Advance Journal of Food Science and Technology 12(2): 82-87, 2016
DOI:10.19026/ajfst.12.2844
ISSN: 2042-4868; e-ISSN: 2042-4876

Submitted: September 9, 2015 Accepted: September 25, 2015 Published: September 15, 2016

Research Article

A Comparison of Methods for Classification of Flue-cured Tobacco Aroma Types

Fenghua Ma and Wei Wu
College of Computer and Information Science, Southwest University, Chongqing 400716, China

Abstract: It is well acknowledged that flue-tobacco aroma types were divided into light, medium and heavy in China. For the sake of singling out an optimal scheme to discriminate the spatial distribution of flue-cured tobacco aroma type, in the current study, different amounts of chemical indices data with various methods including Back-Propagation Neural Networks (BP NN), Support Vector Machine (SVM) and Discriminant Analysis (DA) were presented and compared. All the experimental results indicated that, by and large, the number of chemical indices have nothing to do with the accuracy. Additionally, the classification effects of BP NN are superior to the others. On a whole, the best scheme with accuracy reaching to 81.18% and kappa value up to 0.72 was drawn only when the BP model combined with 9 kinds of chemical indices. In the end, the optimal spatial distribution was established in ArcGIS9.3.

Keywords: BP NN, DA, flue-cured tobacco aroma, spatial classification, SVM

INTRODUCTION

According to Food and Agricultural Organization (FAO), flue-cured tobacco is planted and sold in more than 150 countries and regions around the world. Among them, China is the major tobacco production and consumption country. Generally, there are three aroma types (light, medium and heavy) of flue-cured tobacco in China. The types are usually determined by experts. However, flue-cured tobacco aroma types are inevitably associated with leaf chemical compositions. Only few studies focused on identifying flue-cured tobacco aroma types using routinely measured chemical compositions at national scale (Bi *et al*., 2006; Yang *et al*., 2014; Zhang *et al*., 2013). For example, Bi *et al*. (2006) combined eight chemical variables with discriminate analysis method for classifying the FCT aroma types of Yunnan, Henan and Liaoning provinces and a relative good result was achieved. Yang *et al*. (2014) received a very good classification effects adopting the methods of SVM with various kinds of chemical compositions. By building the Fisher discriminant formula with 67 kinds of chemical compositions of FCT, Yan *et al*. has discriminated the aroma types of FCT, in 11 main tobacco production provinces of China.

It is well acknowledged that Artificial Neural Network (ANN), Support Vector Machine (SVM) and Discrimination Analysis (DA) are the most popular methods for supervised classification. With the great capability of nonlinear approximation and pattern recognition, ANN has been widely used in many fields, such as engineering, medical science, agriculture, finance and national defense (Widrow, 1988). Currently, Back Propagation (BP) neural network has been one of the most widely used ANNs. For example, Kavdır (2004) differentiated between 2 and 3 weeks old sunflower plants and common cocklebur weeds of similar size, shape and color by a back propagation neural network classifier. Being a powerful classifier, SVM has been also widely used in the fields where ANNs have dominated. For instance, Kolios and Stylios (2013) used a series of traditional and modern algorithms to investigate the Land Use and Land Cover (LULC) changes in a coastal area. They reported that the SVM classifier gave the best overall accuracy for the study area (Kolios and Stylios, 2013). Zheng *et al*. (2015) applied Support Vector Machines (SVMs) to discriminate various crop types in a complex cropping system in the Phoenix Active Management Area and the models achieved very high overall classification accuracy. DA is also a widely used supervised classifier. For example, Marey-Pérez and Rodríguez-Vicente (2011) revisited the factors determining forest management by farmers in northwest Spain using the discriminant analysis. Riveiro-Valino used discriminant analysis to validate the types of dairy farms obtained from the combinatorial method for Galicia (Riveiro-Valiño *et al*., 2009). Nieuwenhuizen *et al*. (2010) compared discriminant analysis and neural network to determine the reflectance properties of sugar beet and volunteer potato. They found that the neural network gave the best classification results (Nieuwenhuizen *et al*., 2010).

Corresponding Author: Wei Wu, College of Computer and Information Science, Southwest University, Chongqing 400716, China

In the current study, ANN, SVM and DA are compared to identify the flue-cured tobacco aroma types at national scale in China based on routinely measured chemical compositions of flue-cured tobacco leaves. The results are expected to provide valuable information on regional planning and decision making for producing high-quality flue-cured tobacco with different aroma types.

METHODOLOGY

BP neural network: Back Propagation (BP) neural network algorithm has been a fashion way for classification because of its strong nonlinear mapping ability and high learning accuracy. On the basis of the error function gradient of network, error against propagation algorithm is used to train the BP neural network. In this study, a multilayer feed-forward network including input layer, hidden layer and output layer was applied to classify the aroma types of flue-cured tobacco (Fig. 1).

Figure 1, numbers of neurons are included in each layer. $X_k = [x_{k1}, x_{k2}, x_{k3}, \ldots, x_{kM}]$ and $Y_k = [y_{k1}, y_{k2}, y_{k3}, \ldots, y_{kP}]T$ are the k-th input and output samples of the BP network. The number of input, hidden and output layer neurons is M, I and P, respectively.

The sigmoid function was the continuous differentiable non-linear activation function used for hidden and output layers. The function is defined as follows:

$$f(x) = 1/(1 + e^{-x}) \quad (1)$$

The input and output formulas of the i-th neuron in hidden layer are defined as:

$$\mu_i = \sum_{m=0}^{M} w_{mi}\, x_{km} + \theta_i \quad (2)$$

$$v_i = f(\sum_{m=0}^{M} w_{mi}\, x_{km}) \quad (3)$$

The input and output formulas of the p-th neuron in output layer are defined as:

$$\mu_p = \sum_{i=1}^{I} w_{ip}\, v_i + \theta_p \quad (4)$$

$$y_p = f\left(\sum_{i=1}^{I} w_{ip}\, v_i\right) \quad (5)$$

The output error of the p-th neuron in output layer is defined as:

$$e_{kp}(n) = t_{kp}(n) - y_{kp}(n) \quad (6)$$

The formula of weight modifying is defined as:

$$w_{ip}(n+1) = w_{ip}(n) + \eta \sum_{k=1}^{p} \delta_{ip}^{k}\, x_{ip} \quad (7)$$

where, w is the weight between neurons. v_i and y_p are the input and output values of output layer, separately. $T_k = [t_{k1}, t_{k2}, t_{k3}, \ldots, t_{kP}]$ is the expected output. Here, M, I and P were 18, 10 and 3, respectively. While n is the number of iterations, H represents the learning step size. And δ_{jp}^{k}stands for the local gradient, k is on the behalf of the k-th sample. More information on back propagation neural network could be found in Hecht-Nielsen (1989) and Johnson and Wichern (1992).

SVM: Support Vector Machine (SVM) developed by Vapnik (Li *et al*., 2009) is a statistical learning technique based on the VC (Vapnik-Chervonenkis) dimension theory which minimizes prediction error and model complexity (Li *et al*., 2009). SVMs overcome efficiency problems of ANNs, such as over-fitting and local minimum. Figure 2a the input vectors are mapped to a high feature space from the input space by a nonlinear transformation function. An optimal separating hyperplane can be structured successfully in this feature space. Figure 2b, circles and triangle represent different classification samples, respectively and the samples in solid line are support vectors. The

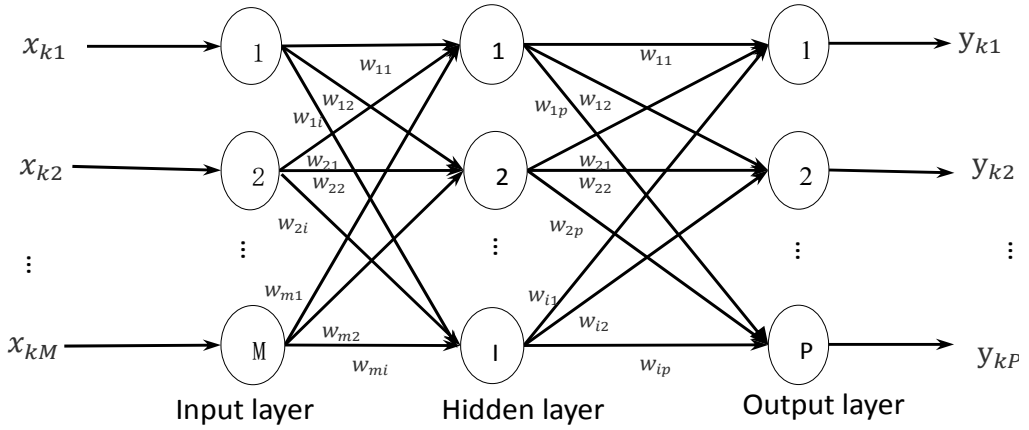

Fig. 1: Structure of three-layer BP neural network

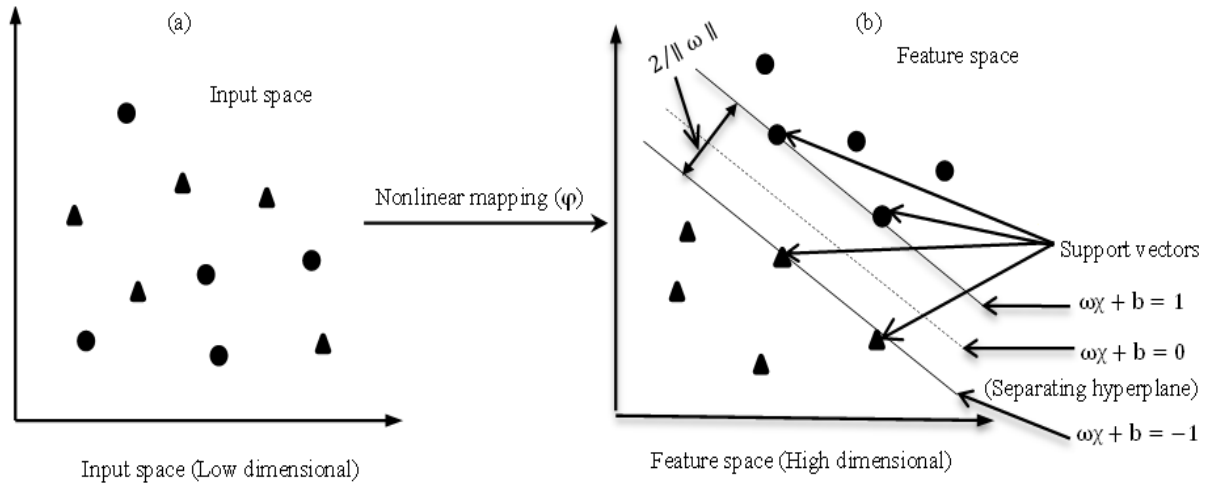

Fig. 2: Nonlinear mapping from input space to high dimensional space

classification interval between two parallel lines is $2/\| w \|$. The intention of SVM is to maximize the interval. The formulas $\omega\chi + \mathrm{b} = \pm 1$ and $\omega\chi + \mathrm{b} = 0$ stand for the classified lines and the separating hyperplane, respectively.

Given training samples $x_i \in R^n, i = 1,2,3, \ldots, l$, where l stands for the size of training set. In two classes, a vector $\mathrm{y}_i \in \{1, -1\}$, $(\mathrm{y} \in \mathrm{R}^l)$ delegates the label of category. Thus the initial problem can be described as follows:

$$\min_{\mathrm{w,b},\xi} \frac{1}{2}\mathrm{w}^{\mathrm{T}}\mathrm{w} + \mathrm{c}\sum_{i=1}^{l}\xi_i \qquad (8)$$

Subject to $\mathrm{y}_i(\mathrm{w})^{\mathrm{T}}\phi(\chi_{\mathrm{i}}) + \mathrm{b} \geq 1 - \xi_i$ and $\xi_{\mathrm{i}} \geqq 0, i = 1, 2, 3, \ldots, l.$

where, C is punish coefficient. The bigger the c value is, the more severe the penalty would be.

SVMs are binary classifiers. Several methods have been designed to deal with multi-class classification problems (Hu *et al*., 2010; Hsu and Lin, 2002). One-against-one method is adopted in the current study. For a k-class classification problem, a total of k (k-1)/2 (k>2) classifiers are constructed and each of them trains the data derived from two different classes:

$$\min_{\mathrm{w^{ij}},\mathrm{b^{ij}},\xi^{\mathrm{ij}}} \frac{1}{2}(\mathrm{w^{ij}})^{\mathrm{T}}\mathrm{w^{ij}} + \mathrm{c}\sum_{t=1}^{l}\xi_t^{ij} \qquad (9)$$

Subject to $\frac{1}{2}(\mathrm{w^{ij}})^{\mathrm{T}}\phi(\chi_{\mathrm{t}}) + \mathrm{b^{ij}} \geq 1 - \xi_{\mathrm{t}}^{\mathrm{ij}}$, if $y_t = i$ and $\frac{1}{2}(\mathrm{w^{ij}})^{\mathrm{T}}\phi(\chi_{\mathrm{t}}) + \mathrm{b^{ij}} \leq -1 + \xi_{\mathrm{t}}^{\mathrm{ij}}$ if $y_t = j$ and $\xi_{\mathrm{t}}^{\mathrm{ij}} \geq 0, \mathrm{t} = 1,2, \ldots, \mathrm{l}, \mathrm{i}, \mathrm{j} = 1,2, \ldots \mathrm{K}.$

In this case, three classifiers were constructed to identify light, medium and heavy aroma types:

$$f_{ij} = sgn\left(\sum_{p=1}^{l} y_p\, a_p^{ij} K(x_p, x) - b^{ij}\right), \quad i, j = 1,2, \ldots, l \qquad (10)$$

Discriminant analysis: Discriminant analysis is a multivariate statistical analysis procedure where a data set containing p variables is separated into a number of previously defined groups using a linear combination of features. Given a set of p independent variables with known class k, discriminant analysis attempts to find linear combinations of the predictors (discriminant function, D). The function D Eq. (11) is expected to differentiate the k groups of samples as well as estimate groups membership and possibility according to the Fisher's discriminant procedure. For each group i (i = 1, …, k), the discriminant function D is defined as follows:

$$\mathrm{D_i} = \mathrm{b_{1i}x_{1i}} + \mathrm{b_{2i}x_{2i}} + \ldots + \mathrm{b_{pi}x_{pi}} + \mathrm{c_i},\ \mathrm{i} = 1, \ldots, \mathrm{k} \qquad (11)$$

where, b1, b2, …, bp are discriminant coefficients or scores, x1, x2, …, xp are independent variables and c is a constant. The centroids summarizing the group information are calculated as follows:

$$Xi = \begin{bmatrix} X1,i \\ X2,i \\ \ldots \\ Xk,i \end{bmatrix} \qquad (12)$$

where, X1, X2, …, Xp denote the mean values of the independent variables in the corresponding discriminant function for group i. The discriminant function D is designed with the aim of maximum distance between the centroids. The group membership for a new case is calculated based on the centroids. When the average discriminant score is lower than zero, the new case will be assigned to the group with lower centroids and vice versa.

Accuracy evaluation: Overall accuracy and kappa coefficient (Fleiss, 1971) are used to evaluate the

Table 1: Descriptive statistics of chemical compositions of tobacco leaves

Chemical compositions of tobacco leaf	Min.	Max.	Mean	S.D.	C.V (%)
Water-soluble total sugar (%)	16.11	39.50	28.67	4.36	15.22
Total plant alkaloid (%)	1.33	4.28	2.63	0.54	20.49
Protein (%)	3.38	6.77	4.94	0.74	15.14
Total nitrogen (%)	1.35	3.33	1.96	0.335	17.15
Reducing sugar (%)	13.03	35.04	25.08	3.68	14.68
Total volatile acid (%)	0.05	0.575	0.152	0.08	54.27
Total volatile alkali (%)	0.16	0.52	0.29	0.06	21.83
Ratio of nitrogen to nicotine	0.52	13.65	0.89	1.01	113.46
Ratio of sugar to nicotine	3.60	22.22	10.31	2.84	27.60
Ratio of potassium to chlorine	0.86	49.00	9.63	7.035	73.07
Petroleum ether extracts (%)	3.59	16.90	5.70	1.24	21.80
pH	3.38	5.67	5.37	0.418	7.83
Potassium (%)	0.90	3.35	1.93	0.47	24.16
Chlorine (%)	0.06	1.49	0.32	0.196	60.33
Nitrate (%)	0.002	0.63	0.067	0.074	112.64
Sulfate (%)	0.18	7.06	1.53	0.56	37.52
Ash content (%)	5.94	19.66	11.38	1.7	14.94
Alkalinity of water-soluble ash content (%)	0.11	1.48	0.54	0.26	48.64

Min.: Minimum; Max.: Maximum; S.D.: Standard deviation

models' accuracy. The results of classification were compared with their well acknowledged aroma types. Kappa coefficient is as follows:

$$\text{kappa} = \frac{P(A)-P(C)}{1-P(C)} \quad (13)$$

where, P(A) is the proportion of times that the methods agree and P(C) is the proportion of times that one expects them to agree by chance. Almost perfect agreement was yielded when 0.81 < kappa < 1; substantial agreement if 0.61 < kappa < 0.8; moderate agreement if 0.41 < kappa < 0.6; fair agreement if 0.21 < kappa < 0.4; slight agreement if 0.01 < kappa < 0.2 and poor agreement if kappa<0 (Landis and Koch, 1977). The relative improvement of overall accuracy and kappa coefficient were used to measure the improvement on the classification accuracy of the better performed models over the reference methods:

$$RI = \frac{CA_E - CA_R}{CA_R} \quad (14)$$

where, CAE and CAR are the overall accuracy or kappa coefficient of the better performed models and the reference method, respectively.
All analyses were done in Matlab 7.0.

Data: During the period of 2003 to 2007, 186 tobacco leaf samples with grade of C3F were collected from the representative counties planting flue-cured tobacco across China. Among them, 27 records with light, medium and heavy aroma types were used to train the classifiers and the remaining were unclassified samples. Eighteen routinely measured chemical compositions of flue-cured tobacco leaves including water-soluble total sugar, total plant alkaloid, total nitrogen, protein, reducing sugar, total volatile acids, total volatile alkali, ratio of nitrogen to nicotine, ratio of sugar to nicotine, ratio of potassium to chlorine, petroleum ether extracts, pH, potassium, chloride, nitrate, sulfate, ash content and alkalinity of water-soluble ash content were used to classify the aroma types in the current study. The descriptive statistics of these chemical parameters was given in Table 1.

The first five chemical compositions (water-soluble total sugar, total plant alkaloid, protein, total nitrogen and reducing sugar) were the most widely used indicators in evaluating flue-cured tobacco leave quality (Hu *et al*., 2010; Wang *et al*., 1998; Du *et al*., 2000) Therefore, these five chemical compositions were used as basic indicators. The others were added to the classifiers one after another. Finally, 42 classifiers developed by BP, SVM and DA were evaluated in the current study (Table 2).

RESULTS

The overall accuracy and kappa coefficients of the classifiers were shown in Table 2. The mean values of overall accuracy and kappa coefficient were 78% and 0.66 for BPs, 67% and 0.50 for SVMs and 67% and 0.50 for DAs. The results showed that BP models gave better performance than SVM and DA methods. The values of relative improvement of BP on SVM and DA were 16.5 and 33.6% for overall accuracy and kappa coefficient, respectively.

Compared with the basic BP classifier (BP5), BP models with more indicators have higher performance. The average relative improvements were 13 and 25% for overall accuracy and kappa coefficient, respectively (Table 2). Among them, the BP models with 9, 12 and 18 indicators had higher classification performances with relative improvement higher than 15 and 30% for overall accuracy and kappa coefficient, respectively. Hence, the BP model with fewer indicators and higher

Table 2: Overall accuracy (OA) and kappa coefficients of the classifiers

Model	OA (%)	Kappa	Model	OA (%)	Kappa	Model	OA (%)	Kappa
BP5	69.35	0.54	SVM5	61.83	0.43	DA5	66.67	0.50
BP6	73.12	0.59	SVM6	61.83	0.43	DA6	68.82	0.53
BP7	77.42	0.66	SVM7	69.35	0.54	DA7	69.35	0.54
BP8	79.57	0.69	SVM8	63.98	0.45	DA8	63.98	0.45
BP9	81.18	0.72	SVM9	65.05	0.47	DA9	65.05	0.47
BP10	77.42	0.66	SVM10	65.05	0.47	DA10	65.05	0.47
BP11	77.96	0.67	SVM11	68.82	0.53	DA11	68.82	0.53
BP12	80.11	0.70	SVM12	66.13	0.49	DA12	66.13	0.49
BP13	77.42	0.66	SVM13	67.20	0.50	DA13	67.20	0.50
BP14	77.42	0.66	SVM14	66.67	0.49	DA14	66.67	0.49
BP15	76.88	0.65	SVM15	62.37	0.43	DA15	62.37	0.43
BP16	79.03	0.68	SVM16	67.74	0.51	DA16	67.74	0.51
BP17	79.57	0.69	SVM17	70.97	0.56	DA17	70.97	0.56
BP18	80.65	0.71	SVM18	75.27	0.63	DA18	69.35	0.53
Mean	77.65	0.66		66.59	0.49		67.01	0.50

Table 3: Relative improvement of OA and kappa coefficients for BP classifiers

Model	OA (%)	Kappa (%)
BP5	-	-
BP6	5.44	9.26
BP7	11.64	22.22
BP8	14.74	27.78
BP9	17.06	33.33
BP10	11.64	22.22
BP11	12.42	24.07
BP12	15.52	29.63
BP13	11.64	22.22
BP14	11.64	22.22
BP15	10.86	20.37
BP16	13.96	25.93
BP17	14.74	27.78
BP18	16.29	31.48
Mean	12.89	24.50

performance (BP9) was the optimal classifier for identifying flue-cured tobacco aroma types. The confusion matrix was shown in Table 3. A total of 151 samples were classified into their well-known aroma types. The overall accuracy and kappa coefficient of BP9 were 81.18% and 0.72, respectively. These results suggested that the classification produced substantial agreement.

Table 4: Confusion matrix of BP9

Aroma type	Heavy	Medium	Light	Total
Heavy	46	5	0	51
Medium	4	53	6	63
Light	6	14	52	72
Total	56	72	58	186

The spatial distribution map of the flue-cured tobacco aroma types was built by ArcGIS 9.3 based on the optimal result produced by BP9 (Table 4). Most of the samples were classified into their well accepted types. The dominating areas producing flue-cured tobacco with heavy aroma type are Henan, Anhui, Jiangxi, Hunan, eastern regions of Shandong provinces. Areas producing flue-cured tobacco with

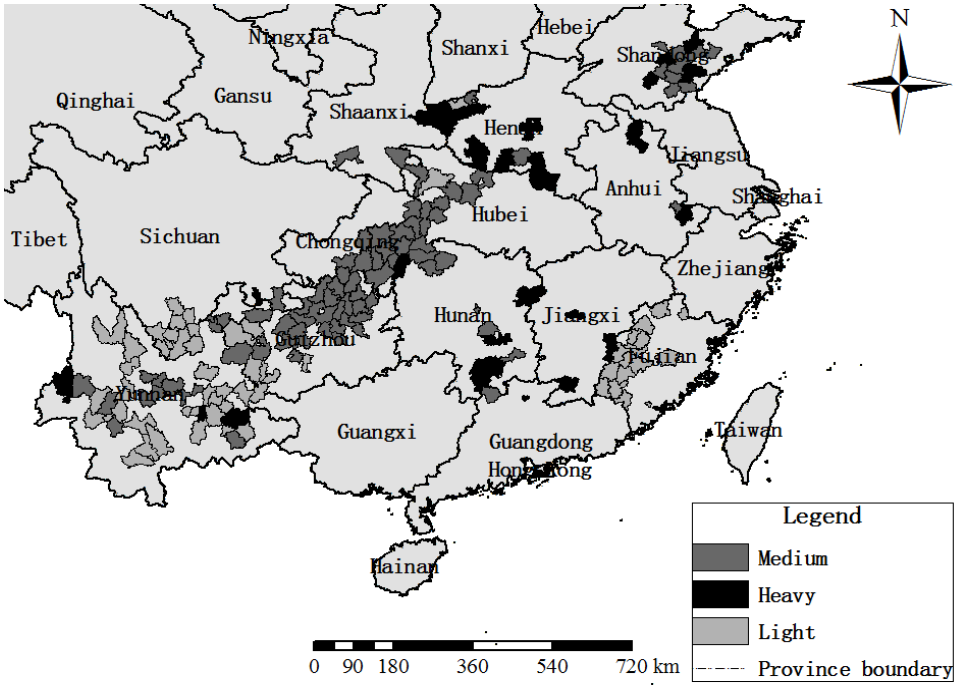

Fig. 3: Spatial distribution of flue-cured tobacco aroma types

medium aroma type are mostly concentrated in Guizhou, Chongqing, Hubei, central Shandong, northwestern Hunan and northeastern Yunnan provinces. Areas producing flue-cured tobacco with light aroma type are mainly distributed in Yunnan and Fujian provinces. However, not all results of the current study were similar to their well-known types. For instance, Yunnan province located in southwestern China is well-known for yielding flue-cured tobacco with light aroma type. In the current study, above half of samples in Yunnan were classified into light aroma type and others were medium and heavy aroma groups (Fig. 3). These results were agreement with the previous study (Yang *et al*., 2014) and further confirmed that care should be taken in regional planning and decision making for flue-cured tobacco production in these areas.

CONCLUSION

In this study, three kinds of supervised classifiers were evaluated for identifying flue-cured tobacco aroma types using routinely measured chemical compositions. The results showed that the BP model with 9 indicators outperformed others. The overall accuracy and kappa coefficient of BP9 was 81.18% and 0.72. The spatial distribution map showed that most samples were classified into their well-accepted aroma types. In the future, we will futherly explore the special chemical indice that can make the optimal model achieve the best effect.

REFERENCES

Bi, S.F., X.L. Zhu and C.Z. Ma, 2006. Application of stepwise discriminatory analysis in distinguishing aromas of flue-cured Tobacco in China. Chinese J. Trop. Crop., 27(4): 104-107.

Du, Y.M., C.F. Guo, H.B. Zhang, Y. Shang, X.L. Wang, J. Qiu and H.L. Ai, 2000. Study on relationship between content of water soluble sugar, alkaloid, total nitrogen and taste quality of flue cured tobacco. Chinese Tobacco Sci., 1: 7-10.

Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychol. Bull., 76(5): 378-382.

Hecht-Nielsen, R., 1989. Theory of the backpropagation neural network [C]. Proceeding of the IEEE International Joint Conference on Neural Networks (IJCNN, 1989), 1: 593-605.

Hsu, C.W. and C.J. Lin, 2002. A comparison of methods for multiclass support vector machines. IEEE T. Neural Networ., 13(2): 415-425.

Hu, J.J., M. Ma, Y.G. Li and C.Y. Yu, 2010. Grey incidence analysis on the correlation between main chemical components and sensory quality of flue-cured Tobacco. Tobacco Sci. Technol., Vol. 1, 2010.

Johnson, R.A. and D.W. Wichern, 1992. Applied Multivariate Statistical Analysis. 3rd Edn., Prentice-Hall, Englewood Cliffs, NJ, pp: 644.

Kavdır, I., 2004. Discrimination of sunflower, weed and soil by artificial neural networks. Comput. Electron. Agr., 44(2): 153-160.

Kolios, S. and C.D. Stylios, 2013. Identification of land cover/land use changes in the greater area of the Preveza peninsula in Greece using Landsat satellite data. Appl. Geogr., 40: 150-160.

Landis, J.R. and G.G. Koch, 1977. The measurement of observer agreement for categorical data. Biometrics, 33(1): 159-174.

Li, Z.H., N.R. Wang, D.S. Wang, X.L. Zhu and H.L. Zhou, 2009. Preliminary study of aroma type styles of flue-cured tobacco in different ecological scale regions. Chinese Tobacco Sci., 30(5): 67-70, 76.

Marey-Pérez, M.F. and V. Rodríguez-Vicente, 2011. Factors determining forest management by farmers in northwest Spain: Application of discriminant analysis. Forest Policy Econ., 13(5): 318-327.

Nieuwenhuizen, A.T., J.W. Hofstee, J.C. van de Zande, J. Meuleman and E.J. van Henten, 2010. Classification of sugar beet and volunteer potato reflection spectra with a neural network and statistical discriminant analysis to select discriminative wavelengths. Comput. Electron. Agr., 73(2): 146-153.

Riveiro-Valiño, J.A., C.J. Álvarez-López and M.F. Marey-Pérez, 2009. The use of discriminant analysis to validate a methodology for classifying farms based on a combinatorial algorithm. Comput. Electron. Agr., 66(2): 113-120.

Wang, Y.B., B.H. Wang, C.F. Guo, F.L. Wang and J. Zhou, 1998. Study on the main chemical components related to smoking quality in flue-curred tobacco. Sci. Agr. Sinica, 31(1): 89-91.

Widrow, B., 1988. DARPA Neural Network Study. AFCEA Int. Press, Fairfax, VA.

Yang, C., W. Wu, S.C. Wu, H.B. Liu and Q. Peng, 2014. Aroma types of flue-cured tobacco in China: Spatial distribution and association with climatic factors. Theor. Appl. Climatol., 115(3): 541-549.

Zhang, J., F.F. Zhou, G.B. Deng, C.T. Mao, C.Y. Bao, J.C. Rao and X.L. Zhang, 2013. Discriminant analysis of aroma types of upper leaf in flue-cured tobacco based on chemical constituents and aroma components. J. Hunan Agric. Univ. Nat. Sci., 39(3): 232-241.

Zheng, B.J., S.W. Myint, P.S. Thenkabail and R.M. Aggarwal, 2015. A support vector machine to identify irrigated crop types using time-series Landsat NDVI data. Int. J. Appl. Earth Obs., 34: 103-112.