Analytica Chimica Acta 807 (2014) 36–43

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Chemometrics

A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration

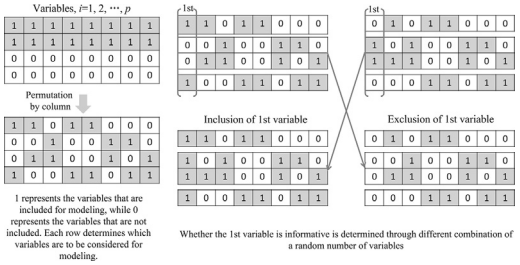
Yong-Huan Yun [a], Wei-Ting Wang [a], Min-Li Tan [a], Yi-Zeng Liang [a,*], Hong-Dong Li [a], Dong-Sheng Cao [b], Hong-Mei Lu [a], Qing-Song Xu [c]

[a] College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, PR China
[b] College of Pharmaceutical Sciences, Central South University, Changsha 410083, PR China
[c] School of Mathematics and Statistics, Central South University, Changsha 410083, PR China

HIGHLIGHTS

- Considers the possible interaction effect among variables through random combinations.
- Four kinds of variables were distinguished.
- It was more efficient when compared with CARS, GA-PLS and MC-UVE-PLS.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 21 August 2013
Received in revised form
13 November 2013
Accepted 14 November 2013
Available online 21 November 2013

Keywords:
Variable selection
Informative variables
Partial least squares
Iteratively retaining informative variables
Random combination

ABSTRACT

Nowadays, with a high dimensionality of dataset, it faces a great challenge in the creation of effective methods which can select an optimal variables subset. In this study, a strategy that considers the possible interaction effect among variables through random combinations was proposed, called iteratively retaining informative variables (IRIV). Moreover, the variables are classified into four categories as strongly informative, weakly informative, uninformative and interfering variables. On this basis, IRIV retains both the strongly and weakly informative variables in every iterative round until no uninformative and interfering variables exist. Three datasets were employed to investigate the performance of IRIV coupled with partial least squares (PLS). The results show that IRIV is a good alternative for variable selection strategy when compared with three outstanding and frequently used variable selection methods such as genetic algorithm-PLS, Monte Carlo uninformative variable elimination by PLS (MC-UVE-PLS) and competitive adaptive reweighted sampling (CARS). The MATLAB source code of IRIV can be freely downloaded for academy research at the website: http://code.google.com/p/multivariate-calibration/downloads/list.

1. Introduction

Variable (feature or wavelength) selection techniques have become a critical step in the analysis for the datasets with hundreds of thousands of variables in several areas. These areas include genomics, bioinformatics, metabonomics, near infrared and Raman spectroscopy, and quantitative structure-activity relationship (QSAR). The goal of variable selection can be summarized in three aspects: (1) improving the prediction performance of the predictors, (2) providing faster and more cost-effective predictors by reducing the curse of dimensionality, (3) providing a better understanding and interpretation of the underlying process that generated the data [1,2]. In the face of the situation that the number of samples is much smaller than the number of variables (large p, small n) [3,4], a large amount of variable selection methods has been employed to tackle this challenge in the field of multivariate calibration. There are two directions in these methods. One is

* Corresponding author. Tel.: +86 731 8830824/+86 13808416334; fax: +86 731 8830831.
E-mail addresses: yizeng_liang@263.net, yizengliang@gmail.com (Y.-Z. Liang).

0003-2670/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.aca.2013.11.032

based on statistical features of the variables through some kind of criteria, such as correlation coefficient, *t*-statistics and Akaike information criterion (AIC), and selects the significant variables. This kind of method includes uninformative variable elimination (UVE) [5], Monte Carlo based UVE (MC-UVE) [6], competitive adaptive reweighted sampling (CARS) [7,8], successive projection algorithm (SPA) [9], random frog [10,11]. Although they have computational efficiency, a common disadvantage is that they select the variables individually while ignoring the joint effect of variables. The other kind is based on the optimized algorithm to search the optimal variables. However, since the space of variable subsets grows exponentially with the number of variables, a partial search scheme is often used, such as stepwise selection [12], forward selection [12,13], backward elimination [12–14], genetic algorithm (GA) [15–21] and simulated annealing (SA) [22]. It is unpractical to conduct the greedy search by exhaustively searching through all possible combinations of variable sets. The process of finding the best variable subsets, not only using an effective search scheme but also considering the interaction of variables in the search space, has been an important part in variable selection research. Margin influence analysis (MIA) proposed by Li et al. [23] considers the combination of variables using Monte Carlo sampling technique and ranks all the variables based on *P* value of hypothesis testing, but the number of variables to be sampled is often set to be a fixed parameter predefined by user's input. Sampling a fixed set number of variables result in the chance of every variable to be sampled not being the same. Some are selected more frequently, but some not. Binary matrix shuffling filter (BMSF) [24] provides effective search for informative variables in the infinite dimensional search space, and also considers the synergetic effect among multiple variables through random combination. It consists of a guided data-driven random search algorithm using a binary matrix to convert the high-dimensional variables space search into an optimization problem. Different from MIA, BMSF assures that all the variables have the same chances to be sampled with the binary matrix. Furthermore, BMSF is an iterative filtering algorithm [25] which eliminate about half of the variables on each round of filtering using a criterion.

Based on the core idea of BMSF, in this study, we propose a novel variable selection strategy, called Iteratively Retaining Informative Variables (IRIV). Additionally, we also divided the variables into four categories as strongly informative, weakly informative, uninformative and interfering variables, which is regarded as an improvement of the definition on the classification of variables by Wang and Li [26], where the variables are only classified into three categories, as informative, uninformative and interfering variables. Their definition is illustrated in next section. As the name of IRIV implies, it conducts many rounds until it reaches convergence. For every round, a large amount of models is generated. Here, we introduce a novel methodology called model population analysis (MPA) [23,26–28] to distinguish the four kinds of variables through the analysis of a 'population' of models. Generally, strongly informative variables indicate that they are always necessary for an optimal subset. Some variable selection methods use them to constitute the optimal variable set ignoring the effect of weakly informative variables. However, weakly informative variables contribute as well due to their beneficial effect. Based on this idea, IRIV retains both the strongly and weakly informative variables in every iterative round until no uninformative and interfering variables exist. When compared to obtaining the informative variables in one round, this approach can avoid that some uninformative and interfering variables appear in the strongly and weakly informative variables set by chance. Finally, backward elimination is conducted to obtain the most optimal variable set.

In this work, IRIV coupled with partial least squares (PLS) [29] was investigated through the analysis of three real near infrared spectral datasets and better results were obtained than with the three outstanding methods. This demonstrates that IRIV is a good alternative of variable selection algorithm for multivariate calibration. It should be pointed out that although IRIV was just tested by NIR dataset, it is a general strategy and can be coupled with other regression and classification methods and applied for other kinds of data, such as genomics, bioinformatics, metabonomics and quantitative structure–activity relationship (QSAR) [30].

2. Methods

Given that the data matrix **X** contains *N* samples in rows and *P* variables in columns, and **y**, of size $N \times 1$, denotes the measured property of interest. In this study, PLS was used as multivariate calibration method in the IRIV strategy. The detailed approach of IRIV is illustrated as follows:

Step 1: In order to consider the combination of multiple variables to be used for modeling, a binary matrix **M** that just contains either 1 or 0 with dimensions $K \times P$ is generated. *K* is the number of random combination of variables. *P* is the number of variables. The number "1" represents the variables that are included for modeling, while 0 represents the variables that are not included. It should be mentioned that the binary matrix **M** contains *KP*/2 ones and *KP*/2 zeros, which guarantees that there are the same choices for all variables to be included or excluded for modeling. Besides, the **M** is permuted by column. The number of ones and zeros in each column is the same. Each row of **M** determines which variables are to be considered for modeling. Different rows of **M** give different variables sets that contain different combinations of a random number of variables. This process is clearly illustrated in Fig. 1A.

Step 2: Each row of **M** is used for modeling with PLS method. Here, the mean squared error of five-fold cross validation (RMSECV) is used to assess the performance of each variables subset. Thus, a RMSECV vector ($K \times 1$), denoted as RMSECV_0, can be obtained.

Step 3: For the sake of assessing each variable's importance through its interaction with other variables, a novel strategy is used. That is, in a variable set (a row of **M**), the performance of the inclusion and exclusion of one variable is compared, while the state (inclusion or exclusion) of other variables are left unchanged. Thus, for the *i*th ($i = 1, 2, \ldots, P$) variable, we can obtain matrix **M1** by changing all the ones in *i*th column of **M** to zero and all the zeros in *i*th column to one, while keeping other columns of **M** unchanged (shown in Fig. 1B). RMSECV_i ($K \times 1$) can be obtained after finishing all rows of $\mathbf{M}_i$. Then, Φ_0 and Φ_i are collected to assess the importance of each variable as the following formulas:

$$\Phi_{0k} = \begin{cases} k^{\mathrm{th}}RMSECV_0 & \text{if } M_{ki} = 1 \\ k^{\mathrm{th}}RMSECV_i & \text{if } M1_{ki} = 1 \end{cases}, \quad \Phi_{ik} = \begin{cases} k^{\mathrm{th}}RMSECV_0 & \text{if } M_{ki} = 0 \\ k^{\mathrm{th}}RMSECV_i & \text{if } M1_{ki} = 0 \end{cases} \tag{1}$$

Where M_{ki} is the value in the *k*th row and *i*th column of **M**, and $\mathrm{M1}_{ki}$ is the value in the *k*th row and *i*th column of **M1**. Briefly, Φ_0 collects the value of RMSECV_0 and RMSECV_i that the *i*th variable is included in the variable set for modeling, while Φ_i collects the value of RMSECV_0 and RMSECV_i that the *i*th variable is excluded in the variable set for modeling. Notice that the values of Φ_0 and Φ_i are put in the same order as the rows of **M**. Such arrangement ensures that the Φ_0 and Φ_i are in pairs so as to focus on how significant the *i*th variable makes a contribution with various combinations of multiple variables in the modeling, while the conditions of the other variables are held the same. Both Φ_0 and Φ_i contain the results of K models. Therefore, model population analysis (MPA) [23,26–28,31] can be used to analyze the variable importance with these two 'population' models.

Step 4: Calculate the average value of Φ_0 and Φ_i, denoted as $\mathrm{MEAN}_{i,\mathrm{include}}$, $\mathrm{MEAN}_{i,\mathrm{exclude}}$. The difference of the two mean values can be calculated as:

$$\mathrm{DMEAN}_i = \mathrm{MEAN}_{i,\mathrm{include}} - \mathrm{MEAN}_{i,\mathrm{exclude}} \tag{2}$$

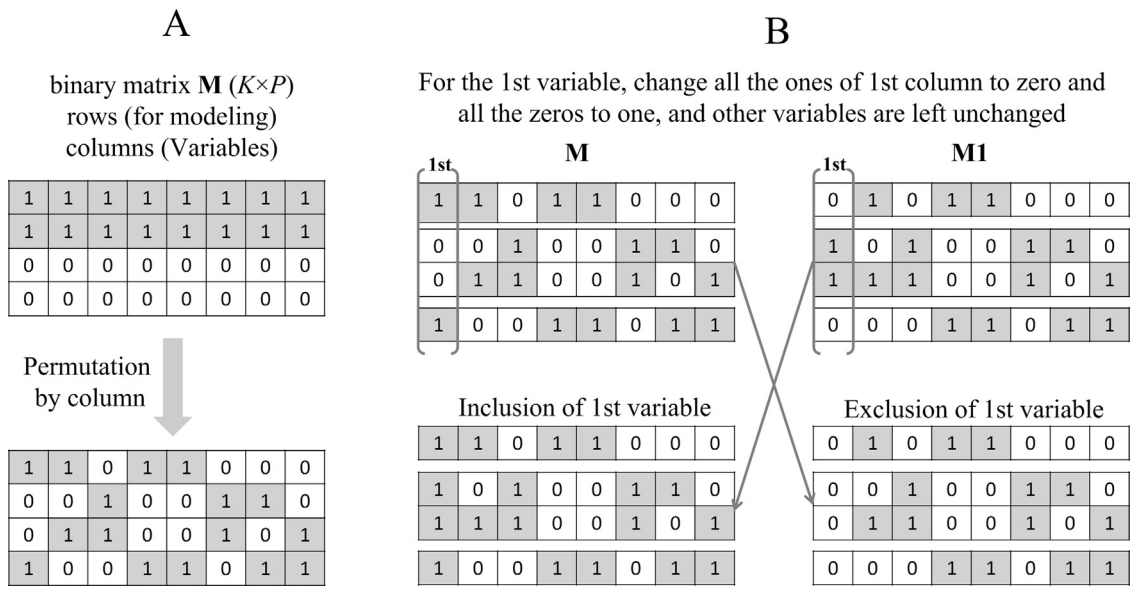


Fig. 1. (A) the process of generating binary matrix; (B) the performance of the inclusion and exclusion of one variable is compared.

Here, the four kinds of variables, such as strongly informative, weakly informative, uninformative and interfering variables, can be distinguished by Eq. (2) and hypothesis testing. In this study, the nonparametric test method, the Mann–Whitney *U* test [32], is used. *P* value, denoted as P_i, is computed by the Mann–Whitney *U* test with the distribution of Φ_0 and Φ_i. With respect to statistics, the smaller P_i value, the more significant difference between the two distributions. $P = 0.05$ is predefined as the threshold. In this sense, with the *P* value and DMEAN_i, we can easily classify the variables into four categories as follows:

$$V_i \in V_{\text{strongly informative}} \quad \text{if } \text{DMEAN}_i < 0, \quad P_i < 0.05 \tag{3}$$

$$V_i \in V_{\text{weakly informative}} \quad \text{if } \text{DMEAN}_i < 0, \quad P_i > 0.05 \tag{4}$$

$$V_i \in V_{\text{uninformative}} \quad \text{if } \text{DMEAN}_i > 0, \quad P_i > 0.05 \tag{5}$$

$$V_i \in V_{\text{interfering}} \quad \text{if } \text{DMEAN}_i > 0, \quad P_i < 0.05 \tag{6}$$

where $V_{\text{strongly informative}}$, $V_{\text{weakly informative}}$, $V_{\text{uninformative}}$ and $V_{\text{interfering}}$ indicate that strongly informative, weakly informative, uninformative and interfering variables, respectively. From the above definition, we can see that the strongly informative variables not only have a beneficial effect on the modeling ($\text{DMEAN}_i < 0$), but also are more significantly important ($P_i < 0.05$). The weakly informative variables have the beneficial effect but not significantly. The interfering variables have a significantly bad effect because of $\text{DMEAN}_i > 0$ and $P_i < 0.05$, while the uninformative variables have a bad effect but not significantly. Finish all variables $i = 1, 2, \ldots, P$ by step 3 and step 4.

Step 5: Remove the uninformative and interfering variables, and retain the strongly informative and weakly informative variables according the above definition. Go back to step 1 to perform the next round until no uninformative and interfering variables exist.

Generally, only the strongly informative variables are selected as the optimal variable set. Although they have significant positive effect, they are not always the optimal ones because of their ignoring the positive effect of the weakly informative variables. Thus, the weakly informative variables should be retained. Conducting many iterative rounds not one round intends to explore the uninformative and interfering variables that are not significant from the last round. Thus, IRIV strategy can search the significant variables through many rounds until no uninformative and interfering variables exist.

Step 6: Conduct backward elimination strategy with the retained variables as the below procedure:

(a) Denote *j* to be the number of retained variables
(b) With all *j* variables, obtain the RMSECV value with five-fold cross validation using PLS, denoted as θ^j.
(c) Leave out the *i*th variable and use the remaining $j-1$ variables in five-fold CV to obtain the RMSECV value θ_{-i}. Conduct this for all variables $i = 1, 2, \ldots, j$.
(d) If $\min\{\theta_{-i}, 1 \le i \le j\} > \theta^j$, go to step (g)
(e) When excluding the *i*th has the minimum RMSECV value, remove the *i*th variable and change the *j* to be $j-1$
(f) Repeat (a)–(e)
(g) The retained variables are the final list of informative variable.

After several iterative rounds, the number of the retained variables is relatively small. Fine evaluation of conducting backward elimination strategy can perform better because each variable is taken into account its interaction with other variables. The whole process of IRIV is briefly illustrated in Fig. 2.

3. Datasets and software

3.1. Diesel fuels dataset

This benchmark dataset is freely available at http://www.eigenvector.com/data/SWRI/index.html. The original data set includes 20 high leverage samples and the remaining samples are split into two random groups, which were measured at 401 wavelength channels from 750 to 1550 nm with 2 nm intervals. The property of interest is boiling point at 50% recovery (BP50). As the original author suggested, the 20 high leverage samples and one of the two random groups (113 samples) were used to make up the calibration set (133 samples) and the other random group (113 samples) was used as the independent test.

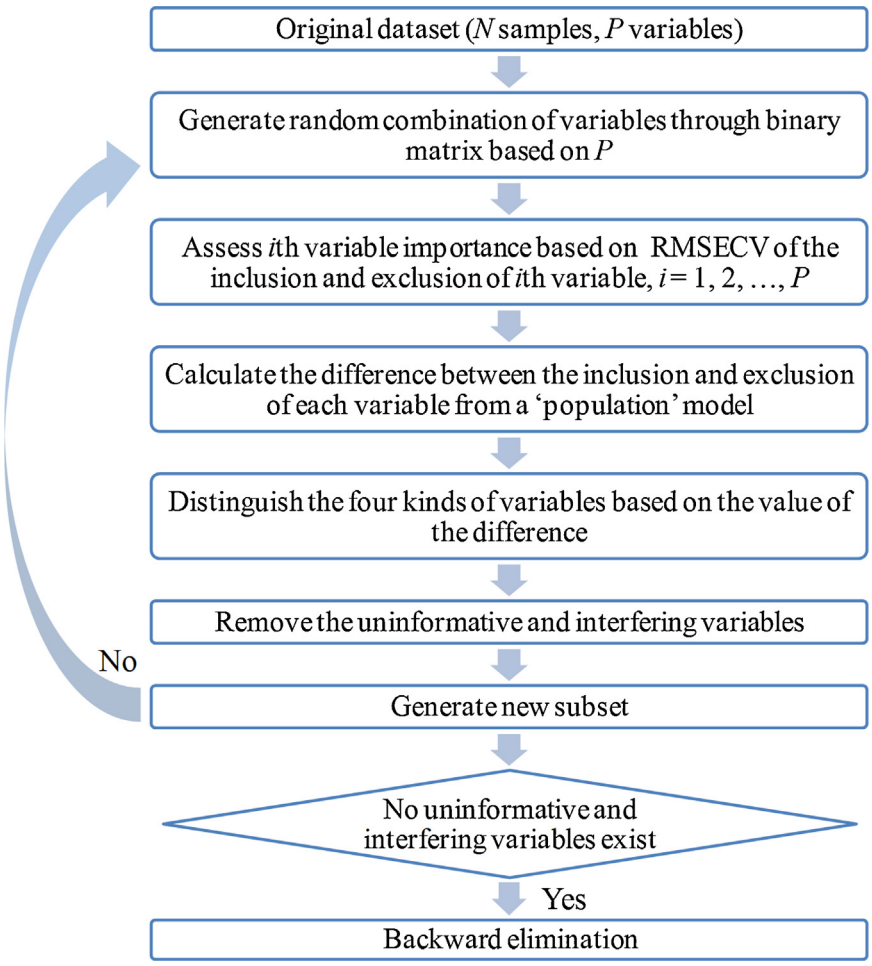


Fig. 2. The flow chart of IRIV algorithm. The number of the variables, *P*, changes after removing the uninformative and interfering variables.

3.2. Soy dataset

Data set Soy [33]: NIR spectra of soy flour samples (54 samples), on which oil responses is used for this study. The spectra have been recorded from 1104 to 2496 nm with an interval of 8 nm (175 wavelengths). Forty samples were used to make up the calibration set and the other 14 samples were used as the independent test according to the reference.

3.3. Corn dataset

The corn dataset is available in the website: http://www.eigenvector.com/data/Corn/index.html. This dataset consists of 80 samples of corn measured on three different NIR spectrometers. Each spectrum contains 700 spectral points at intervals of 2 nm within the wavenumbers range 1100–2498 nm. In the present study, the NIR spectra of 80 corn samples measured on m5 instrument were considered as **X** and the moisture value was considered as property of interest **y**. In addition, the dataset was divided into calibration set (80% of the dataset, 64 × 700) and independent test set (20% of the dataset) on the basis of Kennard–Stone (KS) method [34].

3.4. Software

All the computations were performed by using in-house code in MATLAB (Version 2010b, the MathWorks, Inc.) on a general-purpose computer with Inter® Core® i5 3.2 GHz CPU and 3GB of RAM. The Microsoft Windows XP was used as the operating system of the computer. The Matlab code of IRIV is available at the website: http://code.google.com/p/multivariate-calibration/downloads/list.

4. Results and discussion

In order to demonstrate the performance of IRIV strategy, three good variable selection methods were employed for comparison, including CARS, GA-PLS [21,35], and MC-UVE-PLS. The defined parameters of GA are listed in Table 1.

In this work, the calibration set was used for variable selection and modeling, and the independent test set was used for validation

Table 1
The parameters of the GA-PLS.

- Population size: 30 chromosomes
- On average, five variables per chromosome in the original population
- Response: cross-validated explained variance % (five deletion groups; the number of PLS components is determined by cross-validation)
- Maximum number of variables selected in the same chromosome: 30
- Probability of cross-over: 50%
- Probability of mutation: 1%
- Maximum number of components: 10
- Number of runs: 100
- The amount of evaluations: 200
- Backward elimination after every 100th evaluation

Table 2
The results of different methods on the diesel fuels dataset.

Methods	*n*VAR	*n*LVs	RMSEC	RMSEP
PLS	401	10	3.0643	3.8597
GA-PLS	61.0 ± 20.5	8.8 ± 1.4	2.7438 ± 0.1334	3.4113 ± 0.1872
CARS	17.1 ± 3.9	7.9 ± 1.4	2.7856 ± 0.1067	3.5022 ± 0.1733
MC-UVE-PLS	110.7 ± 51.2	9.8 ± 0.4	2.8538 ± 0.1178	3.4477 ± 0.1367
IRIV	47.0 ± 9.5	9.1 ± 0.9	2.3609 ± 0.0591	3.1568 ± 0.0727

of the calibration model. The maximum latent variable was set to 10. The number of latent variables was determined by 5-fold cross validation. All the data was first centered to have zero mean before modeling. Also, the root mean square error of calibration set (RMSEC) and the root mean square error of prediction of test set (RMSEP) were used to assess the performance of the model.

4.1. The problem on setting the dimension of the binary matrix **M** ($K \times P$)

There is no doubt that the number of random combinations should not be too small, assuring that the interaction of one variable with other variables is enough. In this work, for these three dataset, the K of each round was set based on the number of variables P as follows:

if $500 = P < 1000$, $K = 500$

if $300 = P < 500$, $K = 300$

if $200 = P < 300$, $K = 200$

if $100 = P < 200$, $K = 100$

if $50 = P < 100$, $K = 50$

if $P < 10$ go to the backward elimination step (Step 6)

It should be noted that the number of K has just a little impact on assessing the overall importance of each variable, as a result of conducting many iterative rounds to filter the uninformative and interfering variables.

4.2. Diesel fuels dataset

Table 2 shows results obtained with the different methods. Note that CARS is a very fast method but not always stable due to Monte Carlo sampling. MC-UVE-PLS is also based on Monte Carlo sampling strategy. For GA-PLS, the chromosome of the existing population is selected randomly. Furthermore, because of random combination, IRIV may also obtain different results when conducted many times. Thus, all methods were conducted 50 times to obtain statistical results. The mean and standard deviation are both given in Table 2. The frequencies of variables selected by these methods within 50 times are shown in Fig. 3.

When compared to the results of using the full spectrum, the RMSEC and RMSEP values of IRIV decrease 23% and 18%, respectively. Clearly, IRIV has highly improved the prediction performance with a reduced dimensionality. A boxplot is used in Fig. 4A for a better display of the prediction performance of the four methods. From Table 2 and Fig. 4A, we also can observe that IRIV obviously has the best prediction performance than other methods with the lowest RMSEC and RMSEP.

As we can see from Fig. 3, for GA-PLS, the shadow regions, which represent very significant variables, are consistent with the other three methods. However, some variables are not significant for GA-PLS, while they were frequently selected by the other three methods, such as 770 and 772 nm. Additionally, the

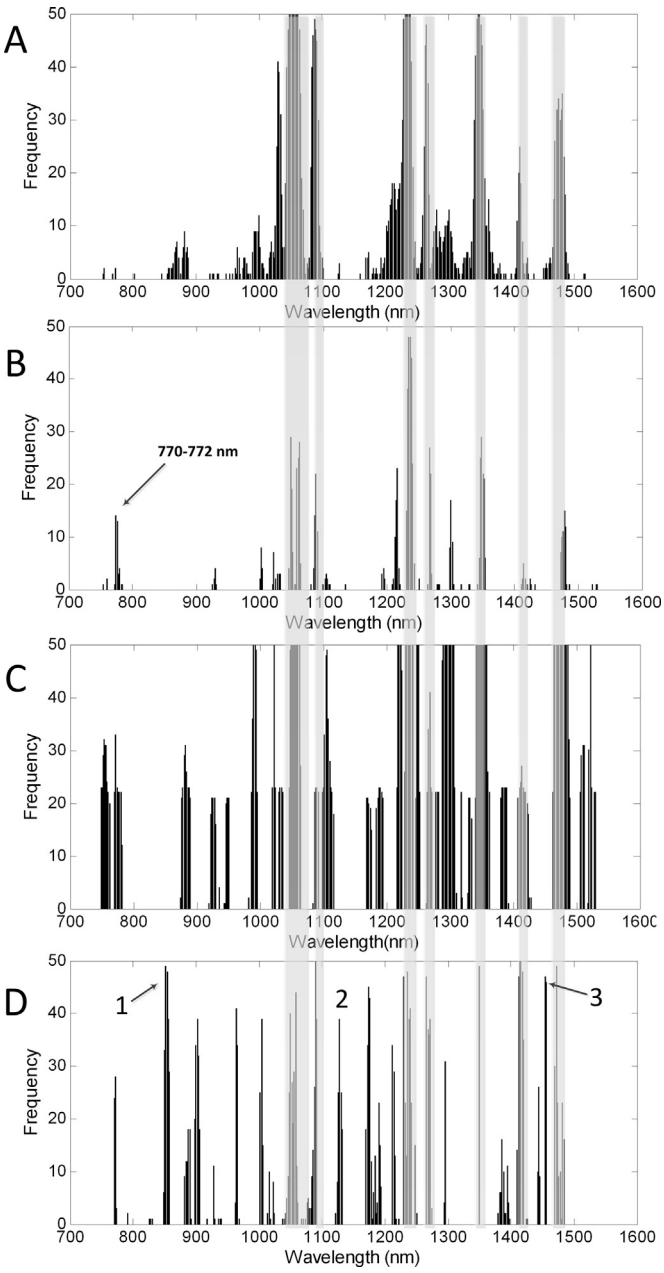


Fig. 3. The frequencies of selected variables within 50 times for the diesel fuels dataset. (A) GA-PLS; (B) CARS; (C) MC-UVE-PLS; (D) IRIV. The gray shadows in the figure represent the common selected wavelengths by the four methods. The variables, including 852–854, 1412–1418 and 1454–1456 nm wavelengths (denoted as 1–3, respectively), which were frequently selected by IRIV, were not selected by the other three methods.

variables at 852–854, 1412–1418 and 1454–1456 nm (denoted as 1–3 in Fig. 3D, respectively), which were frequently selected by IRIV, were not selected by GA-PLS, CARS and MC-UVE-PLS. Along with the good results of IRIV, it was demonstrated that these variables play an important role in the modeling. Although MC-UVE-PLS made a slightly better performance than CARS, it selected too many variables that were about 111 variables on average. Many uninformative variables may still exist like around 1522 and 1104 nm wavelengths. CARS selected the fewest variables, but it had the worst performance due to the fact that some informative variables were not selected. With regard to IRIV, the frequencies of the selected variables were overall high. Nineteen variables were selected more than 40 times. It is worth mentioning that when the

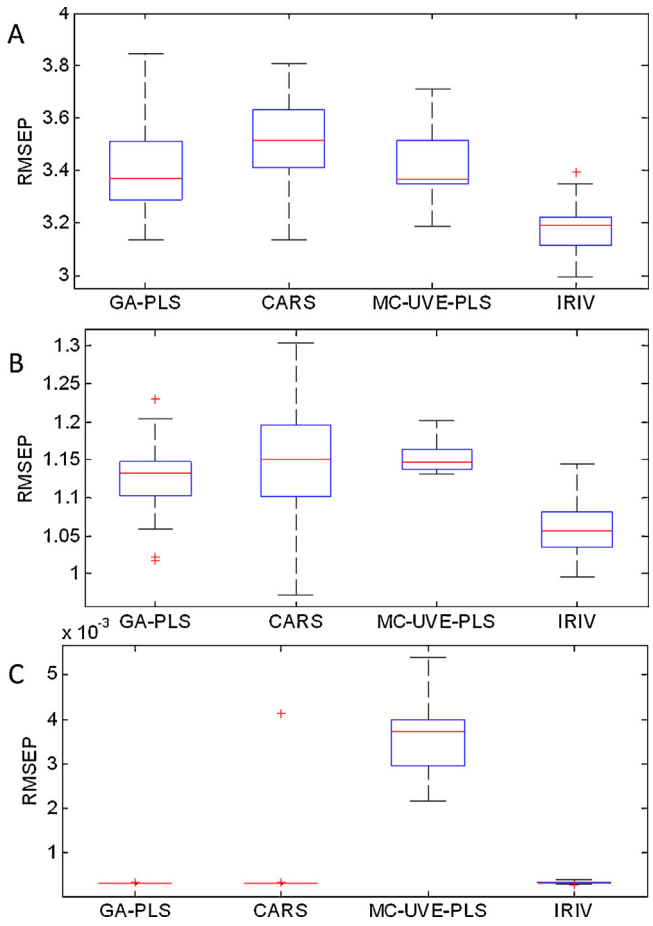


Fig. 4. The boxplot of 50 RMSEP values for the four methods. (A) diesel fuels dataset; (B) soy oil dataset; (C) corn moisture dataset. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentile, the whiskers extend to the most extreme data points are the maximum and minimum, and the "+"plotted individually represents outliers.

variables that were selected more than 40 times are used for modeling, the RMSEC and RMSEP are 2.6396 and 3.6826, respectively. Comparing this result with the one of IRIV, we observe that the joint effect between the most significant variables and other variables made contributions. Thus, we can say that IRIV can work well by using the random combination to consider the joint effect of variables.

In addition, it is also interesting to analyze the process of IRIV. IRIV is a strategy that retains both the strongly and weakly informative variables with iterative rounds. For *i*th variable, four kinds of variables could be distinguished by MPA and Eqs. (3)–(6) with the two RMSECV distribution of Φ_0 and Φ_i (inclusion and exclusion of *i*th variable, respectively). Fig. 5 depicts the differences between them. For instance, the variable 1236 nm that satisfies Eq. (3) (1236 nm, DMEAN < 0 and P value = 4.8×10^{-94}) is identified as strongly informative; 1050 nm (DMEAN < 0 and P value = 0.4332 > 0.05) is weakly informative; 1468 nm (DMEAN > 0 and P value = 0.6489 > 0.05) is uninformative; 1502 nm (DMEAN > 0 and P value = 0.0128 < 0.05) is interfering.

Fig. 6 shows the change in the number of retained variables in each round (out of 50 times) for the three datasets. As for the diesel fuels dataset, at the beginning, about half of the original variables were removed. Due to removing a large number of uninformative and interfering variables, it drops very fast in the first three rounds, and then appears mild. It converges gradually at 9th round with 72 strongly and weakly informative variables. After backward elimination, the number of the final selected variables is 42.

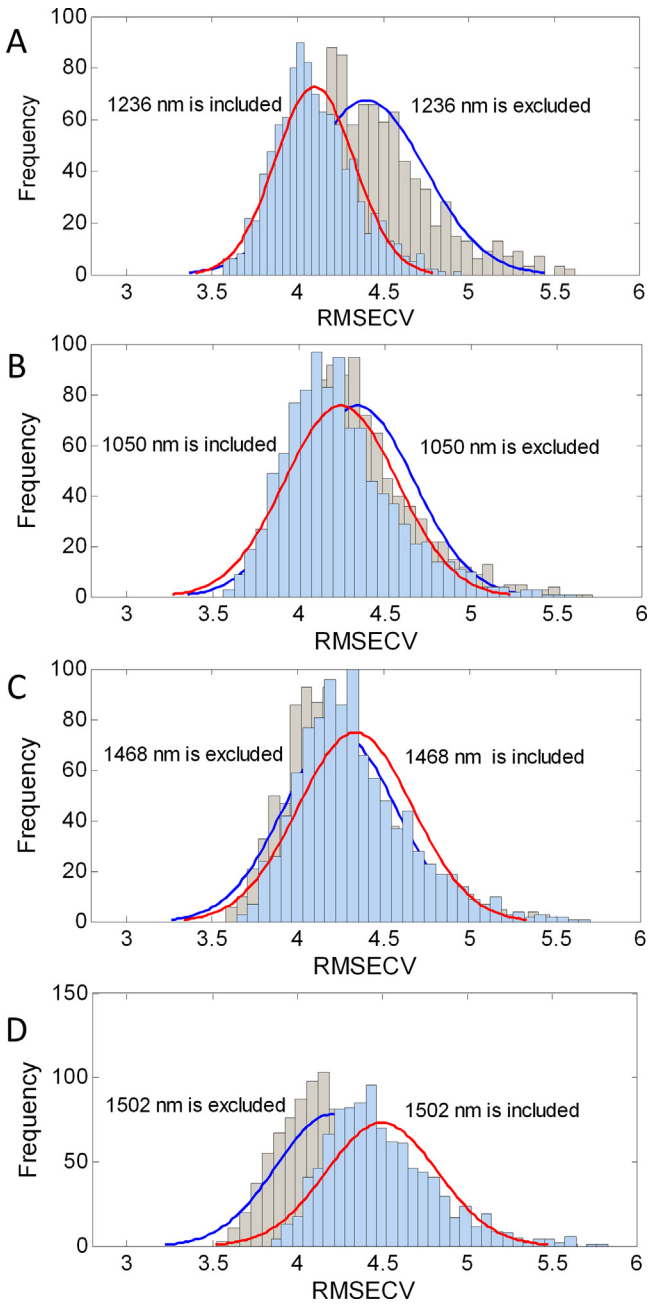


Fig. 5. The illustration of strongly informative, weakly informative, uninformative and interfering variables discriminated by MPA for the diesel fuels dataset; light blue distributions represent the inclusion of *i*th variable, while gray distributions represent the exclusion of *i*th variable. (A) strongly informative variables; (B) weakly informative variables; (C) uninformative variables; (D) interfering variables. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.3. Soy oil dataset

The statistical results obtained by conducting the four methods 50 times are reported in Table 3, including the mean and standard deviation values. Fig. 4B clearly exhibits the difference of the results by means of a boxplot. It can be seen that IRIV performs much better than the other three methods. Fig. 7 visually displays the frequencies of selected variables of the four methods revealing their differences. The variables were selected by all methods around 1144–1152, 1216–1224, 1552–1560 and 1632–1640 nm (marked by the shadow region). But as for MC-UVE, it selected

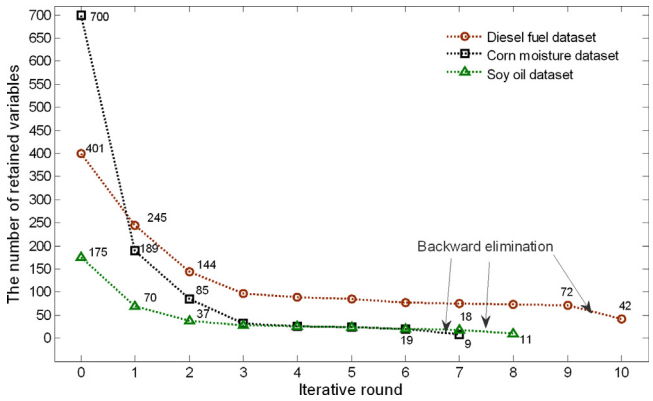


Fig. 6. The change in the number of retained informative variables in each round. The 0 value in the *X* axis represents the number of the variables in the original dataset.

Table 3
The results of different methods on the soy oil dataset.

Methods	*n*VAR	*n*LVs	RMSEC	RMSEP
PLS	175	8	0.8217	1.2252
GA-PLS	14.0 ± 5.6	4.1 ± 0.3	0.7905 ± 0.0113	1.1236 ± 0.0435
CARS	11.1 ± 4.8	5.3 ± 0.7	0.8036 ± 0.0258	1.1518 ± 0.0751
MC-UVE-PLS	36.5 ± 5.6	7.2 ± 0.9	0.8158 ± 0.0312	1.1507 ± 0.0159
IRIV	12.0 ± 2.6	4.5 ± 0.6	0.7789 ± 0.0138	1.0578 ± 0.0316

many uninformative variables like the variables 2288–2296 and 2344–2360 nm. Regarding to the variables around 1832 and 1920 nm, they were selected by IRIV more than 10 times but not by the other methods. We believe that these variables may be not very significant when singly considered, but they may have some synergetic effects with other variables improving the prediction performance. Fig. 6 shows the trend that the number of retained variables reduces in each round. For the soy oil dataset, it reduces very rapidly in the first three rounds from 175 to 37, and then it has a convergence that no uninformative and interfering variables exist.

4.4. Corn moisture dataset

As we did in the first two datasets, GA-PLS, CARS, MC-UVE-PLS and IRIV were also conducted 50 times to obtain statistical results. Table 4 and Fig. 4C clearly present the results. We can verify that only MC-UVE-PLS performed very badly as a result of selecting many uninformative variables like 1422–1428 nm. The performance of the other methods is nearly the same. The 1908 and 2108 nm wavelengths, which are corresponding to the water absorption [36] and the combination of O–H bond [37], were proved to be the key wavelengths in this dataset [7]. Fig. 8 displays the frequencies of selected variables of the four methods. From the Fig. 8, both of them were selected very frequently by all the methods. Furthermore, GA and IRIV selected them with 50 times. It can be stated that IRIV not only selected the most significant variables but also produced better prediction performance. Additionally, for the coin moisture dataset, the change in the number of retained variables in each round is also shown in Fig. 6. It is consistent with the general

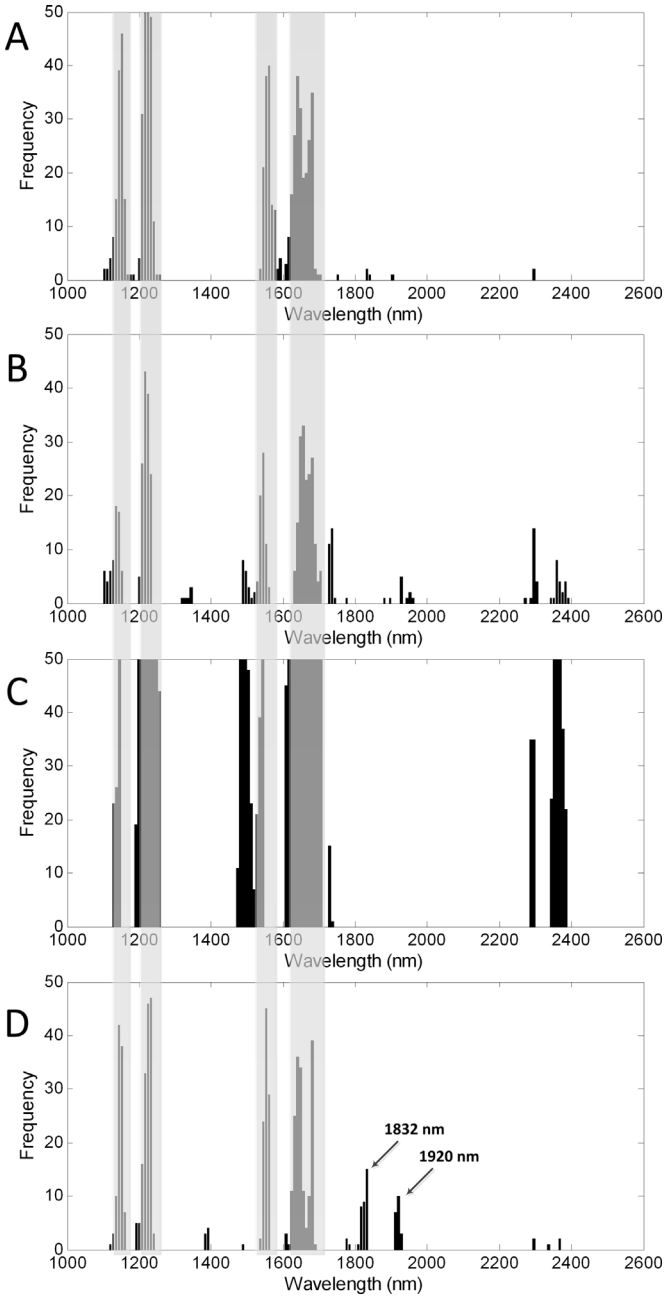


Fig. 7. The frequencies of selected variables within 50 times on the soy oil dataset. (A) GA-PLS; (B) CARS; (C) MC-UVE-PLS; (D) IRIV. The gray shadows in the figure represent the common selected wavelengths by the four methods. Around 1832 and 1920 nm wavelengths were selected by IRIV more than 10 times but not by the other methods.

convergence process of the IRIV. The number of retained variables reduces from 700 to 9, which means that a large amount of variables is not useful. Therefore, it can be said that variable selection is very important and necessary for multivariate calibration with the high dimensional data.

Table 4
The results of different methods on the corn moisture dataset.

Methods	*n*VAR	*n*LVs	RMSEC	RMSEP
PLS	700	9	0.0149	0.0201
GA-PLS	3.2 ± 0.7	3.2 ± 0.7	$2.8 \times 10^{-4} \pm 5.5 \times 10^{-7}$	$3.4 \times 10^{-4} \pm 2.7 \times 10^{-6}$
CARS	2.6 ± 1.4	2.5 ± 1.5	$3.1 \times 10^{-4} \pm 3.1 \times 10^{-4}$	$4.0 \times 10^{-4} \pm 4.6 \times 10^{-4}$
MC-UVE-PLS	52.7 ± 4.7	10.0 ± 0	$3.1 \times 10^{-3} \pm 6.9 \times 10^{-4}$	$3.6 \times 10^{-3} \pm 8.3 \times 10^{-4}$
IRIV	7.7 ± 3.6	7.0 ± 3.0	$2.6 \times 10^{-4} \pm 1.6 \times 10^{-5}$	$3.5 \times 10^{-4} \pm 2.5 \times 10^{-5}$

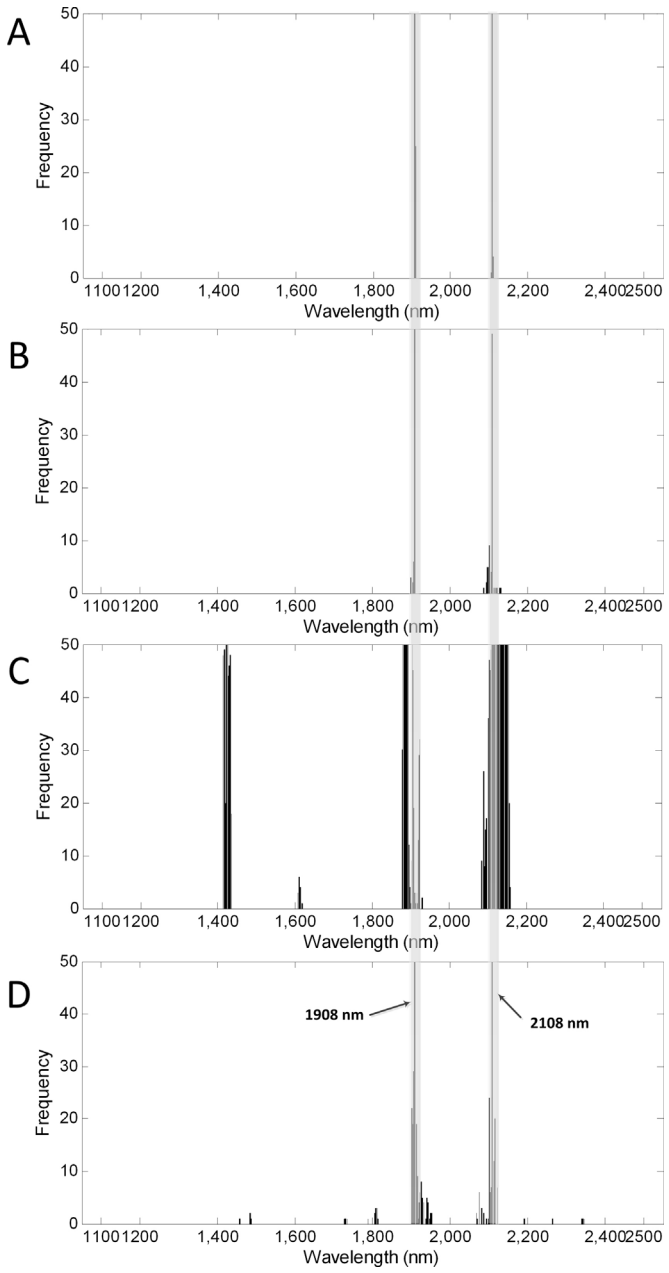


Fig. 8. The frequencies of selected variables within 50 times on the corn moisture dataset. (A) GA-PLS; (B) CARS; (C) MC-UVE-PLS; (D) IRIV. The gray shadows in the figure represent the common selected wavelengths by the four methods.

5. Conclusion

Based on the idea of Binary matrix shuffling filter (BMSF), a variable selection strategy, called Iteratively Retaining Informative Variables (IRIV), was proposed in this study. Additionally, we also divided the variables into four categories as strongly informative, weakly informative, uninformative and interfering variables. IRIV strategy considers the synergetic effect among variables through random combination. By means of this, only strongly informative and weakly informative variables are retained in each round. This is due to their positive effect under the condition of random combinations among variables. When compared with three outstanding variable selection methods, including GA-PLS, MC-UVE-PLS and CARS, IRIV was demonstrated as a better strategy with the good results. The outstanding performance of IRIV indicates that it is a good alternative of variable selection in multivariate calibration.

Although variable selection was performed by IRIV coupled with PLS in this study, it is a general strategy that can also be coupled with other regression and classification methods and applied into other fields, such as genomics, bioinformatics, metabonomics and quantitative structure–activity relationship (QSAR).

Acknowledgments

This work is financially supported by the National Nature Foundation Committee of PR China (grants nos. 21275164, 21075138, 21175157 and 11271374). The studies meet with the approval of the university's review board. Specifically, the authors would like to acknowledge Pedro Marques de Sousa in University of Algarve for his great help in the language.

References

[1] I. Guyon, A. Elisseeff, J. Mach. Learn. Res. 3 (2003) 1157.
[2] A. Lorber, B.R. Kowalski, J. Chemom. 2 (1988) 67.
[3] H. Zou, T. Hastie, J. R. Stat. Soc., Series B Stat. Methodol. 67 (2005) 301.
[4] E. Candes, T. Tao, Ann. Stat. 35 (2007) 2313.
[5] V. Centner, D.-L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Anal. Chem. 68 (1996) 3851.
[6] W. Cai, Y. Li, X. Shao, Chemom. Intell. Lab. 90 (2008) 188.
[7] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, Anal. Chim. Acta 648 (2009) 77.
[8] K. Zheng, Q. Li, J. Wang, J. Geng, P. Cao, T. Sui, X. Wang, Y. Du, Chemom. Intell. Lab. 112 (2012) 48.
[9] M.C.U. Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, Chemom. Intell. Lab. 57 (2001) 65.
[10] H.-D. Li, Q.-S. Xu, Y.-Z. Liang, Anal. Chim. Acta 740 (2012) 20.
[11] Y.-H. Yun, H.-D. Li, L.R.E. Wood, W. Fan, J.-J. Wang, D.-S. Cao, Q.-S. Xu, Y.-Z. Liang, Spectrochim. Acta. A 111 (2013) 31.
[12] H. Martens, T. Naes (Eds.), Multivariate Calibration, Wiley, New York, NY, 1989.
[13] F.G. Blanchet, P. Legendre, D. Borcard, Ecology 89 (2008) 2623.
[14] J.M. Sutter, J.H. Kalivas, Microchem. J. 47 (1993) 60.
[15] A.S. Bangalore, R.E. Shaffer, G.W. Small, M.A. Arnold, Anal. Chem. 68 (1996) 4200.
[16] J. Yang, V. Honavar, IEEE Intell. Syst. 13 (1998) 44.
[17] M. Arakawa, Y. Yamashita, K. Funatsu, J. Chemom. 25 (2011) 10.
[18] J. Ghasemi, A. Niazi, R. Leardi, Talanta 59 (2003) 311.
[19] C.B. Lucasius, M.L.M. Beckers, G. Kateman, Anal. Chim. Acta 286 (1994) 135.
[20] K. Sasaki, S. Kawata, S. Minami, Appl. Spectrosc. 40 (1986) 185.
[21] R. Leardi, J. Chemom. 14 (2000) 643.
[22] J.H. Kalivas, N. Roberts, J.M. Sutter, Anal. Chem. 61 (1989) 2024.
[23] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, IEEE/ACM Trans. Comput. Bioinf. 8 (2011) 1633.
[24] H. Zhang, H. Wang, Z. Dai, M.-s. Chen, Z. Yuan, BMC Bioinf. 13 (2012) 1.
[25] R. Diaz-Uriarte, S. Alvarez de Andres, BMC Bioinf. 7 (2006) 3.
[26] Q. Wang, H.-D. Li, Q.-S. Xu, Y.-Z. Liang, Analyst 136 (2011) 1456.
[27] H.-D. Li, Y.-Z. Liang, D.-S. Cao, Q.-S. Xu, TrAC, Trends Anal. Chem. 38 (2012) 154.
[28] H.-D. Li, M.-M. Zeng, B.-B. Tan, Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, Metabolomics 6 (2010) 353.
[29] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. 58 (2001) 109.
[30] M. Goodarzi, Y.V. Heyden, S. Funar-Timofei, TrAC, Trends Anal. Chem. 42 (2013) 49.
[31] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, J. Chemom. 24 (2010) 418.
[32] H.B. Mann, D.R. Whitney, Ann. Math. Stat. 18 (1947) 50.
[33] M. Forina, G. Drava, C. Armanino, R. Boggia, S. Lanteri, R. Leardi, P. Corti, P. Conti, R. Giangiacomo, C. Galliena, R. Bigoni, I. Quartari, C. Serra, D. Ferri, O. Leoni, L. Lazzeri, Chemom. Intell. Lab. 27 (1995) 189.
[34] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137.
[35] R. Leardi, A. Lupiáñez González, Chemom. Intell. Lab. 41 (1998) 195.
[36] D. Jouan-Rimbaud, D.-L. Massart, R. Leardi, O.E. De Noord, Anal. Chem. 67 (1995) 4295.
[37] J.-H. Jiang, R.J. Berry, H.W. Siesler, Y. Ozaki, Anal. Chem. 74 (2002) 3555.