Chemometrics and Intelligent Laboratory Systems 180 (2018) 84–87

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Relational variable for more accurate prediction of models

Zhe Yuan [a,b,1], Liangxiao Zhang [a,c,e,f,*,1], Ruinan Yang [a,b], Jin Mao [a,e], Qi Zhang [a,d], Peiwu Li [a,c,d,e,**]

[a] Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, Wuhan, 430062, China
[b] Key Laboratory of Biology and Genetic Improvement of Oil Crops, Ministry of Agriculture, Wuhan, 430062, China
[c] Laboratory of Risk Assessment for Oilseed Products (Wuhan), Ministry of Agriculture, Wuhan, 430062, China
[d] Key Laboratory of Detection for Mycotoxins, Ministry of Agriculture, Wuhan, 430062, China
[e] Quality Inspection and Test Center for Oilseeds Products, Ministry of Agriculture, Wuhan, 430062, China
[f] Hubei Collaborative Innovation Center for Green Transformation of Bio-Resources, Wuhan, 430062, China

ARTICLE INFO

Keywords:
Relational variable
Model
Variable selection
Metabolomic
High accuracy

ABSTRACT

In natural science, models could grant new insights into phenomena or scientific problems which are hard to be observed or otherwise explained to overcome the limitations of human beings. Routinely, scientists strive to develop new methods for data acquisition, preprocessing, variable selection, modeling and valuation with the help of statistics and machine learning theories. Theoretically, the aim of these methods is global or local optimization in the space of variables and linear/nonlinear combinations for classification or regression. However, the relationships between responses and features are often complex and therefore sometimes far from linear or fixed nonlinear model. In this study, we proposed the relational variable (e.g. ratio between two variables) for more accurate prediction performance of models and illustrated its application on three classic data. We found that the selected relational variables could significantly improve the accuracy of prediction. The software was complemented on the MATLAB R2015a platform in Windows Server 2012 R2 standard. The Matlab codes used in this study are publicly available at http://www.libpls.net.

1. Introduction

Classification rules could transfer labels from annotated samples to unknown data points to classify two or more than two classes of samples and therefore uncover the mechanism of biological phenomena such as diseases [1,2]. Traditionally, statisticians or computational biologists mainly focus on development and rational use of new methods on data acquisition, preprocessing, variable selection, modeling and valuation. Theoretically, the aim of these methods is global or local optimization in the space of variables and their linear/nonlinear combinations for building a high performance classification or regression model [3,4]. Taking partial least squares regression (PLS regression) as an example, its aim to build a linear regression model by projecting the predicted variables and the observable variables to a new space [5]. However, whatever natural phenomena or diseases, the relationships between response and features are often complex and therefore sometimes far from linear or fixed nonlinear model. In this case, if the relationships between original variables could be described before data preprocessing and modeling, it might improve the prediction performance of classification or regression model. We term the descriptor of relationships between variables as Relational Variable (RV).

2. Materials and methods

2.1. Relational variable

Relational Variable is a kind of quantitative descriptor of relationships of variables. In this study, we demonstrated the efficiency of two kinds of simplest relational variables (the ratio and product of two original variables) on improvement of classification. As illustrated in Fig. 1, the ratio and product of two original variables were calculated, respectively. Then, these ratios and products of each pair of original variables were taken as relational variables. Finally, original data matrix ($X_{m\times n}$) could be transferred to data matrix of relational variable ($R^2_{m\times n}$).

* Corresponding author. Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, Wuhan, 430062, China.
** Corresponding author. Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, Wuhan, 430062, China.
E-mail addresses: liangxiao_zhang@hotmail.com (L. Zhang), peiwuli@oilcrops.cn (P. Li).
[1] These authors contributed equally to this study.

https://doi.org/10.1016/j.chemolab.2018.07.010
Received 21 March 2018; Received in revised form 14 June 2018; Accepted 27 July 2018
Available online 29 July 2018

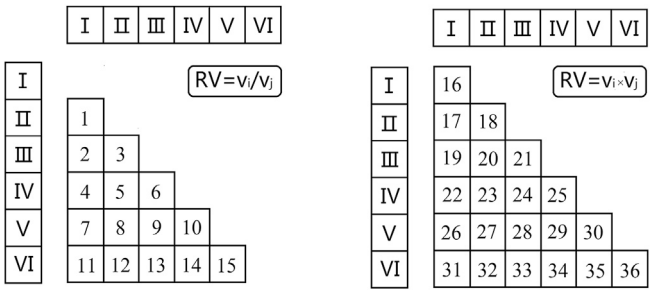

Fig. 1. Multivariate analysis of proteomics data from Gaucher patients (Green triangle) and healthy controls (Red point).

2.2. Datasets

SELDI-TOFMS data [6] were of blood samples collected from 20 patients with Gaucher disease (type I) and 20 healthy volunteers. Basic proteins were analyzed by SELDI-TOF MS making use of the anionic surface of CM10 ProteinChip® Arrays (Ciphergen Biosystems Inc., Fremont, CA, USA). Mass spectra with 590 mass to charge ratios were acquired in the positive-ion mode. Spot-to-spot calibration, baseline subtraction and peak detection were conducted by using the ProteinChip® Software. The detailed information of samples and experiments were described in the previous study [6].

Type 2 diabetes mellitus (T2DM) data [7] are of overnight fasting plasma samples collected from 45 T2DM patients and 45 healthy controls. Comprehensive analysis of plasma free fatty acids (FFAs) profiling of T2DM and healthy controls are performed based on GC-MS analysis after extraction and esterification reaction. Finally, 20 FFAs have been identified and quantified. The detailed information of samples and experiments were described in the previous study [7]. Metabolomic cancer diagnostics data contained spectroscopy 1H NMR spectroscopy (CPMG and NOESY-Presat), the fluorescence data as PARAFAC scores and Biomarker measurements (TIMP-1 and CEA) on Human plasma samples. This data includes 94 samples (47 patients undergoing large bowel endoscopy and 47 healthy volunteers) and 476 variables. The detailed information of samples and experiments were depicted in the previous study [8].

2.3. Implementation

The software was complemented on the MATLAB R2015a platform in Windows Server 2012 R2 standard. The codes used in this study are publicly available in http://www.libpls.net. This work did not analyze new data. SELDI-TOFMS data of Gaucher is publicly available in http://www.bdagroup.nl/content/Downloads/datasets/datasets.php. T2DM data is available in http://www.libpls.net/download.php. Metabolomic cancer diagnostics data is publicly available in http://www.models.life.ku.dk.

3. Results and discussion

Taking Gaucher patients and healthy controls as an example [6], surface enhanced laser desorption ionization time-of-flight mass spectrometry (SELDI-TOFMS) of 20 Gaucher patients and 20 healthy controls were employed to build a classification model for diagnose the Gaucher. RVs were generated by calculating the product and ratio of two original variables. Then, graphical index of separation (GIOS) [9] was employed to select 7 important RVs. As shown in Supplementary Fig. 1, fraction % reflects the distribution of one variable in two classes. If fraction equals 100, it means that all values of this variable in the one class are bigger or smaller than those in the other class. From this figure, it is found that fraction values of RVs are higher than those of original variables, indicating that they are more effective than original variables. Compared with the overlapping of two classes in Principal Component Analysis (PCA) scores plot of original data matrix ($X_{40\times590}$) (Fig. 2A), the same samples could be completely separated in PCA scores plot of data matrix of relational variables (Fig. 2B). The complete separation obtained by just using three relational variables (Fig. 2C) demonstrates that relational variables are informative for more accurate prediction of models. Monte Carlo cross-validation results of 10,000 permutations indicate that the prediction error of the model built by partial least squares linear discriminant analysis (PLS-LDA) could decrease to near zero for both classes of Gaucher patients and healthy controls, which is better than the result (the sensitivity of 89% and the specificity of 90%) in the previous study [6] and also better than the result of Kernel Principal Component Analysis (KPCA) (Fig. 2D). It means that surface enhanced laser desorption ionization time-of-flight mass spectrometry (SELDI-TOFMS) could be employed to diagnose the Gaucher with the advanced chemometrics with relational variables.

The similar results were obtained for metabolomic data. Type 2 diabetes mellitus (T2DM) [7] and metabolomic cancer diagnostics [8] were taken as two examples to illustrate this method. The competitive adaptive reweighted sampling (CARS) [10] method coupled with partial least squares linear discriminant analysis (PLS-LDA) was used to select the potential biomarkers or important RVs. 10-fold double cross validation [11] is employed to assess the prediction ability (accuracy, sensitivity and specificity). As shown in Table 1, accuracy, sensitivity and specificity of the model for identifying T2DM built by RVs are significantly better than the one built by original variables by using the same variable selection and classification modeling methods.

Meanwhile, area under curve (AUC, 0.99) of cancer diagnostics model built by RVs selected by GIOS and competitive adaptive reweighted sampling (CARS) [10] was also higher that the AUC in the previous study (see Table 2). Graphical index of separation (GIOS, 2478 RVs selected) [9] and the competitive adaptive reweighted sampling (CARS, 31 RVs selected) [10] were used to select the important RVs. Important original variables were selected based on VIP-score. Bootstrapped area under curve (AUC) and AUC in 10-fold double cross validation [11] is employed to measure classification ability. Moreover, we compared this method with non-linear and kernel based classification methods, random forest (RF) and support vector machine (SVM) with the help of MetaboAnalyst [12], respectively. As shown in Figs. S2 and S3, the out-of-bag (OOB) error was 0.372, while error rate of the best model was 36.2%. Compared with the above results, we can find that the model built by relational variables show better classification performance than original variables.

The above three examples indicate that RVs could improve the performance of classification model. Generally, traditional clustering and classification methods optimize linear or nonlinear combination of variables, but ignore the relationships between variables. Therefore, it is necessary to create more kinds of relational variables for more accurate prediction of models to some extent.

4. Conclusions

In this study, we proposed the relational variable for more accurate prediction performance of models and illustrated its application on three classic data. We found that the selected relational variables could significantly improve the accuracy of prediction. In future, we anticipate the relational variable to be a promoter for the following studies: (a) combination with the hundreds of existing modeling methods to build more effective models; (b) development of description and computation algorithms for relationship between variables to overcome the limitation of modeling methods (e.g. PLS just uses the linear combination of variables); (c) further investigation of relationships between variables, and between features/variables and responses for scientific purposes, which becomes the goal of analyzing Data.