Talanta 82 (2010) 1571–1575

Contents lists available at ScienceDirect

Talanta

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

# A random forest of combined features in the classification of cut tobacco based on gas chromatography fingerprinting

Xiaohui Lin[a], Lie Sun[a], Yong Li[b], Ziming Guo[c], Yanli Li[b], Kejun Zhong[c], Quancai Wang[a], Xin Lu[b], Yuansheng Yang[a], Guowang Xu[b,*]

[a] *Department of Computer Science & Engineering, Dalian University of Technology, Dalian 116024, China*
[b] *CAS Key Laboratory of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China*
[c] *Technology Center of China Tobacco Hunan Industrial Corporation, Changsha 410007, China*

ARTICLE INFO

*Article history:*
Received 20 April 2010
Received in revised form 20 July 2010
Accepted 22 July 2010
Available online 30 July 2010

*Keywords:*
Random forest
Combined features
Cut tobacco
GC–TOF MS

ABSTRACT

We applied the random forest method to discriminate among different kinds of cut tobacco. To overcome the influence of the descending resolution caused by column pollution and the subsequent deterioration of column efficacy at different testing times, we constructed combined peaks by summing the peaks over a specific elution time interval $\Delta t$. On constructing tree classifiers, both the original peaks and the combined peaks were considered. A data set of 75 samples from three grades of the same tobacco brand was used to evaluate our method. Two parameters of the random forest were optimized using out-of-bag error, and the relationship between $\Delta t$ and classification rate was investigated. Experiments show that partial least squares discriminant analysis was not suitable because of the overfitting, and the random forest with the combined features performed more accurately than Naïve Bayes, support vector machines, bootstrap aggregating and the random forest using only its original features.

## 1. Introduction

Quality-control in cigarettes is attracting more and more attention recently. The main differences among different cigarette brands or grades are the contents of fragrances and potentially hazardous substances, such as tar and nicotine [1,2]. The quality of cigarettes is usually evaluated using visual aspects and oral and olfactory sensory criteria. In many cases, such inspections strongly rely on the operator's experience. The evaluation results are thus unreliable and may differ with various estimators. Therefore, it is necessary to be able to assess different cigarette brands or grades using more scientific techniques.

Several kinds of instrumental methods, including gas chromatography (GC) [3], gas chromatography–mass spectrometry (GC–MS) [4] and near-infrared (NIR) spectrometry [5,6], have been used to collect chemical fingerprint information for tobacco. Multidimensional separation methods, including comprehensive two-dimensional gas chromatography method [7,8], were also used for tobacco components analysis. However, instrument operation and data processing become more complex when multidimensional separation methods were used for quantitative analysis of tobacco profiling. Gas chromatography-flame ionization detection (GC-FID), with its high sensitivity, good stability and popularity, is a powerful tool for tobacco fingerprint collection. The present work aimed to evaluate the possibility of developing a classification method able to discriminate different kinds of cigarettes based on the GC-FID fingerprint data of cut tobacco.

Many multivariate analysis techniques, such as partial least squares discriminant analysis (PLS-DA) [9], support vector machines (SVM) [10] and Naïve Bayes [11], have been used for class discrimination. In addition, it has been shown that ensemble techniques, such as boosting [12], bootstrap aggregating (bagging) [13] and random forest (RF) [14], can significantly improve learning performance [15].

RF, originally developed by Leo Breiman, is one of the most successful ensemble methods [14]. It is a classifier consisting of a collection of tree-type classifiers [14–17]. Each tree-type classifier uses a unique training set constructed by boostraping. As the RF method introduces randomness on the basis of bagging, it can thus be considered a further development of bagging. It has good predictive performance when compared with current popular classification algorithms such as SVM [18,19]. As a type of supervised method, it has no problems with overfitting due to the use of the strong law of large numbers [14]. Since it was proposed, RF has become a well-known data-analysis method. It has been applied to a wide variety of scientific areas, including microarray data [20], quantitative structure–activity relationship (QSAR) modeling [21,22], land cover [23] and the prediction of protein interactions [24].

[*] Corresponding author. Tel.: +86 411 84379531; fax: +86 411 84379559.
*E-mail address:* xugw@dicp.ac.cn (G. Xu).

0039-9140/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2010.07.053

The RF method combines two different forms of randomness: the random selection of the features and the random linear combination of the original features [14]. The chemical fingerprint information of a sample consists of separated peaks. Each peak represents at least one chemical component. The distribution of peaks and their areas (or heights) is unique for each sample. In the ideal case, the degree of separation between different peaks for a given sample is the same in different trials in the same analytical system and operational conditions. However, in practice they tend to decrease with the number of analyzed samples because of the loss of resolution of the chromatography column. Typically, some small peaks will disappear or merge with their neighbors, and peak alignment among different chromatograms will decline, which can result in an incorrect classification. To solve this problem, in this work we defined combined features (peaks) by combining features (peaks) in a specified elution time zone. A combined random forest (CRF) based on both the original features (peaks) and combined features (peaks) was constructed and applied to process chemical fingerprint information from samples of cut tobacco. The data of three grades of "Furong" series cut tobacco were used to demonstrate the CRF method.

## 2. Theory of random forests

Let $A = \{a_1, a_2, \ldots, a_m\}$ be a set of features (here, peaks); $x$ denotes the input vector, and $f(a_i, x)$ denotes the $i$th feature value of the sample $x$. Here $f(a_i, x)$ is the area of the $i$th peak, which represents the concentration of the peak component(s).

A random forest consists of *ntree* tree classifiers $\{h(x,\theta_k), k = 1,2, \ldots, ntree\}$ in which the $\{\theta_k\}$ are independent, uniformly random vectors [14]. For classification, each tree contributes a unit vote at the input sample $x$. The output of the classifier is determined by the majority votes of all of the trees in the forest, i.e.,

$$c^* = \mathrm{argmax}_c \left|\{\theta_k : h(x, \theta_k) = c, 1 \le k \le ntree\}\right| \quad (1)$$

The random forest benefits from two powerful machine-learning techniques: bagging [12] and random subspace selection [14]. First, on constructing a tree classifier, a bootstrap sample is drawn from the original observations. The samples that are not in the bootstrap sample are called out-of-bag (OOB) data. The OOB data (about 37% of the total data) can be used to estimate prediction error. Second, as each node of a tree is split, the best split is chosen from a random subset of the features. The best split [14] is the one yielding the maximum in the expected reduction of the overall impurity value, which is defined as follows:

$$\Delta I_m(LS, a_i) = I_m(LS) - \sum_a \frac{|LS_a|}{|LS|} I_m(LS_a) \quad (2)$$

where $LS$ is the data set at the node and $LS_a$ is the subset of samples from $LS$ such that the samples in $LS_a$ have the same value $a$ for feature $a_i$, that is, $LS_a = \{x|x{\in}LS, f(a_i, x) = a\}$. There are many tools available to measure the impurity $I_m(LS, a_i)$. Here we use Shannon's entropy [25] as the impurity measure.

By defining the margin function for a random forest and using the law of large numbers as a theoretical foundation, Breiman proved that the generalization error [14] of a random forest tends to a limited upper bound with an increase of the number of trees. The generalization error depends on two aspects: (1) a greater strength of the individual classifiers in the random forest leads to a better the performance of the forest; and (2) a lower correlation between the individual trees yields a better performance of a random forest. Formula (3) gives an upper bound for the generalization error [14]:

$$PE* \le \frac{\bar{\rho}(1 - s^2)}{s^2} \quad (3)$$

where $\bar{\rho}$ is the mean value of the correlation between the individual trees and $s$ is the strength of the individual trees.

To lower the similarity between the individual trees and thus obtain low-bias trees, each tree is grown to the largest size and is unpruned [14].

## 3. Experimental

### 3.1. Materials, chemicals and reagents

A total of 75 cut tobacco samples from three grades of the same cigarette brand were kindly provided by the China Tobacco Hunan Industry Corporation, Changde, Hunan, China (25 samples of each grade). Each grade of cut tobacco samples were produced by using routine methods from the product line with different levels of crude tobacco leaves and fragrances. The cigarettes were unwrapped, and the cut tobacco was collected and milled into powder (40 mesh). The quality-control (QC) sample was a mixture of equal amounts of the 75 testing samples.

The internal standard 2-methylnaphthalene was purchased from Sigma–Aldrich (Beijing, China). Dichloromethane (analytical grade) was purchased from Dikma (Beijing, China).

### 3.2. Accelerated solvent extraction

A Dionex ASE200 accelerated solvent extractor (CA, USA) equipped with 11-mL stainless steel extraction cells, and 60-mL glass collection bottles was used for the accelerated solvent extractions (ASE). Each extraction cell was filled with 4.0 g of tobacco powder, and 200 μL of internal standard solution (0.15 mg mL$^{-1}$) was spiked into the tobacco powder. The cell was then loaded into the autosampler tray. Extraction conditions were as follows: static extraction time, 5 min; extraction cycles, 2; extraction pressure, 1500 psi; and extraction temperature, 100 °C. The extraction solution was collected and then condensed to 1 mL with a rotary evaporator at atmospheric pressure. The condensed solution was filtered and stored in a 1.5-mL screw-capped vial for instrumental analysis.

### 3.3. GC-FID analysis

The gas chromatographic analyses were carried out on an Agilent 6890 GC system (Agilent Technologies, USA) equipped with an autosampler. Separations were conducted using a DB-5 MS, 30 m × 0.25 mm capillary column coated with a 0.25-μm stationary phase film (Agilent Technologies, Palo Alto, CA, USA). Helium was used as the carrier gas at a flow rate of 1.2 mL min$^{-1}$. The column temperature was programmed at 50 °C for 1 min, raised to 220 °C at 8 °C min$^{-1}$, held for 7 min and then raised to 280 °C at 15 °C min$^{-1}$. Finally, the temperature was held at 280 °C for 20 min. All of the samples were analyzed in split mode (1 μL injection, split rate, 10:1). The FID heater temperature was held at 280 °C. The flow rates of hydrogen, air and nitrogen in the FID were 40, 450 and 45 mL min$^{-1}$, respectively.

### 3.4. Peak alignment and data preprocessing

#### 3.4.1. Peak alignment

The RF is a type of supervised learning method in which the data are randomly divided into two parts, a training set, $T_R$, containing about two-thirds of the samples, and a test set, $T_E$, containing the remaining samples. Let $T_R = \{x_1, x_2, \ldots, x_{nR}\}$, where $n_R$ is the size of the training set.

We used our own program written in C++ to align the samples. First, when constructing each tree classifier in the forest, the samples in the training set $T_R$ were aligned to obtain a peak table. A

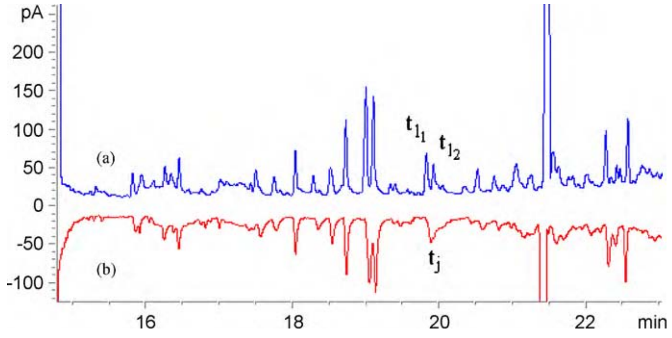

**Fig. 1.** Deterioration of chromatography resolution with the injection time: (a) first injection; (b) 200th injection. Plots of the two injections were mirrored to the *x*-axis to make a distinct comparison. Plots of the two injections were mirrored to the *x*-axis to make a distinct comparison.

typical chromatogram in $T_R$ was selected as the reference. According to the retention time stability of the internal standard, the match window was set to 0.18 min. All of the other samples in $T_R$ were compared with the reference chromatogram. The peaks of the samples lying in the same retention time windows were regarded as the same peaks. The peak areas of a sample are divided by the area of the internal standard.

To remove the unstable peaks (features) in the chromatography data, we used QC samples to filter out the peaks whose peak area RSDs (relative standard deviations) were larger than 25%. The peaks having MSDs (mean squared deviations) in a peak area <0.0005 were also removed because of their small contribution to the classification. Second, during classification, the input sample was aligned to the peak table before the corresponding tree casts a vote on it.

### 3.4.2. Normalization

In multivariate analysis, the influence of the scale of the variables needs to be considered. The peaks in the samples represented a great range of concentrations; hence, before the tree classifiers were constructed, the peak area data were first normalized as follows:

$$f'(a_j, x_i) = \frac{f(a_j, x_i)}{\sum_{k=1}^{m} f(a_k, x_i)} \qquad (4)$$

where $i = 1, 2, \ldots, n_R$ and $j = 1, 2, \ldots, m$.

For prediction, when tree classifier $h(x,\theta_k)$, $k = 1, 2, \ldots, ntree$, receives an input vector $x$, it first aligns $x$ with the peak table and then normalizes $x$ as done in the training set. Finally, each tree casts a unit vote at $x$.

### 3.4.3. Combined peaks

In the separation of complex samples, it is very often observed that the chemical fingerprints obtained at different times for the same sample display differences, and the degree of separation may decrease over time due to the decline in column efficacy from column pollution; some small peaks may disappear or merge with their neighbors (see Fig. 1), leading to poor peak alignment and incorrect classification.

To overcome the influence of resolution deterioration and improve the performance of the RF, we considered the combined peaks, which represent the sum of all of the peaks in a specific retention time zone.

Let $t_j$ denote the retention time of peak (feature) $a_j$ ($1 \le j \le m$). With a given time interval $\Delta t$, for any $0 < l_1 < l_2 \le m$ (Fig. 1), and observation $x$, if $0 < t_{l_2} - t_{l_1} \le \Delta t$, $\Delta t < t_{l_2+1} - t_{l_1}$, and $\Delta t < t_{l_2} - t_{l_1-1}$, we have:

$$f(a_{new}, x) = \sum_{k=l_1}^{l_2} f(a_k, x) \qquad (5)$$

where $a_{new}$ is a new combined peak (feature) that contains all peaks in the time interval $\Delta t$. Assume that there are $m^*$ combined peaks. Letting $A' = \{a_1, a_2, \ldots, a_m, a_{m+1}, \ldots, a_{m+m^*}\}$, we then take $A'$ instead of $A$ to be the feature set. $A'$ contains not only the original separated peaks but also the combined peaks; it can thus compensate for the errors caused by integral and peak mismatch due to resolution change.

Obviously, the selection of the $\Delta t$ value has an important influence on the classification rate of the model. In the next section we investigate the value of $\Delta t$ through experiments.

### 3.5. SVM, Naïve Bayes and bagging

SVM, Naïve Bayes [26] and bagging are three popular multivariate analysis techniques. In order to evaluate our method, we also used them to handle cut tobacco data. In SVM the radial basis function was selected as the kernel function, a “grid search” on the two parameters $C$ and $\gamma$ was used based on cross-validation. For Bagging we took decision trees as its base classifiers. The main difference of it from RF is that all the features (peaks) are used as candidate features on building a tree classifier.

## 4. Results and discussion

### 4.1. GC fingerprint analysis of cut tobacco

A GC method was used to obtain chemical fingerprints of cut tobacco (Fig. 2). To evaluate and regularize the validity of the fingerprinting data, QC samples were inserted into the sample analysis sequence at the rate of 10 test samples for each QC sample. The reproducibility of the QC samples was calculated and is shown in Fig. 3. Usually, peaks with area RSDs over 25% with respect to

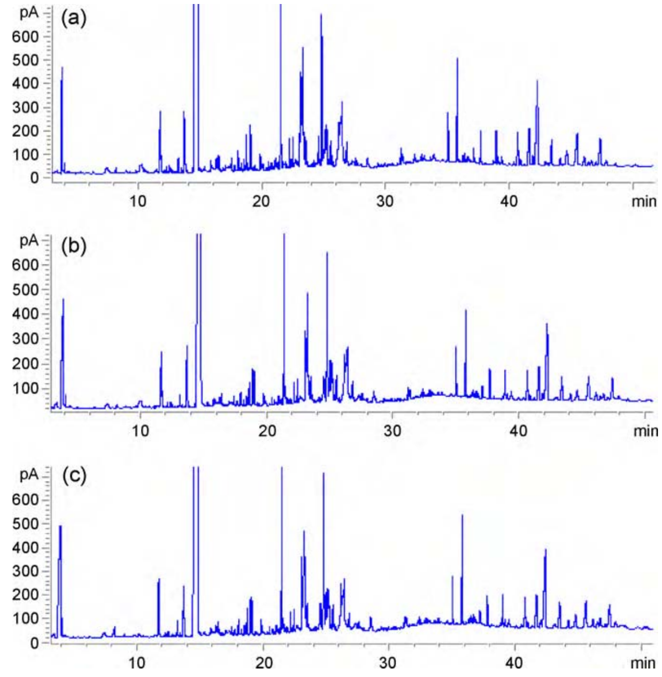

**Fig. 2.** Three typical gas chromatograms of “Furong” series cut tobacco. (a), (b) and (c) were from class 1, class 2 and class 3, respectively.

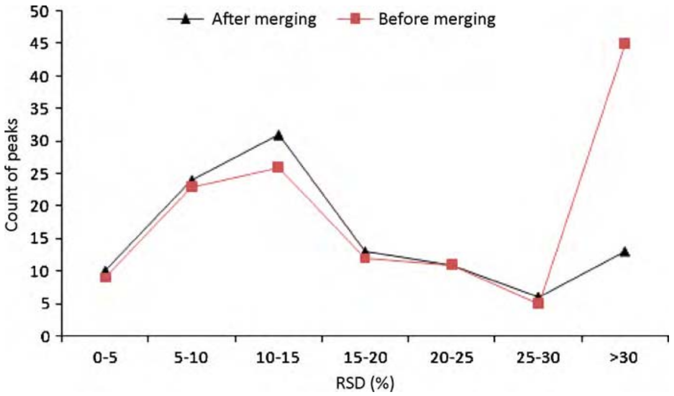

**Fig. 3.** RSD (%) distribution of relative peak areas before and after manually merging incompletely separated peaks.

the QC samples are considered unstable peaks and are not used for classification. In this experiment, 38% of the peaks (comprising 3.4% of the total peak areas) had RSDs greater than 25% in the aligned raw data and should thus be removed, resulting in a great loss of raw information. The unacceptable RSD of these peaks may be attributed to the decrease of separation ability of the chromatography column because of the long analysis period (several days or several months) or the contamination of the column; some peaks were separated in some chromatograms and are overlapping in other chromatograms (Fig. 1). Because of this overlap, the accurate alignment of these unstable peaks and the integration of their areas with the workstation became impossible. To use the information represented by these peaks, merging of the neighboring, incompletely separated peaks is an option. After the merging operation, the number of peaks with RSDs over 30% markedly decreased (Fig. 3).

### 4.2. Parameter optimization

There are three important parameters in CRF: the number of the trees in the forest (*ntree*), the number of the peaks randomly selected as the candidates for splitting at each node ($m_{try}$) and the time interval selected to construct combined peaks ($\Delta t$). We used the OOB error to evaluate the effects of different settings of *ntree* and $m_{try}$. Fig. 4 shows the relationship of the OOB error rate with the two parameters. If *ntree* was large enough, the OOB error tended to be limited by an upper bound. It can be observed that for the "Furong" series cut tobacco data set, 300 trees were sufficient. In addition, $m_{try} = \sqrt{m}$ was the best choice based on the OOB error rate [12].

Because of the complexity of the cut tobacco samples, peak overlap is unavoidable. Furthermore, all of the tests cannot be finished at the same time, so resolution deterioration is also inevitable. The features with serious peak overlap had large quantitative deviations. To overcome this disadvantage, we combined the original independent features into a combined feature in a specified chromatography elution time zone $\Delta t$. The selection of the time interval $\Delta t$ is important for the model performance. Fig. 5 shows how varying $\Delta t$ influenced the predicting results.

When $\Delta t = 0$, the random forest only considered the original individual peaks, and its classification rate was 91.89%. As expected, the random forest CRF performed better than the basic RF. It is found that $\Delta t = 0.25$ min was the best choice, with a classification rate of 93.74%. Therefore, CRF outperformed the RF considering only original peaks. As the degree of separation may be different in different tests, it is not unexpected that peaks may merge with neighboring peaks in a given test while, in other tests, the corresponding peaks are separated. The CRF can process this case perfectly, reducing the effect of misalignment due to resolution change on classification rate.

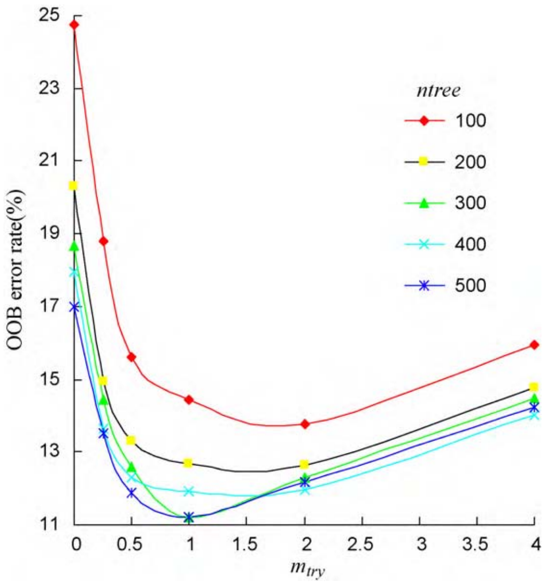

**Fig. 4.** Relationship of OOB error rate with *ntree* and $m_{try}$. The value of the abscissa is the coefficient of $\sqrt{m}$, where *m* is the number of peaks. If the coefficient is 0, the number of peaks tried at each split is 1. *ntree* represents the number of the trees in the random forest.

### 4.3. Comparison with other classification methods

Fingerprint data from gas chromatography contains many variables; a single variable or feature is usually not enough to distinguish different classes. Therefore, many multivariate analysis techniques, such as PLS-DA and SVM, have been applied to discriminate between different sample classes. Here we compared our CRF method with four popular classification methods, PLS-DA, SVM, Naïve Bayes and bagging, in the analysis of the GC-FID data of cut tobacco.

Fig. 6a shows the score plot of PLS-DA for the "Furong" series cut tobacco data. Here, $R^2Y$ and $Q^2$ are 0.928 and 0.779, respectively, and the three classes of cut tobacco are well separated. PLS-DA is a kind of supervised method; thus, the possibility of overfitting must be considered. Cross-validation was used for model validation; 20 permutations for the model were used, and the results are shown in Fig. 6b. Because the $Q^2$-intercept of the PLS-DA model was <0.05 while the $R^2$-intercept was larger than 0.4, overfitting occurred [27]; thus, PLS-DA failed to accurately model the "Furong" series cut tobacco data.

Although RF is also a kind of supervised method, it does not incur the problem of overfitting due to the use of the strong law of large

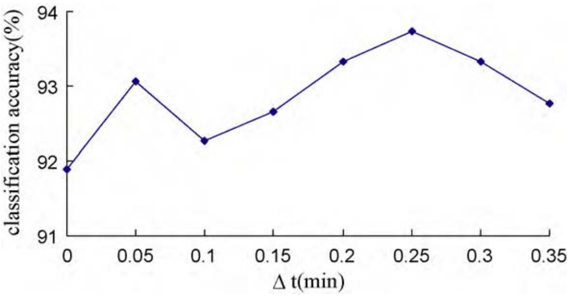

**Fig. 5.** Determination of the optimum $\Delta t$. The prediction classification rate at each $\Delta t$ is the average of the accuracies obtained by running the algorithm 100 times.

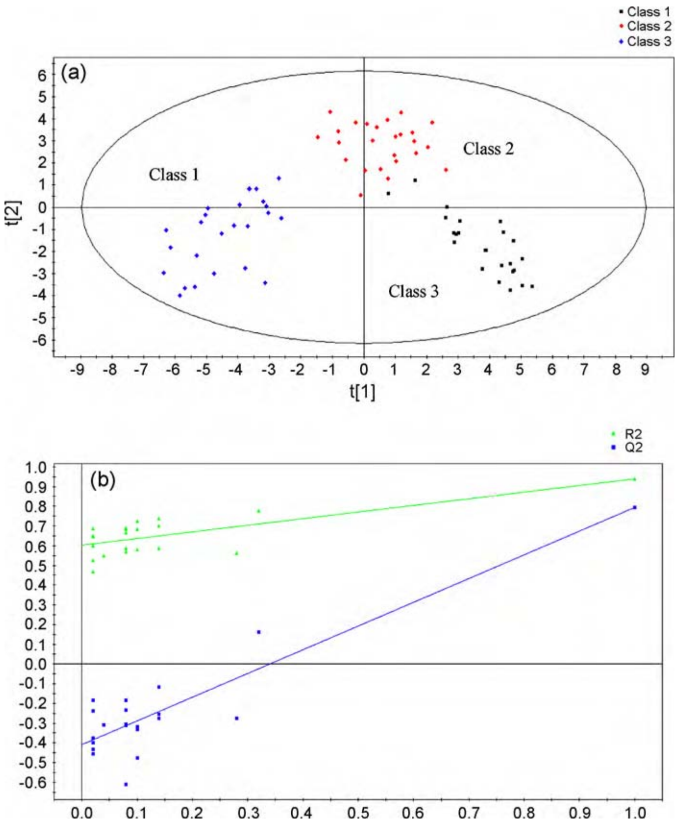

**Fig. 6.** PLS-DA for the overall set of 75 cigarette samples: (a) score plot; (b) permutation test plot on the first Y vector.

numbers [14]. To compare its classification rate with SVM, Naïve Bayes and bagging, we selected the same training and test sets as the CRF did. All of the preprocesses were the same for each algorithm as for the CRF, such as deleting the features for QC samples with RSDs more than 25% or MSDs <0.0005. It was found that the accuracies of the CRF, RF, SVM, Naïve Bayes and bagging methods were 93.74%, 91.89%, 89.22%, 87.00% and 84.59%, respectively.

Although SVM is a very popular classification algorithm that has shown good performance in a variety of classification tasks, its classification rate was 4.52% lower than CRF and 2.67% lower than RF for the cut tobacco data.

The Naïve Bayes's classification rate for this cut tobacco set was 87.00%, lower than both RF and CRF. Bagging is also a kind of ensemble classifier based on tree classifiers, but here it achieved a classification rate of only 84.59%.

Hence the RF, which considers only original peaks, outperformed SVM, Naïve Bayes and bagging in the discrimination of the tobacco fingerprints. Furthermore, our CRF method, which constructs the tree classifiers with both original peaks and combined peaks, obtained a classification rate of 93.74%, which is 1.85% higher than the RF, the variance of the correct classifications of CRF and RF are 5.36% and 5.73%, respectively. We used *t*-test to compare the classification rates of the CRF and RF, and the *p* value was 0.008 meaning that the data from two methods are statistically different.

## 5. Conclusions

We constructed a random forest based on combined features (peaks) according to the characteristics of cut tobacco data. During the building of tree classifiers, we considered both the original peaks and their combinations in a specified time interval $\Delta t$. Then we investigated its application in the classification of cigarette analysis data. In a case study involving three different grades of "Furong" series cigarettes, we first evaluated the effects of the changes of the random forest parameters *ntree* and $m_{try}$ and then discussed the choice of the specified time interval $\Delta t$. The experimental results showed that our CRF algorithm outperformed RF (the random forest with the original individual features), especially when $\Delta t = 0.25$; the classification rate of CRF was 1.85% higher than that of RF. Furthermore, both RF and CRF performed better than SVM, Naïve Bayes and bagging. While PLS-DA also showed good predictive capability, the model suffered from overfitting. Hence, the random forest of combined features was more suitable for analyzing fingerprint data with peak overlap or resolution deterioration.

## Conflict of interest

The authors declared that they have no conflicts of interest.

## Acknowledgements

The study has been supported by the State Key Science & Technology Project for Infectious Diseases (2008ZX10002-019), the National Natural Science Foundation of China (20835006), the Scientific Foundation of State Tobacco Monopoly Administration of China (110200701005) and the project (07A01) from China National Tobacco Corporation Yunnan Provincial Company.

## References

[1] S.D. Bolboaca, L. Jantschi, Int. J. Environ. Res. Public Health 4 (2007) 233–242.
[2] J.F. Pankow, J.E. Henningfield, B.E. Garrett, Nicotin Tob. Res. 6 (2004) 199.
[3] L. Yu, Y.C. Zhang, C.P. Zhou, M.L. Wang, B.Z. Liu, Acta Tabacaria Sin. 13 (2007) 18.
[4] X.L. Zhu, Y. Gao, Z.Y. Chen, Q.D. Su, Chromatographia 69 (2009) 735–742.
[5] L.J. Ni, L.G. Zhang, J. Xie, J.Q. Luo, Anal. Chim. Acta 633 (2009) 43–50.
[6] E.D. Moreira, M.J. Pontes, R.K. Galvao, M.C. Araujo, Talanta 79 (2009) 1260–1264.
[7] X. Lu, J.L. Cai, H.W. Kong, M. Wu, R.X. Hua, M.Y. Zhao, J.F. Liu, G.W. Xu, Anal. Chem. 75 (2003) 4441–4451.
[8] H.F. Li, K.J. Zhong, X. Lu, C.M. Bai, J.G. Huang, H.L. Lu, C.F. Ma, S.K. Zhu, H.W. Kong, M.Y. Zhao, J.P. Xie, S. Niu, G.W. Xu, Acta Chim. Sin. 64 (2006) 1897–1903.
[9] H. Keun, T. Ebbels, H. Antti, M. Bollard, O. Beckonert, E. Holmes, Anal. Chim. Acta 490 (2003) 265–276.
[10] S. Mahadevan, S.L. Shah, T.J. Marrie, C.M. Slupsky, Anal. Chem. 80 (2008) 7562–7570.
[11] L. Fan, K. Pho, P. Zhou, Expert Syst. Appl. 36 (2009) 9919–9923.
[12] Y. Freund, R.E. Schapire, Lecture Notes Comput. Sci. 904 (1995) 23–37.
[13] L. Breiman, Mach. Learn. 24 (1996) 123–140.
[14] L. Breiman, Mach. Learn. 45 (2001) 5–32.
[15] M. Robnik-Sikonja, Machine Learning, ECML 2004, in: Proceedings, Springer, Berlin, 2004, pp. 359–370.
[16] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Chapman & Hall, New York, 1984.
[17] J.R. Quinlan, Mach. Learn. 1 (1986) 81–106.
[18] R. Diaz-Uriarte, S.A. Andres, BMC Bioinform. 7 (2006) 3.
[19] A.G. Heidema, J.M. Boer, N. Nagelkerke, E.C. Mariman, D.L. van der, A.E.J.M. Feskens, BMC Genet. 7 (2006) 23.
[20] T. Shi, D. Seligson, A.S. Belldegrun, A. Palotie, S. Horcath, Mod. Pathol. 18 (2005) 547–557.
[21] R. Guha, P.C. Jurs, J. Chem. Inf. Comput. Sci. 44 (2004) 2179–2189.
[22] V. Svetnik, A. Liaw, C. Tong, J.C. Cullberson, R.P. Sheridan, B.P. Feuston, J. Chem. Inf. Comput. Sci. 43 (2003) 1947–1958.
[23] P.O. Gislason, J.A. Benediktsson, J.R. Sveinsson, Pattern Recogn. Lett. 27 (2006) 294–300.
[24] Y. Qi, Z. Bar-Joseph, J. Klein-Seetharaman, Proteins 63 (2006) 490.
[25] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 2006.
[26] A.A.T. Bui, R.K. Taira, Medical Imaging Informatics, Springer–Verlag, 2009.
[27] L. Eriksson, E. Johansson, N. Kettaneh-Wole, J. Trygg, C. Wikstrom, S. Wold, Multi-and Megavariate Data Analysis – Principles and Applications, Umetrics AB, Umea, 2001.