validation were also presented in Table 3. From the calculated results in the table, it can be concluded that ANFIS has better performance than the PCR models.

5. Conclusion

The purpose of this study was to investigate the application of PCA-ANFIS and NIR spectroscopy in the nondestructive prediction of pharmaceutical samples. ANFIS is a powerful fuzzy logic neural network, which provides a method for fuzzy modeling to learn information about the data set that best allow the associated fuzzy inference system to trace the given input–output data. The PCA was presented and applied to reduce and orthogonalize the input variables of an ANFIS model implemented for thiamphenicol sample prediction. It has been demonstrated that the proposed algorithm is a simple and effective means for the nondestructive determination of thiamphenicol samples by near-infrared (NIR) spectroscopy. The results show that the PCA-ANFIS models of spectra transited by SNV, MSC and the first-derivative correction give more acceptable results than conventional spectral model, and the best one is of MSC spectra.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and helpful suggestions. This work was financially supported by the National Natural Science Foundation of China (41171188) and Doctor Initial Foundation of Jilin Agricultural University (201102).

References

[1] Balabin RM, Safieva RZ. Near-infrared (NIR) spectroscopy for biodiesel analysis: fractional composition, iodine value, and cold filter plugging point from one vibrational spectrum. Energy Fuels 2011;25:2373.
[2] Balabina RM, Smirnov SV. Melamine detection by mid- and near-infrared (MIR/NIR) spectroscopy: a quick and sensitive method for dairy products analysis including liquid milk, infant formula, and milk powder. Talanta 2011;85:562.
[3] Balabin RM, Smirnov SV. Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data. Anal Chim Acta 2011;692:63.
[4] Büning-Pfaue H. Analysis of water in food by near infrared spectroscopy. Food Chem 2003;82:107.
[5] Dou Y, Ren YL, Teng LR, Liang Y. Nondestructive quantitative analysis of cimetidine tablets using artificial neural networks in near-infrared spectroscopy. Spectrosc Lett 2005;38:1.
[6] Qu N, Wang LH, Zhu MC, Dou Y, Ren YL. Radial basis function networks combined with genetic algorithm applied to nondestructive determination of compound erythromycin ethylsuccinate powder. Chemom Intell Lab Syst 2008;90:145.
[7] Sajan D, Laladhas KP, Hubert J, Jayakumar VS. Vibrational spectra and density functional theoretical calculations on the antitumor drug, plumbagin. J Raman Spectrosc 2005;36:1001.
[8] Hemmateenejad B, Rezaei Z, Khabnadideh S, Safari M. A PLS-based extractive spectrophotometric method for simultaneous determination of carbamazepine and carbamazepine-10,11-epoxide in plasma and comparison with HPLC. Spectrochim Acta A Mol Biomol Spectrosc 2007;68:718.
[9] Otsuka M, Kato F, Matsuda Y. Determination of indomethacin polymorphic contents by chemometric near-infrared spectroscopy and conventional powder X-ray diffractometry. Analyst 2011;126:1578.
[10] Blanco M, Coello J, Iturriaga H, Maspoch S, Pagès J. Calibration in non-linear near infrared reflectance spectroscopy: a comparison of several methods. Anal Chim Acta 1999;384:207.
[11] Balabin RM, Safieva RZ. Biodiesel classification by base stock type (vegetable oil) using near infrared spectroscopy data. Anal Chim Acta 2011;689:190.
[12] Balabin RM, Lomakina EI. Support vector machine regression (SVR/LS-SVM)-an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data. Analyst 2011;136:1703.
[13] Qu N, Mi H, Wang B, Ren YL. Application of GA-RBF networks to the nondestructive determination of active component in pharmaceutical powder by NIR spectroscopy. J Taiwan Inst Chem Eng 2009;40:162.
[14] Safavi A, Abdollahi H, Nezhad MR. Artificial neural networks for simultaneous spectrophotometric differential kinetic determination of Co (II) and V (IV). Talanta 2003;59:515.
[15] Zhang YX, Li H, Hou AX, Havel J. Artificial neural networks based on principal component analysis input selection for quantification in overlapped capillary electrophoresis peaks. Chem Intell Lab Syst 2006;82:165.
[16] Balabin RM, Lomakina EI, Safieva RZ. Neural network (ANN) approach to biodiesel analysis: Analysis of biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy. Fuel 2011;90:2007.
[17] Balabin RM, Safieva RZ, Lomakina EI. Near-infrared (NIR) spectroscopy for motor oil classification: from discriminant analysis to support vector machines. Microchem J 2011;98:121.
[18] Balabin RM, Safieva RZ, Lomakina EI. Gasoline classification using near infrared (NIR) spectroscopy data: comparison of multivariate techniques. Anal Chim Acta 2010;671:27.
[19] Ji J, Wang HQ, Chen K, Liu Y, Zhang N, Yan JJ. Recursive weighted kernel regression for semi-supervised soft-sensing modeling of fed-batch processes. J Taiwan Inst Chem Eng 2012;43:67.
[20] Lu WZ, Wang WJ, Wang XK, Yan SH, Lam JC. Potential assessment of a neural network model with PCA/RBF approach for forecasting pollutant trends in mong kok urban air, Hong Kong. Environ Res 2004;96:79.
[21] Song ZS, Yi JQ, Zhao DB, Li XC. A computed torque controller for uncertain robotic manipulator systems: fuzzy approach. Fuzzy Sets Systems 2005;154:208.
[22] Buyukbingol E, Sisman A, Akyildiz M, Alparslan FN, Adejare A. Adaptive neuron-fuzzy inference system (ANFIS): A new approach to predictive modeling in QSAR applications: a study of neuron-fuzzy modeling of PCP-based NMDA receptor antagonists. Bioorgan Med Chem 2007;15:4265.
[23] Bateni SM, Jeng DS. Estimation of pile group scour using adaptive neuron-fuzzy approach. Ocean Eng 2007;34:1344.
[24] Chang FJ, Chang YT. Adaptive neuron-fuzzy inference system for prediction of water level in reservoir. Adv Water Resour 2006;29:1.
[25] Geethanjali M, Slochanal S. A combined adaptive network and fuzzy inference system (ANFIS) approach for overcurrent relay system. Neurocomputing 2008;71:895.
[26] Qin H, Yang SX. Adaptive neuron-fuzzy inference systems based approach to nonlinear noise cancellation for images. Fuzzy Sets Systems 2007;158:1036.
[27] Brereton RG. Introduction to multivariate calibration in analytical chemistry. Analyst 2000;125:2125.
[28] Statheropoulos M, Pappa A, Karamertzanis P, Meuzelaar H. Noise reduction of fast, repetitive GC/MS measurements using principal component analysis (PCA). Anal Chim Acta 1999;401:35.
[29] Denaï M, Palis F, Zeghbib A. Modeling and control of non-linear systems using soft computing techniques. Appl Soft Comput 2007;7:728.
[30] Jang J. ANFIS: Adaptive-network-based fuzzy inference system. IEEE Trans Systems Man Cybernet 1993;23:665.
[31] Chiu S, Fuzzy model identification based on cluster estimation. J Intell Fuzzy Syst 1994;2:267.
[32] Pan Z. Pharmacopoeia of the People's Republic of China, Part II. Beijing: Chemical Industry Press; 2000.