Talanta 117 (2013) 549–555

Contents lists available at ScienceDirect

Talanta

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

Selective of informative metabolites using random forests based on model population analysis

Jian-Hua Huang [a,*,1], Jun Yan [a,1], Qing-Hua Wu [a], Miguel Duarte Ferro [a], Lun-Zhao Yi [a], Hong-Mei Lu [a], Qing-Song Xu [b], Yi-Zeng Liang [a]

[a] Research Center of Modernization of Traditional Chinese Medicines, Central South University, Changsha 410083, PR China
[b] School of Mathematical Sciences and Computing Technology, Central South University, Changsha 410083, PR China

ARTICLE INFO

Article history:
Received 18 May 2013
Received in revised form
22 July 2013
Accepted 27 July 2013
Available online 3 October 2013

Keywords:
Random forests (RF)
Model population analysis (MPA)
Informative metabolite
Feature selection

ABSTRACT

One of the main goals of metabolomics studies is to discover informative metabolites or biomarkers, which may be used to diagnose diseases and to find out pathology. Sophisticated feature selection approaches are required to extract the information hidden in such complex ‘omics’ data. In this study, it is proposed a new and robust selective method by combining random forests (RF) with model population analysis (MPA), for selecting informative metabolites from three metabolomic datasets. According to the contribution to the classification accuracy, the metabolites were classified into three kinds: informative, no-informative, and interfering metabolites. Based on the proposed method, some informative metabolites were selected for three datasets; further analyses of these metabolites between healthy and diseased groups were then performed, showing by *T*-test that the *P* values for all these selected metabolites were lower than 0.05. Moreover, the informative metabolites identified by the current method were demonstrated to be correlated with the clinical outcome under investigation. The source codes of MPA-RF in Matlab can be freely downloaded from http://code.google.com/p/my-research-list/downloads/list

1. Introduction

Metabolomics is an important platform of biologic systems that provides holistic metabolic information of the living system to the clinic and pharmaceutical industry. By the quantitative measurement of the metabolites and their dynamic changes in biological samples, metabolomics has been widely used in many areas, such as drugs discovery and both disease diagnostics and treatment [1–4]. Such studies are of great use for the early diagnosis of diseases and preclinical screening of candidate drugs in the pharmaceutical industry [5,6]. In order to carry out such studies, analytical approaches such as HPLC-MS [7], UPLC-MS [8], NMR [9–12], and GC–MS [13–16] have been widely applied for measuring the global metabolome. With the rapid development of modern analytical instruments, experimental data containing larger amounts of information can be generated, which will bring larger chances for scientists to know more about the analyzed systems. However, these large data sets contain not only useful information, but also redundant, uninformative, and even noisy information.

Therefore, the selection of the relevant features is of great importance in metabolomics research. On one hand, feature selection can improve the performance of the model, during the set up of the classification model, by eliminating redundant information. On the other hand, feature selection may also be used to gain further understanding of the data, helping with the identification of the metabolic biomarkers. Over the last few years, there have been a rapid growing of the number of feature selection methods for biomarkers discover. Although some feature selection methods have been proposed in previous studies, such as sub-window permutation analysis (SPA) [17], support vector machine-recursive feature elimination(SVMRFE) [18] and competitive adaptive reweighted sampling(CARS) [19], there is still an increasing demand for new powerful methods to deal with such challenging task. Furthermore, recent studies in areas where high dimensional sets of data are generated, such as Bioinformatics and Genomics, highlighted the risk of over-fitting, posed by variable selection methods i.e., “selection bias problem” [20]. It is well known that when the number of samples *n* is small, splitting the sample into large training and test set it is usually not feasible. Cross-validation method is one solution to deal with such problem. A typical example of these procedures is: estimate the prediction error of

Abbreviations: RF, Random forests; MPA, Model population analysis

[*] Correspondence to: Department of Chemistry and Chemical Engineering, Central South University, Changsha 410083, P.R. China. Tel./fax: +86731 88830831.
E-mail addresses: huangjh85@gmail.com (J.-H. Huang), yizeng_liang@263.net (Y.-Z. Liang).

[1] The first two authors have equal contribution to this article.

0039-9140/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2013.07.070

the model developed with a leave-one-out or k-fold validation using the full dataset n, which leads to the biased estimation of the discrimination error [21]. So, it is important to ensure that the data used to test the classifier is not part of the data used to train. Besides, another problem in feature selection is the instability of the selection process, which usually outputs diverse selections from different runs for the same dataset.

To cope with these problems, a new feature selection method, named "MPA-RF", was proposed, by coupling model population analysis (MPA) with random forests (RF) for the stable selection of important features. Random forests (RF), which was introduced by Breiman [22], have been successfully applied in various biological problems [23,24]. RF is an ensemble method that uses recursive partitioning to generate many trees, then aggregating the results. In RF, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. During the run, each tree is independently constructed using a bootstrap sample of the training data. For each tree, two-thirds of the training samples are used for tree construction and the remaining one-third of the samples are used to test the tree. This left out data, which is called "out of bag" (OOB) data, is used to calibrate the performance of each tree. The most machine learning methods need to resort to cross-validation for the estimation of a classification error, random forest can natively estimate an OOB error in the process of constructing the forest, and this estimate is claimed to be unbiased in many tests [25,26]. MPA was proposed as a general framework for developing data analysis methods [27]. The MPA-based methods could provide some comprehensive insights to the data since it allows the analysis of some interesting outputs of a large number of sub-models. The selected features can be obtained by taking the mean value of all the sub-models. From the view of our experimental results, selecting features from many sub-models makes the process more robust and stable.

2. Materials and methods

2.1. Datasets

In this study, three metabolomics datasets were used to validate the proposed method:

Dataset 1: "T2DM" dataset, which consists on a matrix X of size 90×21, containing the free fatty acids profile of 90 individuals' plasma samples, collected from 45 type 2 diabetes mellitus (T2DM) patients and 45 healthy controls; also a y classification vector is considered, which is equal to -1 or $+1$, corresponding to T2DM patients and healthy controls, respectively. The plasma samples were obtained from the Xiangya Hospital of Hunan in Changsha, China, and profiled using a gas chromatography–mass spectrometry (GC–MS) [28].

Dataset 2: "POCD" dataset, which consists on a matrix X of size 24×44, containing the metabolic profiles of 24 rats, where 12 were collected with the presence of postoperative cognitive dysfunction (POCD) after isoflurane anesthesia, and 12 with the absence of POCD after isoflurane anesthesia; also a y classification vector for the presence or absence of POCD in rats, respectively, is considered. The rats were purchased from Hunan Agricultural University in Changsha, China, and their serum was profiled by using a GC–MS [29].

Dataset 3: "CHOB" consists on a matrix X of size 29×30, containing the metabolic profiles of 29 children, collected from 13 overweight children and 16 healthy controls; also a y classification vector for overweight or healthy children, respectively, is considered. The children′s plasma samples were obtained from the Xiangya Hospital of Central South University in Changsha, China, and profiled using a GC–MS [30].

2.2. Random forest

Random forest (RF) is a classifier consisting of an ensemble of tree-structured classifiers [22]. RF takes advantages of two powerful machine learning techniques: bagging and random feature selection. In bagging, each tree is trained on a bootstrap sample of the training data, and prediction results are made by the majority vote of the trees which obtained during the training process. RF is a further development of bagging, which instead of using all features in dataset; it randomly selects a subset of features to split at each node when growing a tree. In order to assess the prediction performance of the random forest algorithm, RF performs a type of a cross-validation in parallel with the training step by using the so called OOB samples. Specifically, in the process of training, each tree is grown using a particular bootstrap sample. Since bootstrapping is a sampling method with replacement from the training data, part of the data will be 'left out' of the sample, while other part will be repeated in the sample. The 'left out' data constitute the OOB sample. On average, each tree is grown using about 2/3 of the training data, leaving about 1/3 samples as OOB. Since OOB data have not been used in the tree construction, it can be used to estimate the prediction performance. The RF algorithm implemented in the R-package randomForest was used in this study [31]. The algorithm (for both classification and regression) can be stated as follows:

1. Draw n_{tree} bootstrap samples from the original data, n_{tree} is the number of ensemble trees;
2. For each bootstrap sample, grow an un-pruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all variables, randomly select m_{try} variables and choose the best split among those variables (bagging can be thought as the special case of random forests when $m_{try}=p$, the number of variables). In general, m_{try} is simply a number (positive integer) between 1 and p [22].
3. Predict new data by aggregating the predictions of the n_{tree} (i.e., majority votes for classification, average for regression).
   Variable importance: RF, as an ensemble of trees, inherits the ability to select 'important' features. A measure of how each feature contributes to the prediction performance of RF can be calculated in the course of the training. The important scores can be used to identify biomarkers or as a filter to remove non-informative variables. The frequently used type of RF to measure feature importance is the mean decrease in classification based on permutation. For each tree, the classification accuracy of the OOB samples is determined both with and without random permutation of the values to each variable, one by one. The prediction accuracy of after permutation is subtracted from the prediction accuracy before permutation and averaged over all trees in the forest to give the permutation importance value. In the current research, the mean decrease in classification accuracy was accepted to measure variable importance. The importance of each variable can be calculated as Eq.1

$$Importance\ of\ j = \mathrm{Accuracy}_{j\ normal} - Accuracy_{j\ permuted} \quad (1)$$

2.3. MPA method

The aim of MPA is to extract interesting information from a "population" of sub-models, which are built on different sub-datasets sampled from the original dataset using Monte Carlo sampling (MCS) technique. The details of MPA-RF procedure are described in the following steps:

(i) construct thousands of sub-models by MCS, i.e., suppose that we are given a sample matrix X of size m × p, where each row denotes a sample and each column a variable. The corresponding class label is recorded in the vector y of size m × 1 with element equal to 1 or −1 for the binary classification case. To begin with, two parameters related to the sub-dataset sampling need to be determined (1) *N*, the number of MCS; in this case it was 1000 times; (2) *R*, the ratio of samples to be selected for each MCS; it was set up to 0.8. Each time that MCS selected the 0.8 ratio of samples from the original dataset to establish a sub-model, which would be used to set up models by RF, the procedure would be repeated for 1000 times.

(ii) After setting up the sub-dataset, each classification model would be established by using RF. One parameter related to the classification model needs to be defined (1) n_{tree}, the number of trees grown in each forest, where in this case it was 500; i.e., for each sub-model, 500 trees would be grown for classification.

(iii) Calculate the feature importance for each sub-model; calculate the feature importance of each variable for all the models, then obtain the mean value for the final feature importance; finally, rank the feature importance for each feature.

2.4. Performance assessment

In this study, The parameters employed to evaluate the behavior in this investigation are some commonly used ones in classification problems [32], which are given as follows:

$$\text{Specificity} = \frac{TN}{\text{TN}+\text{FP}} \tag{2}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{3}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{4}$$

where *TP* is the number of positive samples predicted correctly (i.e., true positives), *TN* is the number of negative samples predicted correctly (i.e., true negatives), *FP* is the number of negative samples predicted as positive (i.e., false positives), and *FN* is the positive samples predicted as negative (i.e., false negatives). Accuracy is the classification accuracy of the classifier model for both positive and negative data classes. In addition, after getting the values of sensitivity and specificity, the receiver operating characteristic (ROC) curves plot of sensitivity vs 1 - specificity can be constructed [33]. Then, The area under the ROC curve (AUC), which is denoted AUC, is used as an additional performance estimate. A model with no predictive ability would yield the diagonal line. The closer AUC is to 1, the greater is the predictive ability of the model.

3. Results and discussion

3.1. Selection of the number of trees

The first step in RF modeling is to select the number of trees in the train process. In the current study, the models were tested with 2000 trees. The OOB classification error was plotted against the number of trees grown (Fig. 1). As it can be seen from Fig. 1, the OOB errors for three datasets did not decrease with the number of grown tree, or it could said that the model does not over-fit as the error reaches a certain value, no matter how many trees are built. The optimal number of trees was considered to be the one where a relatively stable trend of the lowest OOB error was reached. When the number of grown trees equaled 500, three of the datasets almost got the lowest errors, therefore, 500 was chosen as the parameter for all models.

3.2. Calculate the feature importance by RF

The variable importance calculated by RF for three datasets is listed in Fig. 2. Since the model accuracy showed a large decrease after the permutation of a variable or feature, it may be thought that this variable is very important as an informative variable. It could be found that some features have large contributions to the accuracy, others have negative contributions to the accuracy while some have no contributions at all, as the accuracy of the model has not changed after permutation.

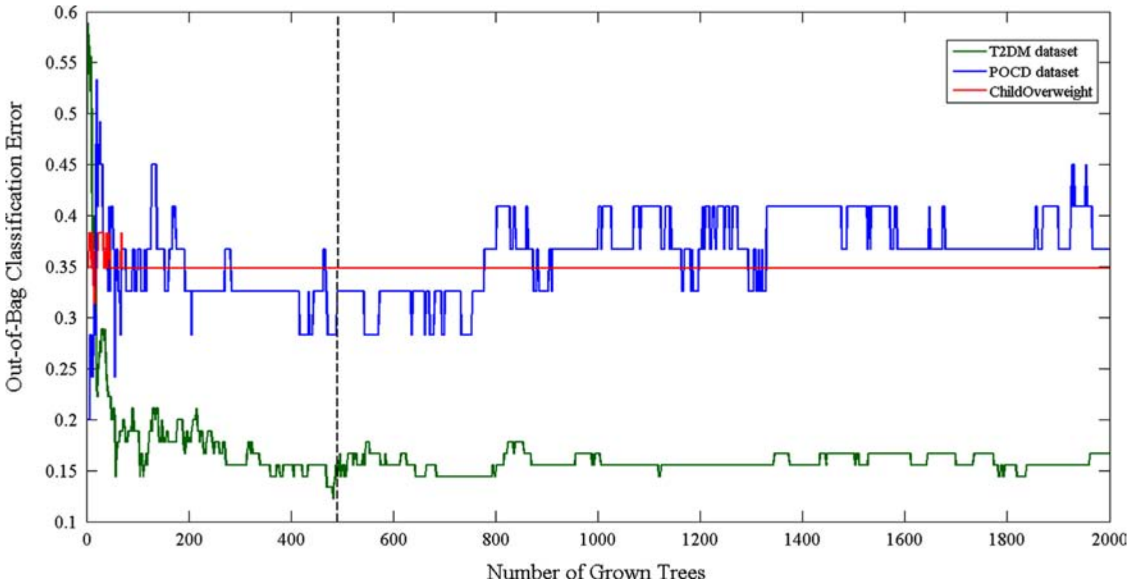

Fig. 1. Identification of the tree number grown in each sub-model.

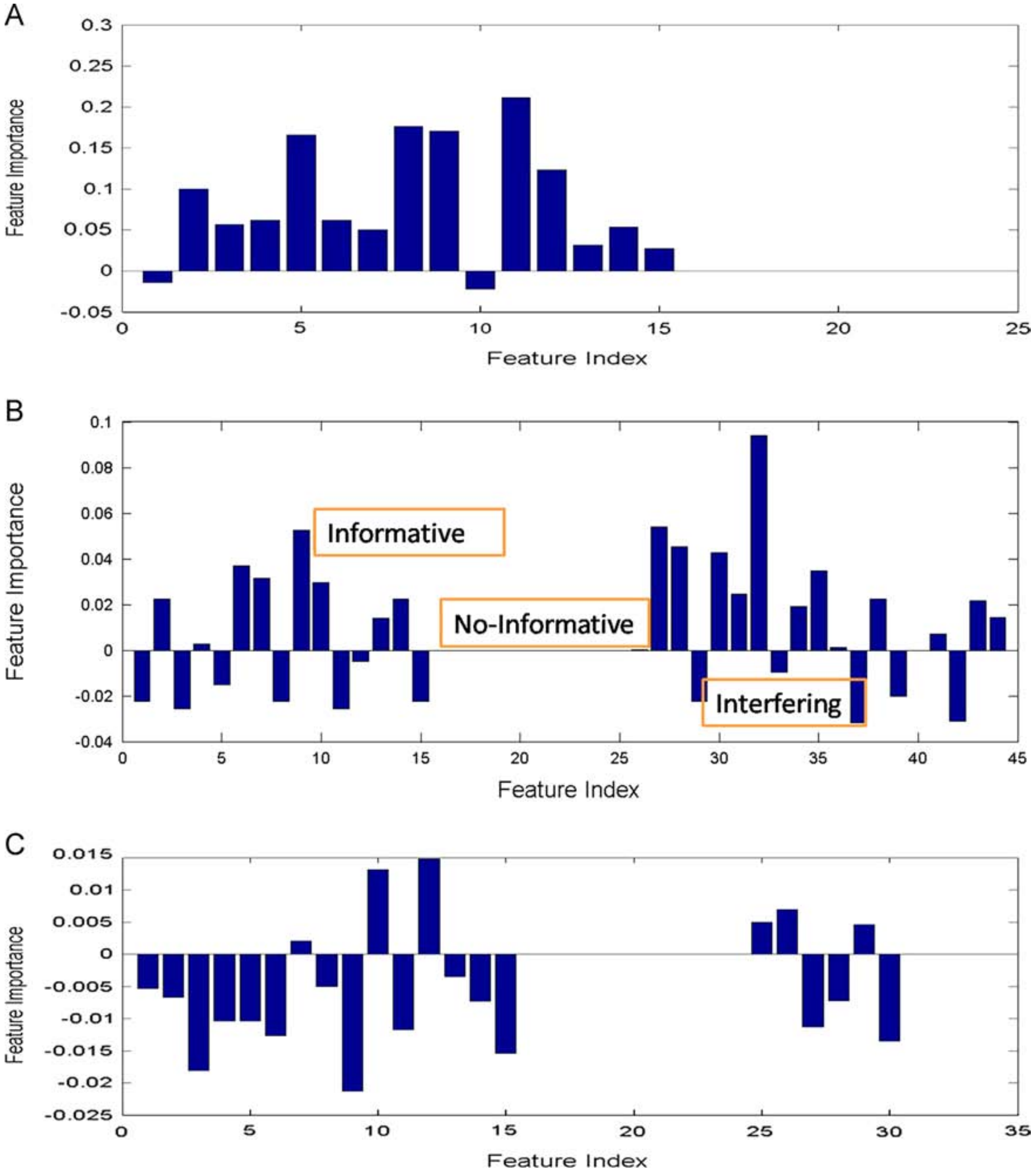

Fig. 2. The variable importance calculated by RF. (A) T2DM dataset, (B) POCD dataset, and (C) CHOB dataset.

By calculating the accuracy decrease after disturbing or permuting one variable, it can be established a reference to estimate the importance of the variable. Based on this reference, the variables may be divided into three detailed classes: informative, no-information and interfering variables. This kind of classification is suitable in metabolomics, since some metabolites in samples, such as urine and serum from different groups (healthy vs cancer patient), always contain these kinds of classification. The informative variables are the metabolites which the difference between healthy and unhealthy samples would be caused by disease, and these metabolites could be taken as potential biomarkers. The no-information variables mean that the concentrations of metabolites between healthy and diseased groups are not changed, since these metabolites have no connection with the diseases. The interfering variables are the metabolites that might be caused by individual differences, the concentrations of these metabolites are not clearly distributed among the groups.

3.3. Calculate variable importance by MPA-RF

In this section, the experiment was designed as two parts. First, it was implemented the variable selection by using the RF program only. All samples in the dataset were used to run the RF program. Fig. 3 shows three results for T2DM dataset by running the RF program three times with the same parameters. As it can be seen from Fig. 3, the variable importance for each variable or metabolite changed with time. Taking the variable 10 as example, it was defined as interference, no-informative, and informative variable, respectively, in three different times of the run. This might cause misjudgment if one does not have an accurate and specific definition of the variable or metabolite, especially in metabolomics research. Besides, the variations of the important variables, such as variable 8, 9, 11, were relatively smaller than other metabolites, these would give us a chance to improve the results by combining with MPA method. Similar results could be observed from other

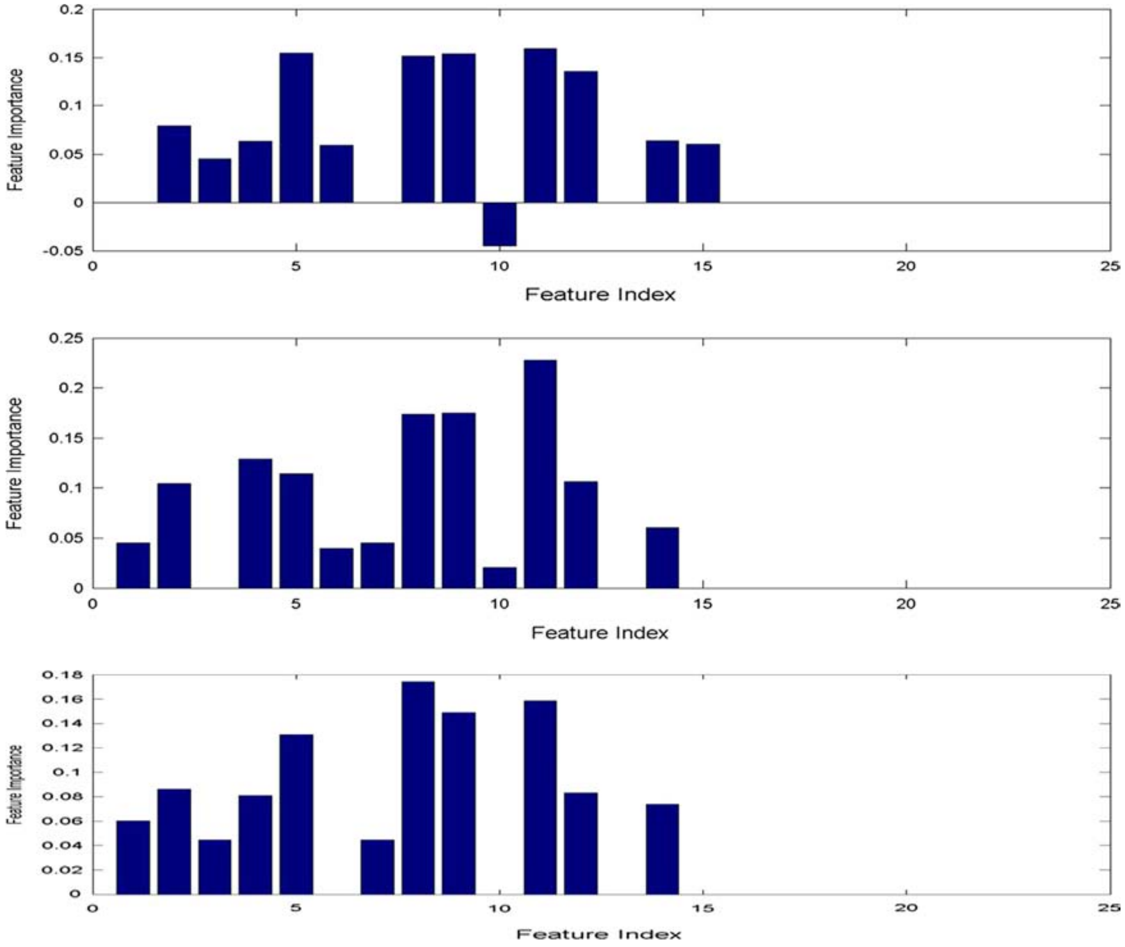

Fig. 3. Three results for T2DM dataset by running the RF program three times with the same parameters.

datasets. Establishing the RF models, the trees grown in the forests are different from each other, which generated different results in different runs. Then, by overcoming these disadvantages caused by the native RF, a chance for RF to have a wide application in metabolomics may be provided.

To overcome the instability of variable selection caused by RF, the MPA method was adopted in the current study. The main idea of MPA is to set up thousand of sub-models for variable selection, then using statistics to evaluate the variable importance for each variable by calculating the mean value among all the models. Selecting features from thousands of sub-models makes the result more robust against small changes in the whole process. The results got by MPA-RF method for three datasets are listed in Fig. 4. There are two advantages by combining RF with MPA. First, one way to decrease the generalization error of RF is to maximize the tree diversity in the forest, i.e., to make sure that the classification trees grown in the forest are dissimilar and uncorrelated from each other; MPA could add the randomness of trees, as each time different samples are used to grow trees in the forest. Second, the variable importance is not stable since a variable will be assigned with a different importance value for each run, then the essence of MPA is to obtain statistically stable results via obtaining the mean value of all the models.

3.4. Compared with other feature selection methods

In this section, it was aimed to compare the performance of the current feature selective method with other feature selection methods. Two reference feature selection methods, SPA[17] and CARS [19], were also used to select important features for the three metabolomics datasets. Table 1 shows the selected features for each dataset and the results obtained with PLS method by using corresponding features. For the MPA-RF, all the variables that have positive contribution to the accuracy i.e., the values with variable importance are larger than zero, have been keep for classification. In order to have a fair comparison with these methods, a PLS model was built on the features selected by each method. The best PLS factors for each model were determined by 10-fold cross-validation. The classification results are also listed in Table 1, including accuracy, sensitivity, specificity, and AUC values for each method. As it can be seen, all the models established by the selected variables were better than those models that were set up without feature selection. For the T2DM dataset, the SPA method got the best results, the accuracy, sensitivity, specificity, and AUC were 100%, 100%, 100%, and 0.9778, respectively. The MPA-RF method also got comparative results, where the accuracy, sensitivity, specificity, and AUC were 98.89%, 100%, 97.78%, and 0.9743, respectively. For the POCD dataset, the MPA-RF got the best result, with accuracy, sensitivity, specificity, and AUC of 91.67%, 91.67%, 91.67%, and 0.8750, respectively. In the CHOB dataset, the accuracy, sensitivity, specificity, and AUC got by MPA-RF method were 80.41%, 84.62%, 75.00%, and 0.8077, respectively. These results proved that the variables selected by MPA-RF had good discrimination ability.

According to the above analysis, some important metabolites were selected. In order to further validate the reliability of the selected features, a *T* test was applied to check the significance of each selected metabolite between two groups, healthy control and

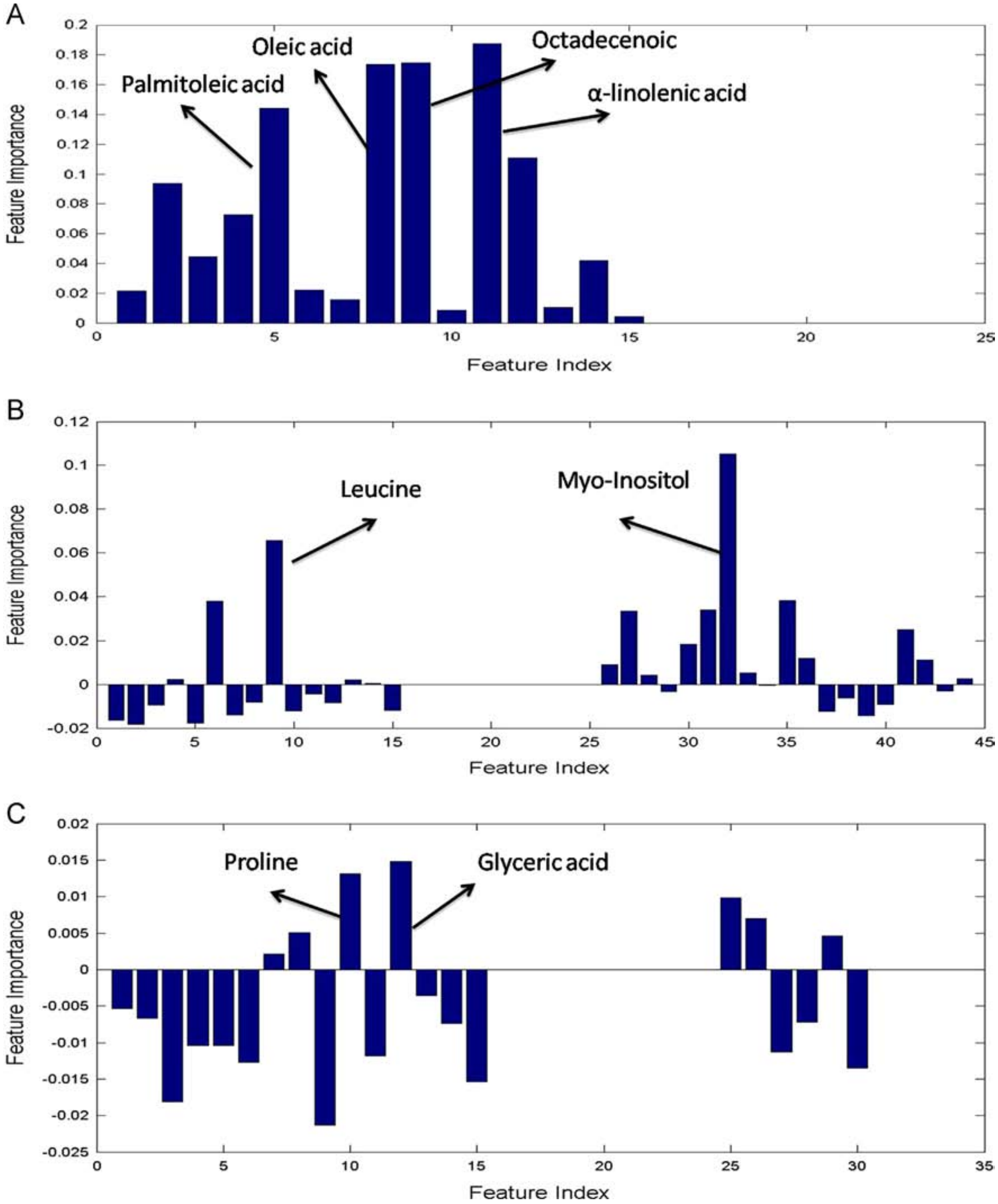

Fig. 4. The variable importance obtained by MPA-RF for three datasets. (A) T2DM dataset, (B) POCD dataset; (C) CHOB dataset.

disease/illness group. For the T2DM dataset, four fatty acid were finally selected as potential biomarkers, which were: Palmitoleic acid (C16:1_{n-9}), Oleic acid (OLA C18:1_{n-9}), Octadecenoic (C18:1_{n-7}), α-linolenic acid (ALA C18:3_{n-3}), and their P values were 1.5899e^{-8}, 1.0605e^{-8}, 2.4558e^{-8}, and 3.8867e^{-10}, respectively. These metabolites were also consistent with the results obtained from biochemical research. Matsuzaka et al., reported that a long-chain fatty acid elongase, elongation of long-chain fatty acids family member-6 (Elovl6), which catalyzes the conversion of palmitate to stearate, plays the major role in insulin resistance [34]. Their experiments demonstrated that the conversion of plaitate (C16) to stearate (C18) was a key step for the emergence of insulin resistance. OLA (C18:1n-9), formed from palmitate by the fatty acid elongase, Elov16, and stearoyl Co-A desaturase (SCD)-1, is the final product of mammalian fatty acid synthesis. Madigan et al., reported that a change from linoleic acid to OLA diet resulted in an improvement of low density lipoprotein and high density lipoprotein in diabetic patients [35]. ALA serves as the substrate to form EPA and is thereby indirectly related to the metabolism of T2DM. For the POCD dataset, two metabolites were selected as potential biomarkers, Leucine and Myo-Inositol, and their P values were 0.0057 and 0.0143, respectively. The result suggested that high levels of myo-Inositol in serum of POCD rats was probably related to unregulated myo-Inositol monophosphate, reflecting possible pathological mechanism concerned with phosphoinositide metabolism. These results were consistent with our group′s previous work [29], and more detailed analysis for the metabolism pathways about the POCD can be found in there. For the CHOB dataset, two metabolites, Proline, and Glyceric acid, were

Table 1
Summary of variable selection results from three datasets using three different methods.

Dataset	Variable	Prediction assessment parameters (%)				Variable choose
		Accuracy	sensitivity	specificity	AUC	
T2DM	None	96.67	95.56	97.78	97.46	All (21)
	CARS	97.78	97.78	97.78	97.51	5, 9, 11, 14, 16, 18, 20
	SPA	100.00	100.00	100.00	97.78	2, 4, 5, 6, 7, 8, 9, 10, 11, 14, 15, 17, 18, 20, 21
	MPA-RF	98.89	100.00	97.78	97.43	1, 2, 3, 4, 5, 6, 8, 9, 10,11, 12, 13, 14, 15
POCD	None	68.97	61.54	75.00	66.35	All (44)
	CARS	87.50	91.67	83.33	84.38	21, 22, 29, 33, 35
	SPA	66.67	66.67	66.67	74.31	8,11,21
	MPA-RF	91.67	91.67	91.67	87.50	6, 9, 26, 27, 28, 30, 31, 32, 35, 36, 41, 42
CHOB	None	68.97	61.54	68.75	69.47	All (30)
	CARS	82.76	86.31	75.00	84.38	1, 2, 4, 5, 7, 9, 10, 14, 15, 19, 23, 24, 26, 27, 28, 30
	SPA	79,31	76.92	81.25	79.09	5, 23, 26
	MPA-RF	80.41	84.62	75.00	80.77	7, 8, 10, 12, 25, 26

selected as potential biomarkers, and their P values were 0.00913, and 0.0221, respectively. Glyceric acid can be derived from oxidation of glycerol and some phosphate derivatives of glyceric acid, such as 2-phosphorlyceric acid and 3-phosphoglyceric acid, are important biochemical intermediates of lipid metabolism. These intermediates have been reported to be correlated with overweight [36,37].

4. Conclusion

In this article, it was presented a new feature selection method, MPA-RF, for the identification of informative biomarkers in complex metabolic datasets. The presented method provided the best use of both RF and MPA method. First, by using selective biomarkers from many sub-models made the result more robust against the changes in the tree growing process; second, avoiding the “selective bias” issue by the useof “OOB” error to evaluate the identified metabolic markers in feature selection process. Finally, comparing with other variable selective methods based on three metabolomics datasets, the current method got quite competitive result; moreover, the informative metabolites identified by the presented method demonstrated to be correlated with the clinical outcome under investigation. Also, wide applications of the proposed feature selection method can be foreseen.

Acknowledgments

This work has been financially supported by the “National Novel Drug Development Foundation 2009ZX09303-008” and Experimental Zooprophylactic Institute of Chinese Academy of Medical Sciences. National Nature Foundation Committee of P.R. China (Grants no. 20875104, No. 21105129, No. 11271374, No. 21175157). The studies meet with the approval of the university′ s review board.

References

[1] J.W. Lee, D. Figeys, J. Vasilescu, Adv. Cancer Res. (2007) 269–298.
[2] J. Silberring, P. Ciborowski, TrAC, Trends Anal. Chem. 29 (2010) 128–140.
[3] D.S.W. Tan, G.V. Thomas, M.D. Garrett, U. Banerji, J.S. de Bono, S.B. Kaye, P. Workman, Cancer J. 15 (2009) 406–420.
[4] J.K. Nicholson, J.C. Lindon, E. Holmes, Xenobiotica 29 (1999) 1181–1189.
[5] S.J. Bruce, I. Tavazzi, V. Parisod, S. Rezzi, S. Kochhar, P.A. Guy, Anal. Chem. 81 (2009) 3285–3296.
[6] M. Oldiges, S. Luetz, S. Pflug, K. Schroer, N. Stein, C. Wiendahl, Appl. Microbiol. Biotechnol. 76 (2007) 495–511.
[7] C. Denkert, J. Budczies, T. Kind, W. Weichert, P. Tablack, J. Sehouli, S. Niesporek, D. Koensgen, M. Dietel, O. Fiehn, Cancer Res. 66 (2006) 10795–10804.
[8] R.S. Plumb, K.A. Johnson, P. Rainville, J.P. Shockcor, R. Williams, J.H. Granger, I. D. Wilson, Rapid Commun. Mass Spectrom. 20 (2006) 2800–2806.
[9] J.C. Lindon, J.K. Nicholson, J.R. Everett, Annu. Rep. NMR Spectrosc. 38 (38) (1999) 1–88.
[10] M.E. Bollard, E.G. Stanley, J.C. Lindon, J.K. Nicholson, E. Holmes, NMR Biomed. 18 (2005) 143–162.
[11] S. Kochhar, D.M. Jacobs, Z. Ramadan, F. Berruex, A. Fuerhoz, L.B. Fay, Anal. Biochem. 352 (2006) 274–281.
[12] E.G. Stanley, N.J.C. Bailey, M.E. Bollard, J.N. Haselden, C.J. Waterfield, E. Holmes, J.K. Nicholson, Anal. Biochem. 343 (2005) 195–202.
[13] J.H. Granger, R. Williams, E.M. Lenz, R.S. Plumb, C.L. Stumpf, I.D. Wilson, Rapid Commun. Mass Spectrom. 21 (2007) 2039–2045.
[14] Q. Zhang, G.J. Wang, Y. Du, L.L. Zhu, A. Jiye, J. Chromatogr. B 854 (2007) 20–25.
[15] K.K. Pasikanti, P.C. Ho, E.C.Y. Chan, J. Chromatogr. B 871 (2008) 202–211.
[16] H.J. Major, R. Williams, A.J. Wilson, I.D. Wilson, Rapid Commun. Mass Spectrom. 20 (2006) 3295–3302.
[17] H.-D. Li, M.-M. Zeng, B.-B. Tan, Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, Metabolomics 6 (2010) 353–361.
[18] X. Lin, F. Yang, L. Zhou, P. Yin, H. Kong, W. Xing, X. Lu, L. Jia, Q. Wang, G. Xu, J. Chromatogr. B (2012).
[19] H. Li, Y. Liang, Q. Xu, D. Cao, Anal. Chim. Acta 648 (2009) 77–84.
[20] C. Ambroise, G.J. McLachlan, PNAS 99 (2002) 6562–6566.
[21] S. Varma, R. Simon, BMC Bioinformatics 7 (2006).
[22] L. Breiman, Mach. Learn. 45 (2001) 5–32.
[23] T. Chen, Y. Cao, Y. Zhang, J. Liu, Y. Bao, C. Wang, W. Jia, A. Zhao, Evidence-Based Compl. Alt. (2013). (doi: 10.1155.2013.298183).
[24] X. Lin, Q. Wang, P. Yin, L. Tang, Y. Tan, H. Li, K. Yan, G. Xu, Metabolomics 7 (2011) 549–558.
[25] G. Martinez-Munoz, A. Suarez, Pattern Recognition 43 (2010) 143–152.
[26] G.-Y. Zhang, C.-X. Zhang, J.-S. Zhang, Commun. Stat. Simula. Comput. 39 (2010) 1877–1892.
[27] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, TrAC, Trends Anal. Chem. 38 (2012) 154–162.
[28] B. Tan, Y. Liang, L. Yi, H. Li, Z. Zhou, X. Ji, J. Deng, Metabolomics 6 (2010) 219–228.
[29] W. Zhang, L. Zhang, H. Li, Y. Liang, R. Hu, N. Liang, W. Fan, D. Cao, L. Yi, J. Xia, Chromatographia 75 (2012) 799–808.
[30] M. Zeng, Y. Liang, H. Li, M. Wang, B. Wang, X. Chen, N. Zhou, D. Cao, J. Wu, J. Pharm. Biomed. Anal. 52 (2010) 265–272.
[31] A. Liaw, M. Wiener, R. News. 2 (2002) 18–22.
[32] B.W. Matthews, Biochim. Biophys. Acta 405 (1975) 442–451.
[33] T. Fawcett, Pattern Recognition Lett. 27 (2006) 861–874.
[34] T. Matsuzaka, H. Shimano, N. Yahagi, T. Kato, A. Atsumi, T. Yamamoto, N. Inoue, M. Ishikawa, S. Okada, N. Ishigaki, H. Iwasaki, Y. Iwasaki, T. Karasawa, S. Kumadaki, T. Matsui, M. Sekiya, K. Ohashi, A.H. Hasty, Y. Nakagawa, A. Takahashi, H. Suzuki, S. Yatoh, H. Sone, H. Toyoshima, J.-i. Osuga, N. Yamada, Nat. Med. 13 (2007) 1193–1202.
[35] C. Madigan, M. Ryan, D. Owens, P. Collins, G.H. Tomkin, Irish J. Med. l Sci. 174 (2005) 8–20.
[36] M.W. Hulver, J.R. Berggren, R.N. Cortright, R.W. Dudek, R.P. Thompson, W. J. Pories, K.G. MacDonald, G.W. Cline, G.I. Shulman, G.L. Dohm, J.A. Houmard, Am. J. Physiol–Endoc M. 284 (2003) E741–E747.
[37] J. He, S. Watkins, D.E. Kelley, Diabetes 50 (2001) 817–823.