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