Journal of Pharmaceutical and Biomedical Analysis 49 (2009) 746–752

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

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

Early prediction of lung cancer based on the combination of trace element analysis in urine and an Adaboost algorithm

Chao Tan [a,d,*], Hui Chen [b,c], Chengyun Xia [c]

[a] Department of Chemistry and Chemical Engineering, Yibin University, Yibin, Sichuan, 644007, PR China
[b] Hospital, Yibin University, Yibin, Sichuan, 644007, PR China
[c] Clinical College, North Sichuan Medical College, Nanchong, Sichuan, 637000, PR China
[d] Yibin University Key Laboratory of Computational Physics, Yibin University, Yibin, Sichuan, 644007, PR China

ARTICLE INFO

Article history:
Received 2 September 2008
Received in revised form
29 November 2008
Accepted 10 December 2008
Available online 24 December 2008

Keywords:
Trace element
Lung cancer
Adaboost
Chemometrics
Diagnostics

ABSTRACT

Early detection of cancer is the key to effective treatment and long-term survival. Lung cancer is one of the most frequently occurring cancers and its early detection is particularly of interest. This work investigates the feasibility of a combination of Adaboost (*ensemble* from machining learning) using decision stumps as weak classifier and trace element analysis for predicting early lung cancer. A dataset involving the determination of 9 trace elements of 122 urine samples is used for illustration. Kennard and Stone (KS) algorithm coupled with an alternate re-sampling was used to realize sample set partitioning. The whole dataset was split into equally sized training and test set, which were then reversed to yield a second operating case, we called them case A and case B, respectively. The prediction results based on the Adaboost were compared with those from Fisher discriminant analysis (FDA). On the test set, the final Adaboost classifiers achieved a sensitivity of 100% for both cases, a specificity of 93.8%, 95.7%, and an overall accuracy of 95.1%, 96.7%, for case A and case B, respectively. In either case, Adaboost always achieves better performance than FDA; also, it is less sensitive to the composition of the training set compared to FDA and easy to control over-fitting. It seems that Adaboost is superior to FDA in the present task, indicating that integrating Adaboost and trace element analysis of urine can serve as a useful tool for diagnosing early lung cancer in clinical practice.

1. Introduction

About 25 elements are recognized as essential for human life. Most of them are present in trace amounts but part of metalloenzymes and participate in biological functions, such as oxygen transport, free radical scavenging, structural organization of macromolecules, and hormonal activity, and are therefore essential for the functioning of the cells [1,2]. For example, iron plays a role in the growth of cancers and iron chelators have anti-proliferative activity on cancer cell line [3]. As many different kinds of interactions among various trace elements exist, for a healthy individual, there always exists a dynamic balance (at optimum biological levels) which is responsible for numerous metabolic and physiological processes in the human body [4]. The disorder of trace element balance is often related to some pathologic conditions that lead to many diseases. So, the studies on the relationship between trace elements and various diseases are valuable and have attracted considerable attention over the past two decades [5–15]. However, such relationships are also quite complicated and are difficult to explain very satisfactorily through the investigation of one or a few trace elements. Therefore, the reliable, robust prediction of a certain disease based on trace element analysis requires special methodology, *i.e.*, suitable chemometrics [16–18], to construct a classification model for distinguishing healthy and unhealthy subjects.

Lung cancer is the second most common cancer in humans and is the most common cause of cancer deaths in the world. The overall 5-year survival rate of patients with lung cancer is no greater than 14%, which is much lower than that for patients with cancers in other organs, such as the bladder, breast, colon, cervix, and prostate [19,20]. So, early detection of lung cancer is crucial for the successful application of specific therapies to reduce mortality rate or to facilitate a full care. However, because early lung cancers or precancers such as dysplasia and carcinoma *in situ* (CIS) are only a few cell layers thick (0.2–1 mm) and present few symptoms, they can be very difficult to be visually detected by conventional diagnostic methods such as medical imaging. In clinical practice, about 80% cases have already evolved into the advanced stage when first discovered and confirmed, accordingly losing the most suitable opportunity of surgical treatment. Evidently, finding lung cancer as early as possible has important clinical significance. Nowadays, it has been discovered that lung cancer incidence is usually related to risk factors

* Corresponding author. Tel.: +86 831 35510808.
E-mail address: chaotan1112@163.com (C. Tan).

0731-7085/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2008.12.010

which range from behavioral, genetic, occupational, nutritional, and other. In the occurrence and development of lung cancer, changes of some trace elements in the urine can be detected, which conversely reflect the status of human nutrition and metabolism. Based on this, it is possible to predict early lung cancer or the risk of its occurrence by the combination of trace element analysis and chemometrics. Moreover, compared to some newly developed techniques such as fluorescence and Raman spectroscopy [20–22], trace element analysis may be more preferable in practice due to its lower cost and no invasion to a subject.

Focus in the present study is to show how Adaboost (an *ensemble* strategy from machining learning) using decision stumps as weak classifier, coupled with trace element analysis of urine, can be applied to accurately predict early lung cancer. A dataset involving the determination of 9 trace elements of 122 urine samples, among which 95 were taken from healthy adults and 27 from patients with lung cancer, was used for illustration. Kennard and Stone (KS) algorithm coupled with an alternate re-sampling was used to partition the whole dataset into equally sized training and test set, which were then reversed to yield a second operating case (named case A and case B, respectively). On the test set, the prediction results based on Adaboost were compared with those from Fisher discriminant analysis (FDA). It was exactly because of the failure of FDA to provide a satisfactory classification that we had to explore new ways, thereby leading to the introduction of Adaboost. The final Adaboost classifiers achieved a sensitivity of 100% for both cases, a specificity of 93.8%, 95.7%, and an overall accuracy of 95.1%, 96.7%, for case A and case B, respectively. In either case, Adaboost always achieves better performance than FDA; also, it is less sensitive to the composition of the training set compared and easy to control over-fitting. It seems that Adaboost is superior to FDA in the present task. These results confirm the benefits of the proposed method, suggesting that integrating Adaboost and trace element analysis of urine can serve as a useful tool of diagnosing lung cancer in clinical practice.

2. Theory and algorithm

2.1. Adaboost

A classification task is actually a classical learning problem, which can be formulated as a search for a good classifier/classification rule, h, using available data $\{x_i, y_i\}$,$i = 1,2, \ldots, n$. Here x is a vector of m predictors, and y, which takes on the value $\{-1, 1\}$, indicates the class of the pattern associated to $\boldsymbol{x}$. A classifier is called a weak classifier if its error rate is slightly better than random guessing and is called a strong classifier if it is very accurate.

In most cases, it may be difficult to achieve a satisfactory accuracy based on a single classifier [23]. In order to improve a weak classifier by stabilizing its decision, a number of techniques could be used, for instance, noise injection [24]. Another approach is to construct many weak classifiers instead of a single one and then combine them in some way into a powerful classifier. Recently a number of such combining techniques have been developed, among which, Adaboost is one of the most popular and effective algorithms, formulated by Freund and Schapire [25]. The purpose of Adaboost is to find a highly accurate classifier by combining many weak classifiers, each of which may be only moderately accurate [26–28]. The main idea of Adaboost algorithm is to define, at each step (for each classifier), a specific probabilistic distribution of learn patterns (the training set), depending on previous results. A weight is assigned to each pattern. It is initialized to $1/N$, and, at each step, the weight of each misclassified patterns is increased (or alternatively, the weight of each correctly classified example is decreased), so that the new classifiers are concentrated on hard patterns. In this way a sequence of training sets and classifiers can be trained. One can therefore obtain the final decision, *i.e.*, an ensemble classifier, by a weighted majority vote.

From the point of computation, Adaboost consists of the following steps:

1. Assign an initial weight to each sample of the original training set.

$$w_i^{(1)} = 1/N, \quad i = 1, 2, \ldots, N$$

2. Do for $t = 1$ to T
   (1) Train a weak classifier $f^{(t)}(\boldsymbol{x})$ based on re-sampling the weighted training set. First, a special training set with N samples is constructed by randomly re-sampling with replacement from the original training set. The chance for a sample to be picked is related to its weight. A sample with a higher weight has a higher probability to be picked. Then, the training set is used to train a weak classifier of "decision stump", as described later.
   (2) Apply the $f^{(t)}(\boldsymbol{x})$ to the original training set. If a sample is misclassified, its error $\mathrm{err}_i^{(t)} = 1$, otherwise its error $\mathrm{err}_i^{(t)} = 0$.
   (3) Compute the sum of the weighted errors on the whole training set: $\mathrm{err}^{(t)} = \sum_{i=1}^{N} w_i^{(t)} \mathrm{err}_i^{(t)}$.
   (4) Calculate the confidence index of the classifier $f^{(t)}(\boldsymbol{x})$, *i.e.*, $\alpha^{(t)} = (1/2)\ln[(1 - \mathrm{err}^{(t)})/(\mathrm{err}^{(t)})]$ on condition that $\mathrm{err}^{(t)} \leq 0.5$; otherwise, go to (1).
   (5) Update the weights $w_i^{(t+1)} = w_i^{(t)} \exp[\alpha^{(t)} \mathrm{err}_i^{(t)}]$, $i = 1, 2, \ldots, N$, in order to maintain a more reasonable distribution over all training samples
   (6) Re-normalize $w_i^{(t+1)}$ so that $\sum_{i=1}^{N} w_i^{(t+1)} = 1$.
   (7) $t = t + 1$. If $t < T$, repeat steps (1–6); otherwise, stop and $T = t - 1$. After T iterations, there will be T weak classifiers $f^{(t)}(x)$, $t = 1, 2, \ldots, T$.
3. Construct the ensemble classifier $f^* = \mathrm{sign}[\sum_{t=1}^{T} (\alpha^{(t)} f^{(t)}(x))]$, through which an unknown sample can be classified.

In Adaboost, the weighted error (*i.e.*, WeightedErr) on each weak classifier is a key index. Freund and Schapire proved that the training error of the ensemble classifier is at most $\Pi[2\sqrt{\mathrm{err}^{(t)}(1 - \mathrm{err}^{(t)})}]$, known as upper error Bound (ErrBound), a useful index [29]. Thus, in this study, we used both WeightedErr and ErrBound for analysis purpose.

2.2. Weak classifier algorithm

In this study, decision stump consisting of a one-level binary decision tree with categorical or numerical class label, is used to train all weak classifiers. Decision stump is defined as follows:

$$f(x; j, b, s) = s \cdot \mathrm{sign}(x_j - b),$$

where s takes on the values $\{-1,1\}$ and b takes on values as defined below. A decision stump is specified by the parameters j, b and s. It is easily seen that for fixed values of s and b, the decision stump is a shifted step function that assigns $\boldsymbol{x}$ a label based on only the jth predictor $\boldsymbol{x}_j$. There exist many candidates decision stumps (*i.e.* combinations of s and b) for each predictor. Given a training set $\{x_i, y_i\}$, $i = 1, 2, \ldots, n$, we prepare a collection of decision stumps for each predictor $\boldsymbol{x}_j$ in the following manner.

Sort all unique values of the jth predictor $\boldsymbol{x}_j$ as $\{x(j)_i\}$, $i = 1, 2, \ldots, n_j$, where n_j is the number of unique values of the jth predictor. Note that $x(j)_i$ is the jth predictor of the ith sample $\boldsymbol{x}_i$.

Find all mid-points between sequential pairs of points in this sorted collection.

For each mid-point (indicated by b), prepare two candidate decision stumps $f(x; j, b, 1)$ and $f(x; j, b, -1)$.

Finally, a total of $\sum_{j=1}^{K} 2(n_j - 1)$ classifiers are prepared in step (3) for K predictors. So, one weak classifier in Adaboost can be obtained.

2.3. Fisher discriminant analysis

Fisher discriminant analysis [30,31] is a linear dimensionality reduction and classification technique, optimal in terms of maximizing the separation between several classes. It determines a set of projection vectors that maximize the scatter between the classes while minimizing the scatter within each class. A short mathematical description follows. Stacking the training data for all classes into a matrix $\boldsymbol{X}$ ($n \times m$) and representing the ith row of $\boldsymbol{X}$ with the column vector $\boldsymbol{x}_i$, the total-scatter matrix is

$$S_t = \sum_{i=1}^{n} (x_i - x_{\text{mean}})(x_i - x_{\text{mean}})^T \quad (1)$$

where x_{mean} is the total mean vector whose elements correspond to the means of the columns of $\boldsymbol{X}$. Define $\boldsymbol{X}_j$ as the set of vectors x_i which belong to the class j, the within-scatter matrix for class j is

$$S_j = \sum_{x_i \in X_j} (x_i - x_{j,\text{mean}})(x_i - x_{j,\text{mean}})^T \quad (2)$$

where $x_{j,\text{mean}}$ is the mean vector for class j. Let c be the number of classes, then

$$S_w = \sum_{i=1}^{c} S_i \quad (3)$$

is the within-class-scatter matrix, and

$$S_b = \sum_{j=1}^{c} n_j (x_{j,\text{mean}} - x_{\text{mean}})(x_{j,\text{mean}} - x_{\text{mean}})^T \quad (4)$$

is the between-class-scatter matrix where n_j is the number of patterns in class j. In fact, the total-scatter matrix is equal to the sum of the between-scatter matrix and the within-scatter matrix, *i.e.* $S_t = S_w + S_b$. FDA attempts to seek an optimal discriminating vector w by maximizing the Fisher criterion:

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \quad (5)$$

It can be shown mathematically that in most cases the vector w is equal to the eigenvector of the generalized eigenvalue problem

$$S_b w = \lambda S_w w \quad (6)$$

where the eigenvalue λ indicates the degree of overall separability. Here, FDA is used as a reference algorithm. By projecting onto direction w, an observation corresponding to a row (*i.e.*, nine concentration values of a person) in Matrix X can be transformed into a scalar from a vector of m-dimension, and then classified only based on a threshold, which is determined to lie between means of training data projected onto direction w.

2.4. Sample set partitioning

It is well-known that, given a dataset, how to select a representative training set for training a prediction model/classifier, is very important. A test set is also needed to evaluate the performance of the model. In the strictest sense, the evaluation is valid only if the test set has the same distribution as the training set, as the samples ascribed to each class often present a certain distribution. In practice, whether in the training set or in the test, samples from the same class should maintain the original distribution as much as possible. Thus, in this study, an especial scheme, *i.e.*, Kennard and Stone [32–34] algorithm coupled with an alternate re-sampling, is used to realize sample set partitioning. That is, the KS algorithm is first used to rank each class of samples to produce a sample sequence. Next, for each sequence, an alternative re-sampling is applied to pick one sample of every two samples into the training set while the remaining samples constitute the test set, which we called case A. By this means, the whole dataset is split into two equal parts with approximately the same distribution. In order to observe the effect of sample partitioning on Adaboost, we also exchange case A's training set and test set to generate case B. In this study, we consider both case A and case B.

The KS algorithm is a well-known representative sample selection algorithm based on maximizing the minimal Euclidean distances between already selected samples and the remaining samples. Its selection rule consists of the following steps:

(1) Select the two most distant samples using the Euclidean distance measure.
(2) Store the shortest Euclidean distances in a distance list with the corresponding sample number for each remaining sample.
(3) Select the sample with the maximum distance from the shortest distances list.

This procedure is repeated until an expected number of samples are selected, *i.e.*, the total number of samples in our case.

3. Experiment

3.1. Sampling and chemical analysis

The dataset used in this study was taken from the work of Ms. Chen D. (Shenyang Pharmaceutical University in China) [35,36]. Here, a brief introduction was provided. It consists of two groups of samples; one is controlled (healthy) group from 95 healthy persons aged 34–81 while the other is patient (cancer) group from 27 lung cancer patients aged 34–81. For each subject, 50 ml of early morning urine samples was collected. Before analysis, all the urine samples were stored in polyethylene bottles at the temperature of −18 °C. The urine samples were digested with a mixture of HNO_3–$HClO_4$–H_2O_2 (10:6:1, v /v) on the heating board of 140 °C after a overnight pre-digestion, and then diluted to the volume of 10 ml volumetric flask with 1% (v/v) HNO_3 and high purity deionized water. Afterwards, the concentrations of Cr, Fe, Mn, Al, Cd, Cu, Zn, and Ni were determined by inductively coupled plasma atomic emission spectrometry (ICP-AES) and instrument parameters, *i.e.*, high frequency power, cooling gas velocity and atomizer hamber gas pressure were set as 1.2 kW, 18 L/min and 45 psi, respectively. The concentration of Se was determined using atomic fluorescence spectrometry (AFS) and the process of sample preparation is slightly different. That is, the digested samples were diluted to volume by 5% (v/v) HCl instead of HNO_3, and then, 2 ml concentrated HCl and 1 ml potassium ferricyanide was added. All the 122 samples are divided into two equal parts (each consisting of 61 samples): a training set and an independent test set; the former was used for constructing a classifier and the latter for validation. It should be pointed out that in the original paper [35], the author focused on investigating the distribution of trace elements and also provided a linear discriminant analysis, but no independent test set was set aside for validation.

3.2. Software and computation

All of the calculations were performed with Matlab version 7.0 under Windows Xp, based on Pentium IV with 256 RAM. Both Adaboost and FDA were performed by the Statistical Pattern Recognition toolbox (http://cmp.felk.cvut.cz/cmp/software/stprtool/index.html).

4. Results and discussion

4.1. Preliminary analysis

It is well-known that correlation, often measured as a correlation coefficient, can indicate the strength of a linear relationship between two random variables. Therefore, the correlation coefficient of each pair of trace elements is first calculated and given in Table 1, suggesting that there exist no obvious linear correlations between elements. For healthy group and cancer group, the descriptive statistics including mean, minimum, maximum, standard deviation and RSM (the ratio of standard deviation to mean) are summarized in Table 2. It can be observed from Table 2 that there are some differences of elemental concentration between two groups and for most of the elements, the concentrations are relatively dispersive. On the average, the concentrations of Fe, Mn and Al for healthy group are higher than those for cancer group, while the concentrations of other elements are higher in the urine of cancer group. To get an overview of data distribution, Fig. 1 gives the frequency histogram and corresponding estimated probability distribution of trace element concentrations for both healthy group and cancer group. Obviously, in most cases, the distributions are not normally distributed but remain considerable overlap. Of these elements, the concentrations of Ni are significantly different between the healthy and cancer groups; it seems that the concentration of nickel can be used as a simple criterion to discriminate the healthy and cancer groups. However, similar to the calibration in analytical chemistry, to use only one element/variable may be dangerous for a diagnosis and fail to achieve an acceptable accuracy. For this reason, more effort is paid to construct a better classifier instead of selecting variables in this study.

Table 1
Correlative coefficient between different pairs of trace elements.

	Cr	Fe	Mn	Al	Cd	Cu	Zn	Ni	Se
Cr	1.00	0.14	0.20	0.07	0.24	0.34	0.41	0.61	0.43
Fe		1.00	0.53	0.66	0.25	0.26	0.12	0.18	0.09
Mn			1.00	0.22	0.06	0.05	0.02	−0.06	−0.21
Al				1.00	0.09	−0.03	0.06	−0.05	−0.13
Cd					1.00	0.59	0.60	0.59	0.25
Cu						1.00	0.49	0.69	0.30
Zn							1.00	0.60	0.56
Ni								1.00	0.50

In order to obtain preliminary indications on the possible clustering of the urine samples in the two groups, principal components analysis (PCA) was used. Using sample set partitioning described before, we generated two kinds of operating cases: case A and case B. In case A, the test set consisted of 47 healthy subjects and 14 lung

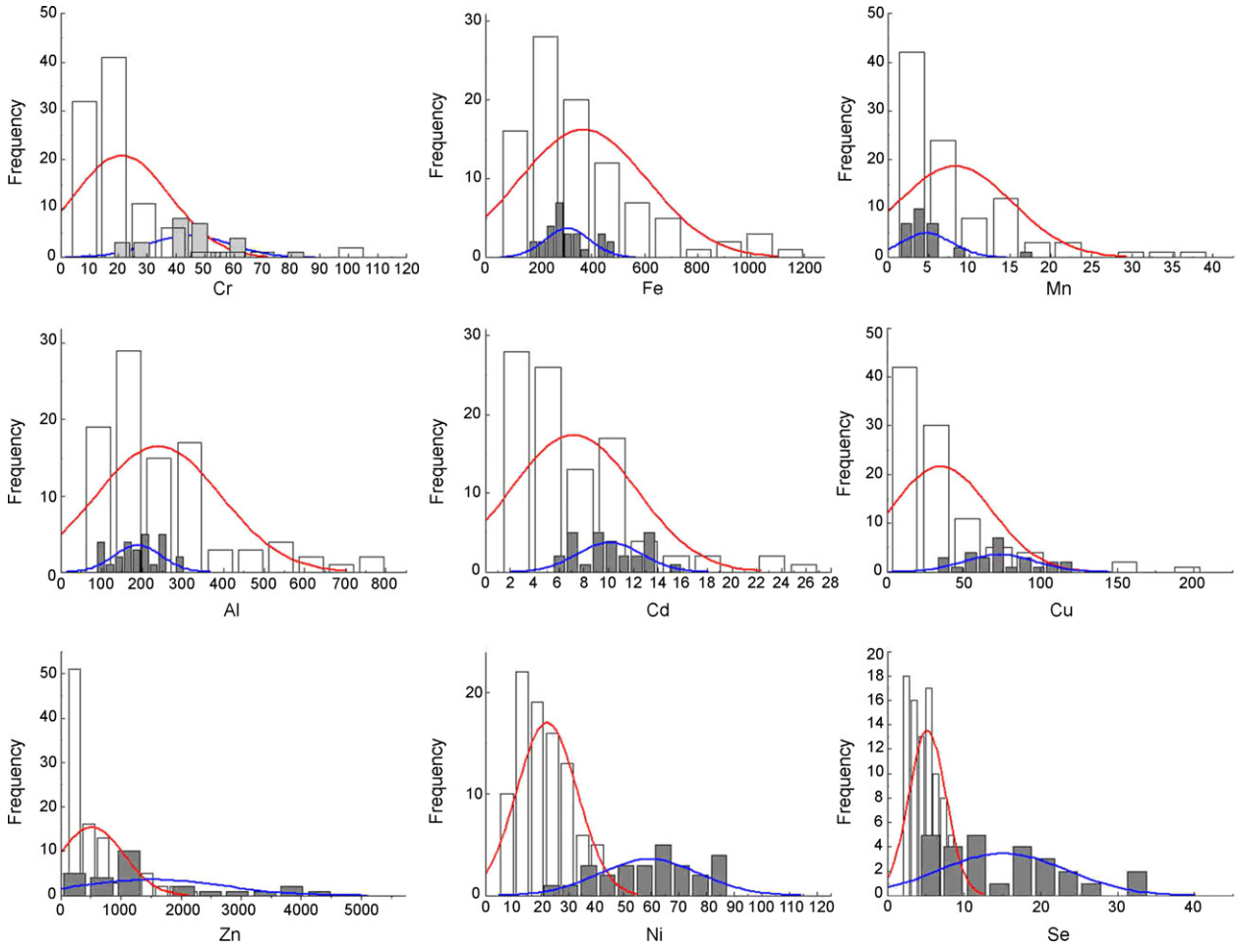

Fig. 1. The frequency histogram and corresponding estimated probability distribution of trace element contents for healthy group (red line) and lung cancer group (blue line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Table 2
Descriptive statistics of trace element concentration in urine (ng/ml).

	Healthy					Lung cancer				
	Minimum	Maximum	Mean	S.D.	RSM	Minimum	Maximum	Mean	S.D.	RSM
Cr	3.1	106	21.2	16.8	0.79	18	84.9	44.6	14.4	0.32
Fe	54.8	1205	368.3	242.1	0.66	164.6	487.0	309.3	85.0	0.27
Mn	1.0	39.5	8.2	7.0	0.85	1.4	17.8	4.7	3.2	0.68
Al	54.5	8.3.3	239.4	154.8	0.64	88.4	303.6	189.1	58.9	0.31
Cd	1.2	27.1	7.1	5.1	0.71	5.5	15.9	10.1	2.7	0.26
Cu	1.2	205.5	34.7	32.1	0.92	32.1	121	73.3	24.0	0.32
Zn	101.3	2524.0	512.3	531.2	1.03	0.3	4526.0	1519.8	1194.8	0.78
Ni	5.1	59.7	22.3	10.9	0.49	20.5	87.9	59.4	18.2	0.31
Se	2.0	11.6	5.2	2.4	0.14	4.1	34.0	15.0	8.3	0.55

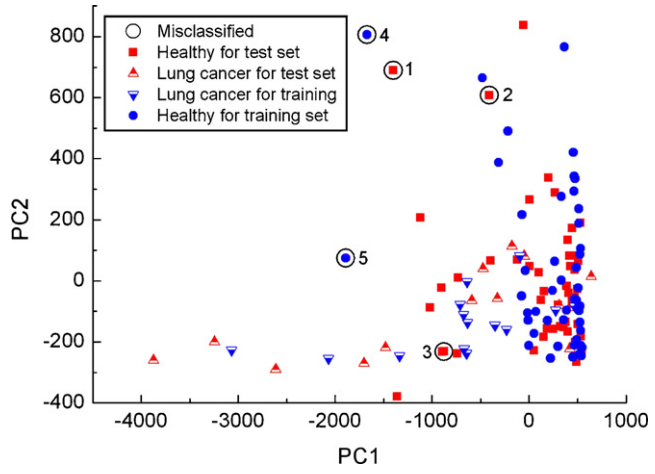

Fig. 2. Score plot of the first two components (PC1 and PC2) for both the training set and the test set (The red (1–3) and blue (4–5) points marked with circle denote the misclassified samples for case A and for case B, respectively). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

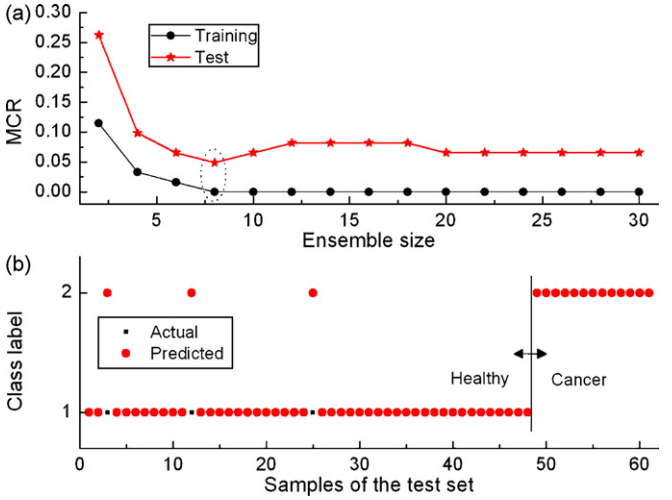

Fig. 3. (a) The curves of MCR (misclassified rate) versus the ensemble size; (b) the predictive performance of the final ensemble classifier on the test set for case A. (Note: Class label "1" signifies health while "2" denotes cancer patient; only three samples of healthy people are misclassified).

cancer subjects while in case B, the test set consisted of 48 healthy subjects and 13 lung cancer subjects. In fact, the training set of case A was exactly the test set of case B while the training set of case B was the test set of case A. That is to say, case B can be obtained only by exchanging case A's training set and test set, vice versa. Taking case A as an example, Fig. 2 shows the score plot of the first two components (PC1 and PC2) for both the training set and the test set. As it can be seen that there exist two clusters corresponding to the two groups; also, the information contained in both the training set and the test set is similar. However, the points associated to cancer group and healthy group spread along PC1 and PC2 directions, respectively, indicating that using PC1 and PC2 is not enough to satisfactorily separate the two kinds of samples. Introducing more PCs has also been tested but failed to improve the results. Taking into account that FDA is a classic classification algorithm, we have also attempted to use it to the task so as to provide a reference. As shown later, FDA has also failed to provide a model with good performance, which is just the reason that makes us move on to Adaboost. These evidences indicate that the classification task is not easy.

4.2. Classification based on Adaboost

Based on the Adaboost strategy using decision stump for training weak classifiers, we built a series of ensemble classifiers with different ensemble size, *i.e.*, the number of weak classifiers. Figs. 3 and 4 show the curves of MCR (misclassified rate) versus the ensemble size and the predictive performance of the final ensemble classifier on the test set for case A and for case B, respectively. It can be noticed that, for each case, with the increase of ensemble size, the

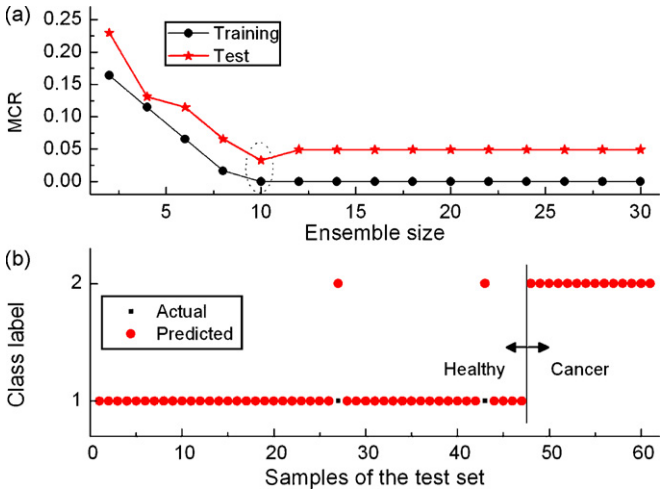

Fig. 4. (a) The curves of MCR (misclassified rate) versus the ensemble size; (b) the predictive performance of the final ensemble classifier on the test set for case B. (Note: Class label "1" signifies health while "2" denotes cancer patient; only three samples of healthy people are misclassified).

MCR values for both the training set and the test set drop quickly. For case A, when integrating only the first eight weak classifiers, the MCR for the training set achieves to zero while the MCR for the test set reaches a minimum (4.9%) and afterward ascend slightly, indicating that the ensemble size should be controlled. For case B, when the MCR for the training set achieve to zero, the ensem-

Table 3
Performance measures of optimal classifiers associated to both Adaboost and FDA.

	Case A (48 + 13)			Case B (47 + 14)		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
FDA	92.3%	89.6%	90.1%	85.7%	89.1%	88.5%
Adaboost	100.0%	93.8%	95.1%	100.0%	95.7%	96.7%

Table 4
The five samples of healthy people misclassified as lung cancer patients.

No.	Cr	Fe	Mn	Al	Cd	Cu	Zn	Ni	Se
1	32.7	107.8	15.8	535.8	13.5	93.8	1994.0	51.5	2.5
2	28.5	967.0	18.3	434.2	10.3	66.2	1010.0	59.7	3.7
3	16.1	228.9	16.7	99.8	18.2	76.8	1530.0	48.9	6.5
4	20.8	161.0	5.5	194.4	23.3	65.5	1394.0	31.1	5.6
5	19.1	411.7	14.9	83.2	13.9	77.3	1666.0	40.1	4.2

Note: No. 1–3 samples corresponds to case A and No.4–5 samples correspond to case B.

ble classifier corresponds to the first 10 weak classifiers and also shows the optimal predictive performance, *i.e.*, the MCR for the test set reaches the minimum (3.3%). It seems that in Adaboost, the optimal ensemble size may be dependant on the training set, but the extent is relatively small due to the use of our especial scheme of sample set partitioning that guarantees both the training set and the test set have a similar information distribution. Obviously and more importantly, in either case, it is convenient to choose an optimal ensemble size, which is exactly the smallest size that makes MCR for the training set equal to zero, as shown in Figs. 3 and 4 Figs. 3(a) and 4(a). These phenomena are consistent to the declaration on the traits of Adaboost. The ensemble size T is the only parameter to be tuned. Generally, as T increases, the training error decreases almost monotonically, even to zero. However, the test error often behaves differently. It may decrease initially but usually increases after a certain number of steps. This phenomenon is called over-fitting. Thus, it is also necessary to control the ensemble size, *i.e.*, the number of weak classifiers. However, in the present task, it seems that over-fitting can easily be avoided by selecting the optimal ensemble size, only according to the MCR for the training set.

The optimal ensemble classifiers for both cases (consisting of the first 8, 10 weak classifiers, respectively) have been evaluated on the corresponding test sets, by means of three measures, *i.e.*, sensitivity, specificity and accuracy. Sensitivity is given as the ratio of TP/(TP + FN), where TP and FN are the number of the true positive (cancer) and false negative results, respectively. Specificity is given by the ratio of TN/(TN + FP), where TN and FP are the number of true negative (healthy) and false positive results. Accuracy is given by the ratio of (TP + TN)/(TP + FP + TN + FN) and equal to 1-MCR. The values of the three measures associated to both Adaboost and FDA are summarized in Table 3. When applying FDA, a few preprocessing methods such as log-transform and auto-scaling were also attempted but failed to obviously reduce the MCR, especially on the test set, and therefore, the results are not reported. It is clear from Table 3 that in either case A or case B, Adaboost always leads to better performance compared to FDA with a single model. Furthermore, Adaboost is less sensitive to the composition of the training set. Such a finding is in accordance with the literature, which claims that an ensemble classifier is often more accurate and robust than a single classifier, even its individual members [37,38]. As shown in Figs. 3(b) and 4(b), of 61 test samples, all cancer patients are correctly classified and only three healthy peoples are misclassified in case A and only two healthy peoples are misclassified in case B (also marked by circles in Fig. 2). Table 4 gives the composition of the five samples, whose concentrations of Cu and Zn seem to be closer to those of cancer group, thus being difficult to classify.

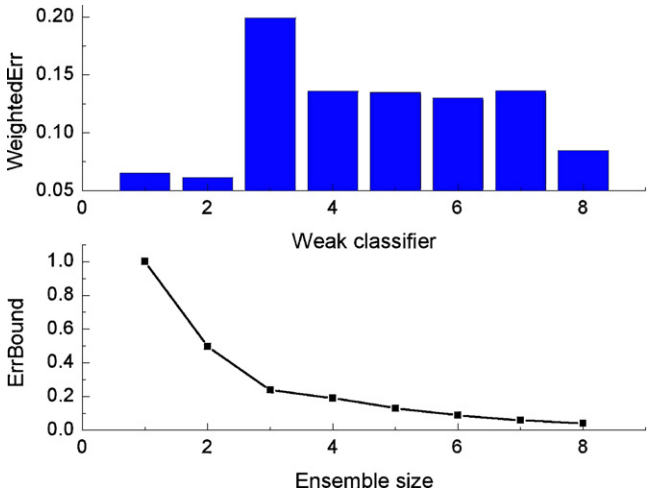

Fig. 5. The bar plot of weighted errors of weak classifiers (upper) and the curve of error bound versus ensemble size (lower) corresponding to the optimal ensemble classifier containing eight weak classifiers for case A.

To obtain a proof on the validity of Adaboost, we have taken case A as an example to check the weighted errors and error bounds of the first eight weak classifiers corresponding to the optimal case, as shown in Fig. 5. Evidently, compared to the first two weak classifiers, the successive ones always take on higher WeightedErr values. This is because they have paid more attention on those "hard" samples to produce the final ensemble classifier. It is just by this means that the superiority of Adaboost can be brought into play. On the other hand, the ErrBound curve can imply the risk of over-fitting to some extent. If the ErrBound values are high, the risk of over-fitting is correspondingly small, vice versa, implying the importance of controlling the ensemble size.

5. Conclusion

It is widely recognized that the primary requirement for successful treatment of lung cancer is early detection. By the time symptoms are present, it is often too late to facilitate a full care. Therefore, there is a concern to develop the methods for early diagnosis. This study demonstrates that the Adaboost using decision stump as the algorithm of weak classifiers, in combination with trace element analysis of urine, could be a potential tool for diagnosing early lung cancer in clinical practice.

Acknowledgements

This work was supported by Scientific Research Startup Fund for Doctor, Yibin University. The authors thank Ms. Chen D. for providing the dataset in this paper.

References

[1] M. Patriarca, A. Menditto, G.D. Felice, F. Petrucci, S. Caroli, M. Merli, C. Valente, Microchem. J. 59 (1998) 194–202.
[2] M.L. Hegde, P. Shanmugavelu, B. Vengamma, T.S.S. Rao, R.B. Menon, R.V. Rao, K.S.J. Rao, J. Trace Elem. Med. Biol. 18 (2004) 163–171.
[3] K. Gurusamy, B.R. Davidson, J. Trace Elem. Med. Biol. 21 (2007) 169–177.
[4] H.L. Zhai, X.G. Chen, Z.D. Hu, Comput. Biol. Chem. 27 (2003) 581–586.
[5] A.M. Ebrahim, M.A.H. Eltayeb, M.K. Ahaat, N.M.A. Mohmed, E.A. Eltayeb, A.Y. Ahmed, Sci. Total Environ. 383 (2007) 52–58.
[6] P. Frisk, P. Darnerud, G. Friman, J. Blomberg, N.G. Ilbäck, J. Trace Elem. Med. Biol. 21 (2007) 29–36.
[7] Z.Y. Zhang, H.L. Zhou, S.D. Liu, P. Harrington, Chemom. Intell. Lab. Syst. 82 (2006) 294–299.
[8] M.T. Douglas, Anal. Bioanal. Chem. 375 (2003) 1062–1066.
[9] F. Bianchi, M. Maffini, A. Mangia, E. Marengo, C. Mucchino, J. Pharm. Biomed. Anal. 43 (2007) 659–665.

[10] H.A. Celik, H.H. Aydin, A. Ozsaran, N. Kilincsoy, Y. Batur, B. Ersoz, J. Clin. Biochem. 35 (2002) 477–481.
[11] Y. Miura, K. Nakai, K. Sera, M. Sato, J. Nucl. Instrum. Methods Phys. Res. B 150 (1999) 218–221.
[12] Y. Miura, K. Nakai, A. Suwabe, K. Sera, J. Nucl. Instrum. Methods Phys. Res. B 189 (2002) 443–449.
[13] G.C. Sturniolo, C. Mestriner, B. Irato, B. Albergoni, G. Longo, R. D'Inca, Am. J. Gastroenterol. 94 (1999) 334–338.
[14] S. Chan, B. Gerson, S. Subramaniam, Clin. Lab. Med. 18 (1998) 673–685.
[15] G. Forte, A. Alimonti, N. Violante, M. Gregorio, O. Senofonte, F. Petrucci, G. Sancesario, B. Bocca, J. Trace Elem. Med. Biol. 19 (2005) 195–201.
[16] Y.L. Ren, Z.Y. Zhang, Y.Q. Ren, W. Li, M.C. Wang, G. Xu, Talanta 44 (1997) 1823–1831.
[17] Z.Y. Zhang, H.L. Zhou, S.D. Liu, P.B. Harrington, Anal. Chim. Acta 436 (2001) 281–291.
[18] O.P. Whelehan, M.E. Earll, E. Johansson, M. Toft, L. Eriksson, Chemom. Intell. Lab. Syst. 84 (2006) 82–87.
[19] R.T. Greenlee, M.B. Hill-Harmon, T. Murray, M. Thun, CA-Cancer J. Clin. 51 (2001) 15–36.
[20] Z.W. Huang, A. Mcwilliams, H. Lui, D. Mclean, S. Lan, H.S. Zeng, Int. J. Cancer 107 (2003) 1047–1052.
[21] M. Zellweger, P. Grosjean, D. Goujon, P. Monnier, H. van den Bergh, G. Wagnieres, J. Biomed. Opt. 6 (2001) 41–51.
[22] L. Nørgaard, G. Sölétormos, N. Harrit, M. Albrechtsen, O. Olsen, D. Nielsen, K. Kampmann, R. Bro, J. Chemom. 21 (2007) 451–458.
[23] A.K. Jain, B. Chandrasekaran, Dimensionality and sample size considerations in pattern recognition practice, in: P.R. Krishnaiah, L.N. Kanal (Eds.), Handbook of Statistics, vol. 2, North-Holland, Amsterdam, 1987, pp. 835–855.
[24] G. An, Neural Comput. 8 (1996) 643–674.
[25] Y. Freund, R.E. Schapire, Proceedings of the Thirteenth International Conference, 1996, pp. 148–156.
[26] R.E. Schapire, Y. Singer, Mach. Learn. 37 (1999) 297–336.
[27] M.H. Zhang, Q.S. Xu, F. Daeyaert, P.J. Lewi, D.L. Massart, Anal. Chim. Acta 544 (2005) 167–176.
[28] C. Tan, M.L. Li, X. Qin, Anal. Bioanal. Chem. 389 (2007) 667–676.
[29] Y. Freund, R.E. Schapire, J. Comput. Syst. Sci. 55 (1997) 119–139.
[30] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., John Wiley & Sons, New York, 2001.
[31] L.H. Chiang, E.L. Russell, R.D. Braatz, Chemom. Intell. Lab. Syst. 50 (2000) 243–252.
[32] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137–148.
[33] M. Daszykowski, B. Walczak, D.L. Massart, Anal. Chim. Acta 468 (2002) 91–103.
[34] R.K.H. Galváo, M.C.U. Araújo, M.N. Martins, G.E. José, M.J.C. Pontes, E.C. Silva, T.C.B. Saldanha, Talanta 67 (2005) 736–740.
[35] D.D. Chen, The methodology on the trace element chemical pattern recognition for early diagnosing lung cancer and cardiovascular diseases (Dissertation), Shenyang Pharmaceutical University in China, 2007 (in Chinese).
[36] D.D. Chen, D. Li, J. Liu, C.J. Zhao, Guangdong Trace Elem. Sci. 14 (2007) 14–17, in Chinese.
[37] D.W. Opitz, R.J. Maclin, Artif. Intell. Res. 11 (1999) 169–198.
[38] R.K.H. Galváo, M.C.U. Araújo, M. do, N. Martins, G.E. José, M.J.C. Pontes, E.C. Silva, T.C.B. Saldanha, Chemom. Intell. Lab. Syst. 81 (2006) 60–67.