Chemometrics and Intelligent Laboratory Systems 132 (2014) 124–132

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

MLRMPA: An R package of multiple linear regression model population analysis based on a cluster sampling technique for variable selection of high dimensional data

Meihong Xie, Fangfang Deng, Xiaoyun Zhang *, Yueli Tian, Peizhen Li, Honglin Zhai

College of Chemistry and Chemical Engineering, Lanzhou University, Lanzhou 730000, Gansu Province, People's Republic of China

ARTICLE INFO

Article history:
Received 24 September 2013
Received in revised form 20 January 2014
Accepted 22 January 2014
Available online 31 January 2014

Keywords:
Feature selection
Cluster sampling
Multiple linear regression model population analysis
R package

ABSTRACT

We develop an R package MLRMPA for fitting a pool of models between response variable and chemical descriptors. It is an embedded method combining feature selection with model building. The feature selection procedure is a cluster sampling method and different from model population analysis (MPA) that was implemented in a previously published study. The modeling process performs multiple stepwise regression analysis using the sampled features from the clustered group. This paper provides the algorithm and method implemented in the R package, which includes VarCor feature selection, cluster sampling, model building and model checking. This package is applied to establish an optimal linear model to predict the response and detect outliers from suboptimal models.

1. Introduction

Model population analysis (MPA) has become an effective method for outlier detection and variable selection since it was proposed by Liang et al. [1,3]. It works mainly in three steps: (a) randomly extract N sub-datasets from the calculated descriptors by Monte Carlo sampling (MCS); (b) build a sub-model for each subset, thus, N sub-models can be built; and (c) analyze the interesting outcome (e.g. prediction error) based on all the N sub-models from a statistical perspective. For outlier detection and variable selection, it has been demonstrated that MPA can provide comprehensive information of the data [4,5].

Motivated by the idea of MCS from the descriptors, an improved method, called multiple linear regression model population analysis (MLRMPA), is developed. The core of the algorithm is clustering all pre-selected descriptors into groups, and then sampling one descriptor by MCS from each group of clustered descriptors, which will constitute a new subset to establish a multiple stepwise regression model. This modified sampling technique is different from MPA that was implemented in a previously published study. The selecting of variable subset (X_{sub}, y_{sub}), known as a sub-window, is randomly produced using MCS from the original dataset (X, y). And the number of variables to be sampled is set to be a fixed value predefined by the user. Then a sub-model can be established using the sampled variables [1,4]. However, sub-window in MLRMPA is randomly sampled from clustered descriptors and then multiple stepwise regression analysis can be used by the sampled descriptors. It is the number of clusters that needs to be set by the user and the number of variables into the model depends on their importance. The sampling process is performed N times to build N models from where one can extract an optimal model.

Besides, MPA method doesn't exclude the obvious uninformative and redundant descriptors, such as descriptors' variance significantly different from that of response variable and those descriptors with lower correlation with the activity are not removed. Therefore, it is reasonable to take feature selection before the clustering procedure. However, throughout all the variable selection methods, they all aim to select a fixed subset of variables to build a linear or nonlinear model including genetic algorithms (GAs) [6,8], Lasso [9], least angle regression (LAR) [10], and elastic net (EN) [11]. And they can't meet the requirement to extract a pool of variable subsets to build a great number of models from a given dataset. So, we propose a new variable selection method, called variance correlation variable selection method (VarCor), to quickly delete uninformative and redundant descriptors based on the variables' variance and correlation distribution without reducing multi-collinear descriptors.

This paper provides both an algorithm on how VarCor-MLRMPA works and an introduction on the approach to implement MLRMPA in the R language environment [12,13]. Because of powerful programming language, R contributed packages have been a rapid expansion in

* Corresponding author. Tel.: +86 931 8912578; fax: +86 931 8912582.
E-mail address: xyzhang@lzu.edu.cn (X. Zhang).

0169-7439/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2014.01.010

different statistical areas, which are closing to 5000 [12]. In order to demonstrate the MLRMPA package, it is helpful to present an overview of the VarCor-MLRMPA process. Finally, in order to verify the performance of the method and package, we perform this process on 101 Setschenow constants of organic compounds that are the same with Xu et al.'s work [14,15]. Another two different linear methods (multiple linear regression and random forest) are compared with the model generated by proposed approach. The results show that, in the same dataset, the MLRMPA approach can establish a better linear model to predict the response and detect outliers from sub-optimal models.

2. Methods

To give an overview of VarCor-MLRMPA, a flow diagram of the algorithm is introduced as shown in Fig. 1b. It works mainly in five steps. For comparison, the workflow of MPA is also illustrated in Fig. 1a.

2.1. Preselection and VarCor selection

In pre-selection step, some descriptors will be excluded if they meet one of the following situations: (a) some samples miss the value; (b) there is a small variation in magnitude for all compounds; (c) the Fisher value of F-criterion is less than one unit; and (d) the value of t-criterion is less than the prior defined (by default 0.1) [17,18].

The VarCor variable selection method includes two procedures (variance variable selection and correlation variable selection) implemented by *VarCor* function. The aim of VarCor variable selection is to remove those descriptors whose variances are significantly different from that of the response variable and those descriptors have lower correlation with the activity. Because the descriptors' variances are significantly different from the activities', the distributions of these descriptors' are not agreements with those of the activity values. The lower correlation descriptors also need be rejected because of little effect on the response variable. However, high multi-collinear descriptors are remained to perform clustering step.

In the *VarCor* function, all descriptors as well as response variable are scaled according to the following formula [19]:

$$X_{\mathrm{ij}}^{\mathrm{n}} = \frac{X_{\mathrm{ij}} - X_{\mathrm{j,min}}}{X_{\mathrm{j,max}} - X_{\mathrm{j,min}}} \qquad (1)$$

where X_{ij} and $X_{\mathrm{ij}}^{\mathrm{n}}$ are the original and scaled values of the *j*th variable of the *i*th compound, respectively; and the superscript n is a sign that represents the scaled data; $X_{\mathrm{j,min}}$ and $X_{\mathrm{j,max}}$ represent minimum and maximum values of the *j*th descriptor. The aim of normalization is to ensure that all the descriptors and response are comparable.

The calculation of the variance is computed according to the formula:

$$s_{\mathrm{j}}^2 = \frac{\sum\left(X_{\mathrm{ij}}^{\mathrm{n}} - \overline{X_{\mathrm{j}}^{\mathrm{n}}}\right)}{n} \qquad (2)$$

where $\overline{X_{\mathrm{j}}^{\mathrm{n}}}$ is the mean of the scaled *j*th descriptor; n is the sample size. Similarly, the variance of activity can be worked out (s_{y}^2).

The correlation between a variable and a dependent variable is calculated as follows:

$$r_{\mathrm{j}} = \frac{\sum\left(X_{\mathrm{ij}}^{\mathrm{n}} - \overline{X_{\mathrm{j}}^{\mathrm{n}}}\right)\left(y_{\mathrm{i}}^{\mathrm{n}} - \overline{y^{\mathrm{n}}}\right)}{\sqrt{\sum\left(X_{\mathrm{ij}}^{\mathrm{n}} - \overline{X_{\mathrm{j}}^{\mathrm{n}}}\right)^2 \sum\left(y_{\mathrm{j}}^{\mathrm{n}} - \overline{y^{\mathrm{n}}}\right)^2}} \qquad (3)$$

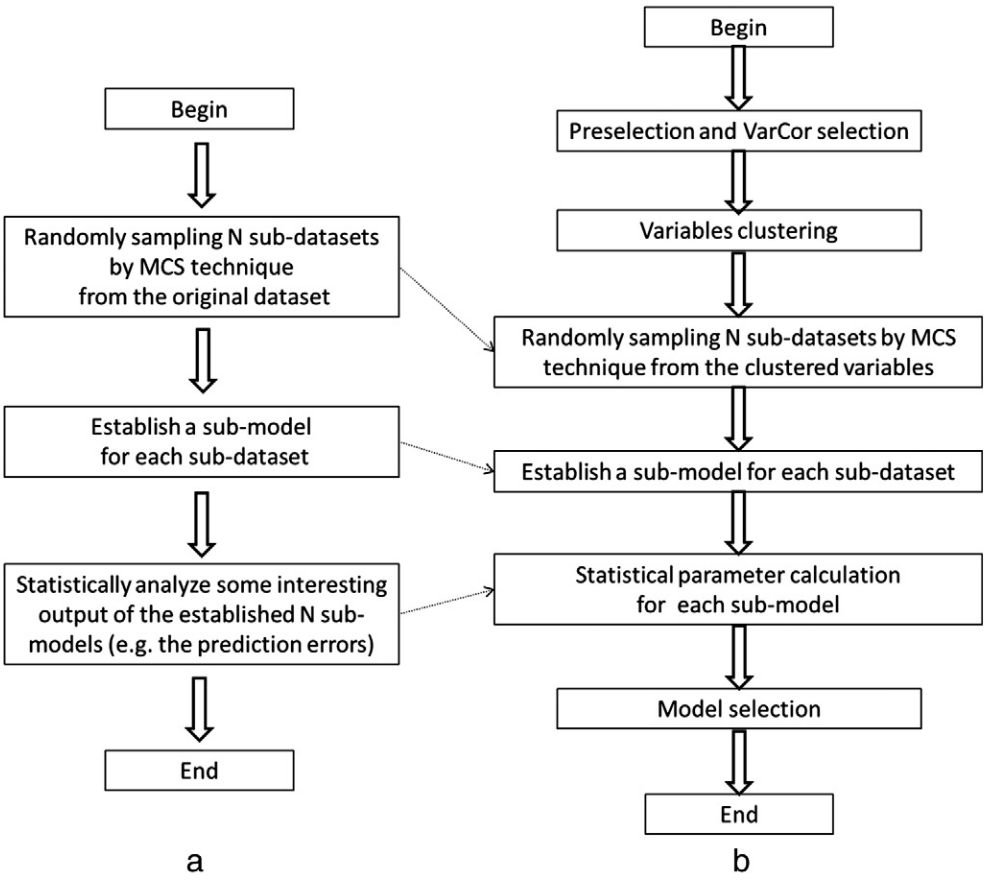
Begin
Randomly sampling N sub-datasets by MCS technique from the original dataset
Establish a sub-model for each sub-dataset
Statistically analyze some interesting output of the established N sub-models (e.g. the prediction errors)
End
a
Begin
Preselection and VarCor selection
Variables clustering
Randomly sampling N sub-datasets by MCS technique from the clustered variables
Establish a sub-model for each sub-dataset
Statistical parameter calculation for each sub-model
Model selection
End
b

Fig. 1. The flow diagram of MPA (a) and MLRMPA (b).