Analytica Chimica Acta 911 (2016) 27–34

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Variable importance analysis based on rank aggregation with applications in metabolomics for biomarker discovery☆

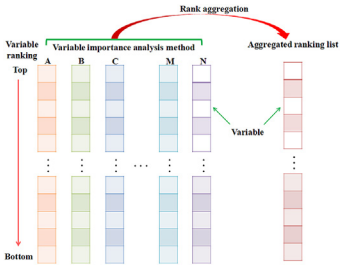
Yong-Huan Yun [a], Bai-Chuan Deng [b], Dong-Sheng Cao [c], Wei-Ting Wang [a], Yi-Zeng Liang [a, *]

[a] College of Chemistry and Chemical Engineering, Central South University, Changsha, 410083, PR China
[b] College of Animal Science, South China Agricultural University, Guangzhou, 510642, PR China
[c] College of Pharmaceutical Sciences, Central South University, Changsha, 410083, PR China

HIGHLIGHTS

- Present a problem of inconsistency between variable ranking methods for biomarker discovery in metabolomics study.
- Rank aggregation is used to merge individual ranking lists into a single "super"-list reflective of the overall preference.
- Rank aggregation has better performance when compared with using all variables and penalized method.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 15 September 2015
Received in revised form
28 December 2015
Accepted 30 December 2015
Available online 7 January 2016

Keywords:
Variable importance
Variable ranking
Biomarker discovery
Rank aggregation
Metabolomics

ABSTRACT

Biomarker discovery is one important goal in metabolomics, which is typically modeled as selecting the most discriminating metabolites for classification and often referred to as variable importance analysis or variable selection. Until now, a number of variable importance analysis methods to discover biomarkers in the metabolomics studies have been proposed. However, different methods are mostly likely to generate different variable ranking results due to their different principles. Each method generates a variable ranking list just as an expert presents an opinion. The problem of inconsistency between different variable ranking methods is often ignored. To address this problem, a simple and ideal solution is that every ranking should be taken into account. In this study, a strategy, called rank aggregation, was employed. It is an indispensable tool for merging individual ranking lists into a single "super"-list reflective of the overall preference or importance within the population. This "super"-list is regarded as the final ranking for biomarker discovery. Finally, it was used for biomarkers discovery and selecting the best variable subset with the highest predictive classification accuracy. Nine methods were used, including three univariate filtering and six multivariate methods. When applied to two metabolic datasets (Childhood overweight dataset and Tubulointerstitial lesions dataset), the results show that the performance of rank aggregation has improved greatly with higher prediction accuracy compared with using all variables. Moreover, it is also better than penalized method, least absolute shrinkage and selectionator operator (LASSO), with higher prediction accuracy or less number of selected variables which are more interpretable.

☆ Selected paper from 15th Chemometrics in Analytical Chemistry Conference, 22–26 June 2015, Changsha, China.

* Corresponding author.
E-mail address: yizeng_liang@263.net (Y.-Z. Liang).

http://dx.doi.org/10.1016/j.aca.2015.12.043
0003-2670/

1. Introduction

Metabolomics is an emerging field which combines strategies to identify and quantify cellular metabolites present in organisms, cells, or tissues using advanced analytical techniques with the application of statistical and multi-variant methods. One target of metabolomics research is to discover biomarkers, which can aid in the diagnosis of many diseases in the metabolic system, biological and clinical guidance. The discovery of biomarkers is typically modeled as selecting the most discriminating variables (metabolites) for classification (e.g., discriminating diseased versus healthy) [1], which is often referred to as variable importance analysis or variable selection in the language of statistics and machine learning. Any variable importance analysis method can be turned into variable selection by introducing a threshold on variable importance values. In the past two decades, a large amount of papers about biomarker discovery in the metabolomics studies have employed the statistical method [2–4] including univariate filtering method and multivariate methods (principal component analysis (PCA), partial least squares-linear discriminant analysis (PLS-DA) [5], support vector machine (SVM) [6], random forest (RForest) [7] and penalized method like least absolute shrinkage and selectionator operator (LASSO) [8] and elastic net [9]). Ranking of variable (metabolite) importance is often carried out with the help of the statistical method. The ranking is assigning a measure of importance to each variable. Then, a subset of all metabolites could be identified by setting a threshold value. Variable importance analysis methods consist of model-free and model-based approaches. One group of the model-free approaches is based on simple univariate test statistics methods without using the model information, such as t test, fold change and Wilcoxon rank-sum test. They reflect whether the difference between healthy and diseased groups' averages is statistically significant. The other group is based on the features of relationship between variables and classification label, such as correlation coefficient, information gain, Euclidean distance and mutual information. As for model-based approaches, they are tied to the model performance including model-fitting and model-prediction approach. Model-fitting approach is commonly used when the established model fits itself, which contains partial least squares (PLS) weights [10], PLS loadings [10], regression coefficient (RC) [11], variable importance in projection (VIP) [12] and selectivity ratio (SR) [13–15]. Model-prediction approach is based on the model prediction performance by means of resampling methods including jackknife, bootstrap and cross validation. Variable importance analysis based on random variable combination (VIAVC) [16], subwindow permutation analysis (SPA) [17], uninformative variable elimination (UVE) [18], random frog (RFrog) [19], margin influence analysis (MIA) [20], RForest and LASSO are assigned to the model-prediction approach. To date, researchers have often applied a lot of different variable important analysis methods to get as much as possible out of their data and present only the most favorable results, and then interpret the biomarkers that have a good performance with high classification accuracy. However, it is possible that there exist many different subsets of variables from different methods that can achieve the same or similar predictive accuracy. Different variable importance analysis methods are mostly likely to generate different variable rankings due to their different principles even when considering top-ranking variables. For instance, method “A” identifies the top three variables as potential biomarkers, whereas these biomarkers are not recognized as top ranking by method “B”. Both methods have discovered two different variable subsets that have the same classification accuracy. Consequently, it cannot assess which method is accurate and feasible. Moreover, inconsistency between variable rankings are often ignored in metabolomics research when presenting a new data set, because they would make the interpretation of results more confusing and arouse doubts on the reliability of their data. As Franklin D. Roosevelt once said that “there are as many opinions as there are experts”. Each method generates a variable ranking list just as an expert presents an opinion. The multiplicity of methods and the question of how to deal with the different ranking results are very general issues in the analysis of variable importance. To address this problem, a simple and ideal solution is to take every ranking into account.

In this study, a strategy that can combine all methods' ranking results, called rank aggregation firstly proposed by Vasyl et al. [21], was introduced. It is an indispensable tool to merge individual ranking lists into a single “super”-list reflective of the overall preference or importance within the population. In other word, it aims to find a “super”-list which would be as “close” as possible to all individual ranking lists simultaneously. This “super”-list is regarded as the final ranking for biomarker discovery. In this work, we employed nine variable importance analysis methods from the model-free and model based approaches, including t test, Wilcoxon rank-sum test, Relief, PLS-RC, PLS-VIP, SPA, RFrog, MIA and RForest. All nine methods were used on the two metabolomics data to generate the rankings of variable importance. Rank aggregation was then used to ensemble all the variable ranking and produce a “super” ranking list. Finally, this ranking list was used to select the best variable subset with the highest predictive classification accuracy.

2. Methods and theory

2.1. Variable importance analysis methods

As illustrated in the section of introduction, the variable importance analysis methods can be separated into two groups as model-free and model-based approaches. In this section, the methods we used for rank aggregation are introduced in brief.

2.1.1. Model-free approaches

Model-free approaches contain two different strategies. One is based on univariate test statistics methods, such as t test and Wilcoxon rank-sum test. The other one is based on the statistical features between variables and classification label such as correlation coefficient, information gain, Euclidean distance and Mutual information.

2.1.1.1. T test statistic. A t test's statistical significance indicates whether or not the difference between healthy and diseased groups' averages. The smaller the P-value of t test, the larger the significance difference of two groups. Each variable of data **X** will have a P-value when calculating the difference between healthy and diseased groups' averages by t test. All variables could be ranked with the P-value by ascend.

2.1.1.2. Wilcoxon rank-sum test. Wilcoxon rank-sum test is a nonparametric which is based solely on the order in which the samples from the two groups. The smaller the P-value of Wilcoxon rank-sum test, the larger the significance difference of two groups. Thus, as like the t test, all variables could be ranked based on the P-value by ascend. Since it compares to rank sums, the Wilcoxon rank-sum test is more robust than the t-test because it is less likely to manifest spurious results based on the existing of outliers.

2.1.1.3. Relief. Relief algorithm [22,23] is a correlation-based method assigning a “relevance” weight to each variable, which is meant to denote the relevance and redundancy analysis of the variable to the target concept. The core idea of Relief is to estimate