Chemometrics and Intelligent Laboratory Systems 149 (2015) 166–176

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Model population analysis in chemometrics

Bai-Chuan Deng [1], Yong-Huan Yun [1], Yi-Zeng Liang *

College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, PR China

ARTICLE INFO

Article history:
Received 13 April 2015
Received in revised form 21 August 2015
Accepted 22 August 2015
Available online 29 August 2015

Keywords:
Model population analysis
Chemometrics
Variable selection
Model evaluation
Outlier detection
Applicability domain

ABSTRACT

Model population analysis (MPA) is a general framework for designing new types of chemometrics algorithms that has attracted increasing interest in the chemometrics community in recent years. The goal of MPA is to extract statistical information from the model, towards better understanding of the chemical data. Two key elements of MPA are random sampling and statistical analysis. The core idea of MPA is quite universal with potential applications in the fields, such as chemoinformatics, biostatistics and bioinformatics.

In this article, we review the development of MPA in chemometrics. We first present the key elements of MPA. Then, the application of MPA in chemometrics is discussed, such as variable selection, model evaluation, outlier detection, applicability domain definition and so on. Finally, the potential application areas of MPA in future research are prospected.

1. Introduction

Model population analysis (MPA) is a general framework for developing a new type of algorithm for modeling which uses statistical tools to extract important information from the model [1]. The concept of MPA was firstly proposed by Li et al. in the field of variable selection [2]. It was found that better insight into the data can be obtained with the aid of random sampling on both variable space and sample space. Before that, a similar idea has been applied to outlier detection by Cao et al. [3] with a discovery suggesting that normal samples and outliers can be separated through their distributions in random sub-models. Recent studies show that the application range of MPA should not be restricted to variable selection [4–7] and outlier detection [3]. It is also a useful tool for model comparison [8] and applicability domain definition [9].

Chemometrics is a chemical discipline that uses mathematical and statistical methods, to design or select optimal measurement procedures and experiments, and to provide maximum chemical information by analyzing chemical data [10]. Chemical modeling is an important content in chemometrics, which is usually referred to as multivariate calibration for regression models and pattern recognition for classification models. The aim of chemical modeling is to develop a quantitative relation between the variables, e.g., wavelengths, molecular descriptors, and properties of interest and e.g., concentration values, molecular activities. MPA, as a powerful tool for modeling, is promising in chemometrics and related fields, because it provides better understanding of the chemical data and improves prediction and interpretation of the model.

It is worth noting that ensemble learning methods can also be formulated into the framework of MPA. Ensemble learning methods, such as bagging [11], boosting [12] and random forests [13], aggregate a large number of models built with sub-datasets randomly generated using a resampling method like bootstrapping. Then predictions are made based on the principle of majority voting for classification or averaging for regression [14]. In our view, part of the idea in ensemble learning methods is the same as that in MPA. However, MPA is an extension of ensemble ideology and is more general. MPA is a framework, in which different blocks can be filled. Firstly, the random sampling techniques can be varied. Secondly, the generating of sub-models is not restricted to sample space and variable space. Besides, various outputs from sub-models can be considered. Finally, different statistical analysis techniques can be applied on the outputs from a large population of sub-models.

2. Model population analysis (MPA)

The core idea of MPA is to statistically analyze the performance of a large population of sub-models generated from random sampling and to extract interesting information from outputs of the sub-models [1]. The key elements of MPA are random sampling and statistical analysis. An important feature of MPA is that it considers the output of interest not as a single value but a distribution [15].

* Corresponding author. Tel.: +86 731 8830824; fax: +86 731 8830831.
E-mail address: yizeng_liang@263.net (Y.-Z. Liang).
[1] The first two authors contributed equally to this work.

http://dx.doi.org/10.1016/j.chemolab.2015.08.018
0169-7439/

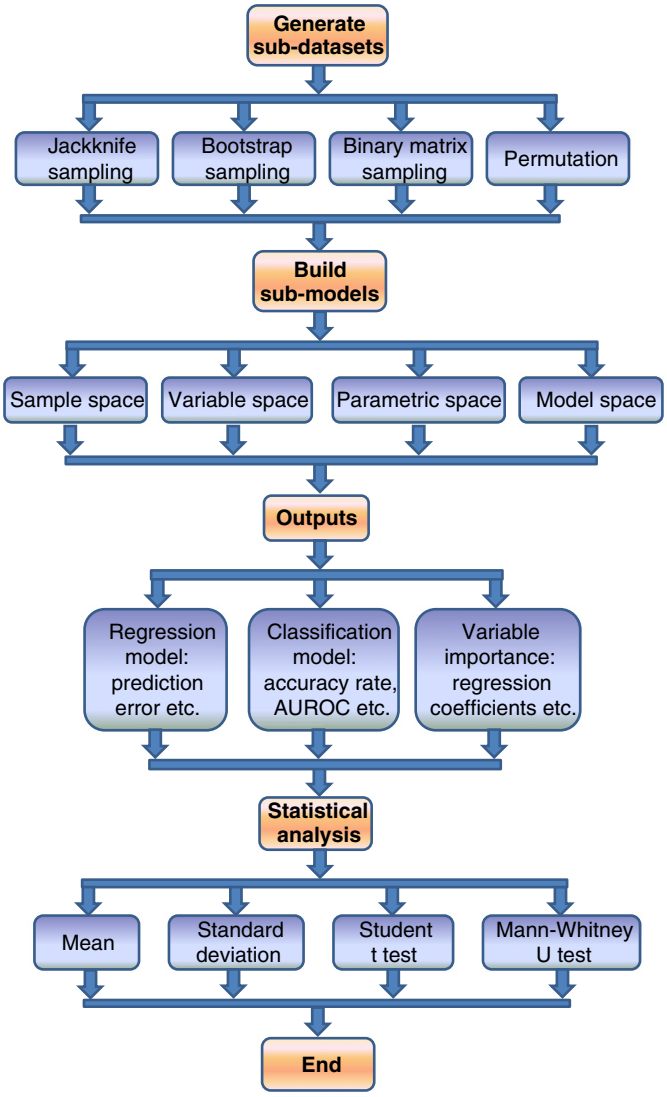

Fig. 1. The framework of model population analysis.

2.1. The framework of model population analysis

MPA generally works as follows (Fig. 1):

(1) random sampling is used to randomly draw *k* sub-datasets (e.g., 5000);
(2) for each sub-dataset, a sub-model is built;
(3) calculate an output of interest (e.g., prediction errors) for each sub-model; and
(4) statistical analysis is applied on the outputs of sub-models.

Sub-models may be built using random sampling in sample space, variable space, parametric space and model space. Accordingly, the information extracted for MPA may be the performance of samples, variables, parameters and models. This framework has proven to be useful for developing algorithms for outlier detection [3], variable selection [2], model evaluation [16] and applicability domain definition [9]. MPA is a part of libpls package, which is implemented in MATLAB and is freely available at www.libpls.net.

2.2. Random sampling techniques

To generate a diverse population of sub-models, random sampling, i.e., Monte Carlo sampling (MCS), techniques are usually applied in MPA to obtain diversity of sub-sample sets or sub-variable sets, among which Jackknife sampling (JNS), bootstrap sampling (BSS) and binary matrix sampling (BMS) are the most popular ones. The core ideas of these random sampling methods are displayed in Fig. 2, where orange cells indicate objects that are selected and white cells denote objects that are not selected.

JNS is a random sampling technique without replacement [17]. The procedure of JNS is displayed in Fig. 2A, where each row of the matrices denotes individual sampling. Assume that we are to randomly select 3 objects from 5 individual objects using JNS. In each row, 5 objects are numbered and the sequence of objects is randomized (Fig. 2A, left). Then, the first 3 objects are picked out (marked in orange) from the pool of objects (Fig. 2A, left). On the right part of Fig. 2A, the selected objects are marked as orange cells and are assigned in sequence. This procedure realizes random sampling of 3 objects from 5 individual objects with no replacement. In this figure, 4 rows of the matrices denote 4 times of random sampling.

BSS is a random sampling technique with replacement [18]. Each row of the matrices denotes individual sampling. In each sampling, objects are picked out from the pool of objects randomly and successively as it is shown in Fig. 2B. The difference of BSS compared to JNS is that samples are replaced during resampling. Consequently, some objects may be selected more than once while the others may not be selected (Fig. 2B left). There is an option either to retain the repeated objects or just retain unique objects. In this schematic diagram, the unique objects are retained and assigned in sequence by orange cells (Fig. 2B, right). A modification of BSS is called weighted bootstrap sampling (WBS), where samples with different possibilities [19] are selected.

BMS realizes random sampling with the help of a binary matrix [20]. The core idea of BMS is to randomly assign the same number of '1' and '0' to each column of the binary matrix (Fig. 2C, left). Each row of the binary matrix corresponds to an individual sampling, where '1' denotes the samples that are selected for modeling and '0' denotes the samples that are not selected. The same number of "1" is assigned to each column. Thus, BMS ensures that each sample is selected at the same frequency after all individual samplings (4 individual samplings are shown in Fig. 2C). The selected objects are marked as orange cells and are assigned in sequence, as it is shown in Fig. 2C (right). Weighted binary matrix sampling (WBMS) is a modification of BMS, where samples are selected at different frequencies [7]. More details on the statistical feature of different random sampling methods can be found in the book [21].

Permutation (or randomization), unlike any random sampling method, destroys the connection between samples and observations [22]. A common way to realize it is to permute the values of responses or observations among samples, so as to break the one-to-one correspondence between responses and observations. It is also a useful technique for MPA, because a variety of models, mostly 'wrong' models, can be built in this way. Useful information, such as variable importance [5,23] and over-fitting [24,25], can be extracted through statistical analysis of the variety of models in permutation.

2.3. Statistical analysis

The core idea of MPA is statistical analysis of an interesting output, of all the sub-models. With a population of sub-models generated by random sampling methods, model comparison is performed by using an empirical distribution derived from the interesting outputs. For example:

1. The distributions of prediction errors for normal samples and outliers follow different patterns;