Chemometrics and Intelligent Laboratory Systems 151 (2016) 89–94

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Short Communication

Improvement on enhanced Monte-Carlo outlier detection method

Liangxiao Zhang [a,d,e,f,*,1], Du Wang [a,e,1], Rongrong Gao [g,1], Peiwu Li [a,c,d,e,*], Wen Zhang [a,b,e,*], Jin Mao [a,d], Li Yu [a,e], Xiaoxia Ding [a,d], Qi Zhang [a,c]

[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] Key laboratory of Detection for Mycotoxins, Ministry of Agriculture, Wuhan 430062, China
[d] Laboratory of Risk Assessment for Oilseeds Products (Wuhan), 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
[g] School of Life Science and Technology, Tongji University, Shanghai, China

ARTICLE INFO

Article history:
Received 29 September 2015
Received in revised form 22 November 2015
Accepted 10 December 2015
Available online 19 December 2015

Keywords:
Outlier detection
Monte-Carlo sampling
Partial least squares regression
Multivariate calibration

ABSTRACT

Highly predictive multivariate calibration model depends on samples in training set. In this study, we introduced an outlier detection method and developed its improvement for shorter run time. Improved Monte-Carlo outlier detection (IMCOD) was proposed to establish cross-prediction models for determining normal samples, which were subsequently used to analyze the distribution of prediction errors for all of dubious samples together. Four real datasets were employed to illustrate and validate the performance of IMCOD. After sample selection for training set of NIR of soy flour samples, the Root Mean Square Error of Prediction (RMSEP) of PLS model decreased from 1.4811 to 0.7650. This method benefits the establishment of a good model for QSAR and NIR datasets.

1. Introduction

The performance of multivariate calibration model depends on samples in the training set. Due to the recording mistakes or influence from exceptional circumstances, some spectra might be different from the majority when analyzing real samples. Sample selection is therefore an important step to identify and subsequently eliminate atypical observations from the training set [1]. For multivariate modeling, the outlier detection methods contain statistical and model based methods. Statistical methods were designed according to the distribution in high dimensional sample space to detect the observations relatively far from the center of the data distribution [2]. Multivariate location and covariance estimation were calculated by data matrix (X) such as Mahalanobis distance (MD), Minimum Covariance Determinant (MCD) [3], Minimum Volume Ellipsoid (MVE) [4], ellipsoidal multivariate trimming (MVT) [5], resampling by half-means (RHM) and smallest half volume (SHV) [6], S-estimators [7], CM-estimators [8], τ-estimators [9], MM-estimators [10], estimators based on multivariate ranks or signs [11], and depth-based estimators [12]. The key to these methods is to find out the main body of an observation matrix and identify the outliers significantly different from the majority of the training set [13].

Model based methods analyzed the distribution in high dimensional model space and negative influence from the outliers significantly different from the majority of the training set. As a classic model based method, Monte-Carlo outlier detection (MCOD) method was proposed to detect three kinds of outliers by establishing many cross-prediction models [14–15]. In MCOD, the dataset was randomly divided into training and testing sets, which were used to establish and validate predictive model, respectively. Since the majority of training set were normal samples, the *X* outlier far from the center of the sample space are considerably variable by Monte-Carlo sampling subset predictive models while predicting the *y* outlier is usually difficult [13–14]. In this case, the distribution, mean value and standard deviation of predictive errors could be employed to detect outlier. However, multiple outliers distort measures of central location and dispersion of models or samples, making the inaccurate results were obtained when there are multiple outliers in the data. This phenomenon is termed the masking effect. To overcome the masking effect and obtain the clear boundary between normal and abnormal samples, we proposed a new strategy, termed as enhanced MCOD (EMCOD), to detect outliers using MCOD to firstly determine normal samples and then individually identify the dubious samples [13]. After validation by one simulated and three real datasets, the results indicated that EMCOD could effectively detect

* Corresponding authors at: Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, Wuhan 430062, China. Tel.: +86 27 86812943; fax: +86 27 86812862.

E-mail addresses: liangxiao_zhang@hotmail.com (L. Zhang), peiwuli@oilcrops.cn (P. Li), zhangwen@oilcrops.cn (W. Zhang).

[1] These authors contributed equally to this study.

http://dx.doi.org/10.1016/j.chemolab.2015.12.006
0169-7439/

outliers and improve the performance of predictive models. However, the run time of EMCOD significantly increases for big data such as NIR data. Sometimes, the run time might need several hours, which is not acceptable for analysts. Therefore, it is necessary to develop more advanced method of outlier detection.

The aim of this study is (1) to improve the EMCOD method by taking the dubious samples as testing set in the model population, and (2) to introduce improved Monte-Carlo outlier detection (IMCOD) method to detect the potential outliers for establishing a highly predictive multivariate model.

2. Theories and methods

2.1 Datasets

Four public available datasets were used to illustrate and validate our method.

Dataset 1: stack loss plant dataset for oxidation of ammonia to nitric acid, provides operational data of a plant, which includes 21 observations on three independent variables (cooling air flow, cooling water inlet temperature and acid concentration) and a dependent variable of stack loss [16–17]. Among all the samples, the outliers are Nos. 1, 3, 4 and 21, and No. 2 is a good leverage point.

Dataset 2: Hawkins–Bradu–Kass data, is another classic dataset for outlier detection and robust regression. The first 14 observations out of 75 are outliers of this dataset [18].

Dataset 3: Oil contents of soybean [19]. This dataset consists of 54 soy flour samples measured on NIR spectrometers. The spectra were recorded at 175 wavelength channels from 1104 to 2496 nm with an interval of 8 nm. Oil content values determined by Soxhlet extraction were used as responses. The oil contents of the first four samples were deliberately changed to make them as outlier.

Dataset 4: Boiling point of diesel fuels. The dataset includes 246 samples, which were measured at 401 wavelength channels from 750 to 1550 nm with 2 nm intervals. The boiling point was measured at 50% recovery (BP50). This dataset is freely available at http://www.eigenvector.com/data/SWRI/index.html.

2.2 Enhanced Monte-Carlo outlier detection

Drawing inspiration from ensemble learning, Monte-Carlo outlier detection (MCOD) method was proposed by multiple learning to obtain the distribution of prediction errors of each sample [14–15]. The detailed algorithm was described elsewhere [14]. Herein, the main procedures of this method were introduced as follows: (1) determine number of principal components by cross-validation; (2) divide randomly the whole dataset into training and validation sets, which were used to build a prediction model and obtain the prediction errors for samples in test set; (3) repeat (2) for N times; (4) mean value (MV) and standard deviation (STD) of the prediction errors were employed to diagnose outliers. Generally, the predictive errors of a *y* outlier have a large mean value, while an *X* outlier (good leverage point) possesses a small mean value of predictive residuals but a large STD. Intuitively, the MV/STD plot provides visual diagnosis for outlier [14–15].

To overcome the masking effect, an enhanced Monte-Carlo outlier detection (EMCOD) method was recently proposed [13]. Due to the masking effect, the boundary of outliers and normal samples was hard to be determined. However, the samples with the smallest MV and STD of prediction errors are easily determined to be normal samples. Therefore, the core idea of EMCOD is to use the strategy of 'Let One In' to diagnose the dubious samples one by one. Firstly, using the MV/STD plot in MCOD, the samples with the smallest MV and STD of prediction errors were selected as determinate normal samples and determine the remaining as dubious samples. Each dubious sample was extracted to form new dataset with determinate normal samples. Then, the MCOD was conducted to obtain the predictive errors of the current dubious sample. Finally, MV and STD of the predictive errors of the dubious samples were used to diagnose outliers. The detailed algorithm was described elsewhere [13].

2.3 Improved Monte-Carlo outlier detection

To overcome the masking effect, the dubious samples were checked by EMCOD method one by one. However, along with increase of samples, the run time of EMCOD significantly increases. Especially, for NIR data of more than 200 samples, the run time of EMCOD sometimes researches to several hours, which is not acceptable for outlier detection. Therefore, improved Monte-Carlo outlier detection (IMCOD) was conducted in this study. As shown in Fig. 1, the procedures of IMCOD contain the following steps: (1) As the same as EMCOD, the samples with the smallest MV and STD of prediction errors were selected as determinate normal samples (Ns) and determine the remaining as dubious samples (Ds); (3) randomly divide Ns into training and test sets; (4) build the multivariate model to predict the samples in the test set and Ds to obtain the prediction errors; (5) after N cycles, MV and STD of predictive errors on the dubious samples and also normal samples were used to diagnose outliers. In theory, EMCOD needs to conduct MCOD procedures once for all samples and Ds times for the data of Ns + 1 samples; while IMCOD runs MCOD procedures once for all samples and once for the data of Ns determinate normal samples. Obviously, compared with EMCOD, the run time of IMCOD significantly decreases to less than 2 times of MCOD procedures once for all samples.

2.4 Data processing and analysis

All programs used were coded in MATLAB 2015a for Windows and all calculations were carried out on a personal computer.

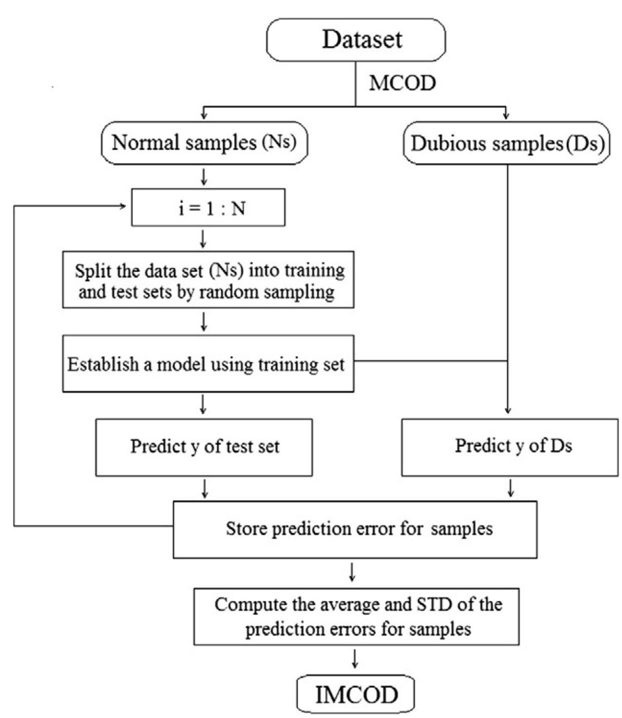

Fig. 1. Flow chart for improved Monte-Carlo outlier detection.

3. Results and discussion

3.1 Improvement for EMCOD

Outlier detection is important to establish a high-performance model. MOCD was recently proposed to detect outliers by establishing many predictive models and analyzing a MV/STD plot of prediction errors. EMCOD was developed a strategy of 'Let One In' to diagnose the dubious samples one by one for obtaining the visualized boundary between normal and abnormal samples. Herein, we develop improved Monte-Carlo outlier detection for shorter run time. To illustrate our method, Dataset 1 and Dataset 2 were used.

Dataset 1 is the stack loss dataset of a plant. In MCOD, the number (N) of Monte-Carlo models and sampling ratio are set to 10,000 and 0.8, respectively. The MV/STD plot of the prediction errors for 21 samples was shown in Ref. [13]. To obtain a clearer result using relatively short run time, IMOCD was proposed and employed to detect outliers in this dataset. As shown in Fig. 2A, the samples including 20, 5, 16, 18, 19, 13, 14, 8, 15, 10, and 17 were normal samples (green square), which had the smallest mean and STD values. We established MC prediction models using these 11 samples and used these models to observe other samples. The number (N) of Monte-Carlo models and sampling ratio are also set to 10,000 and 0.8, respectively. According to the hypothesis that the models built with merely normal samples provide lower prediction errors for normal samples but higher prediction errors for outliers, the distances between normal samples and outliers should be longer. Then, whether selection of the determinate normal samples influences outlier detection was investigated. For

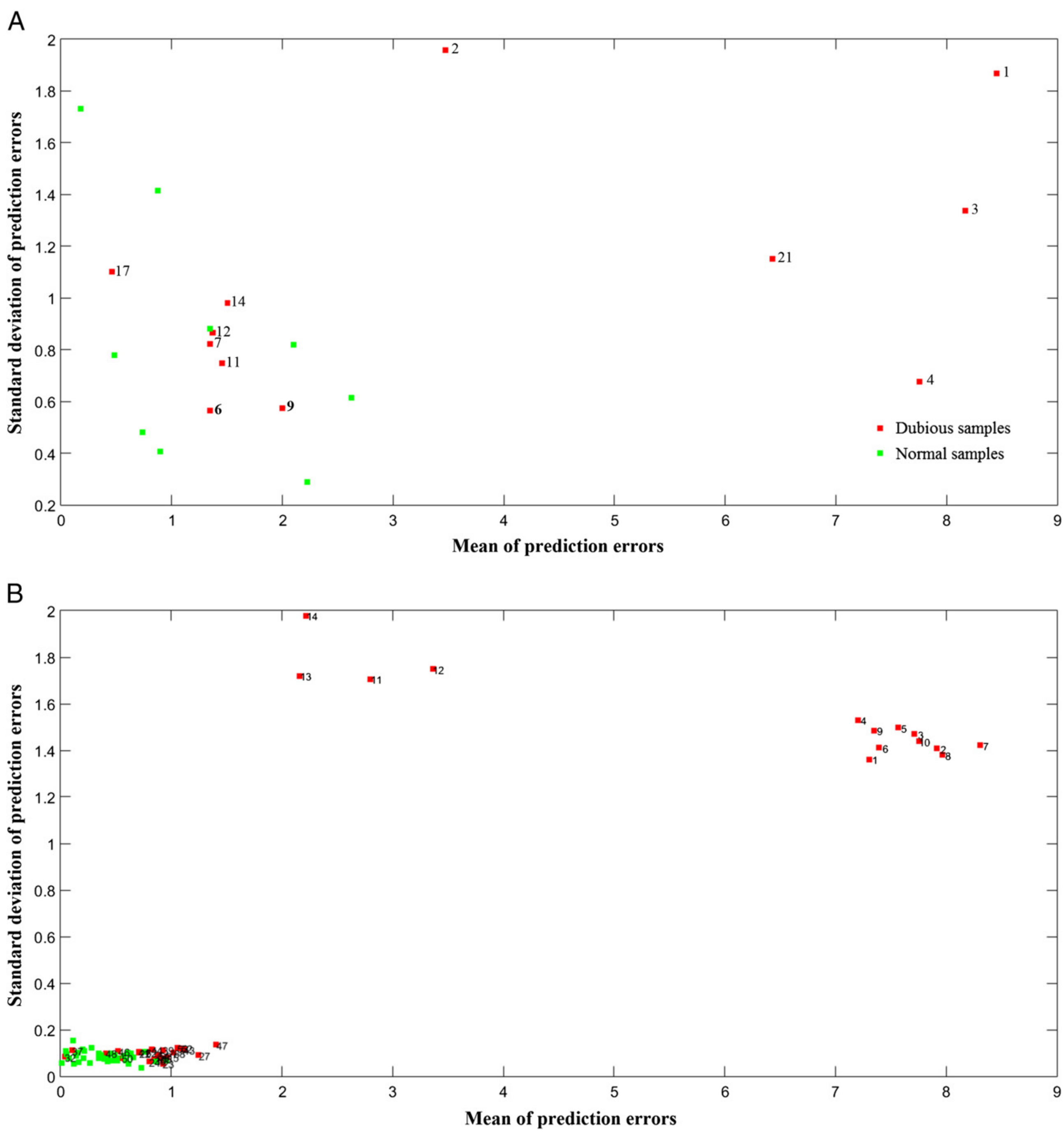

Fig. 2. Improved Monte-Carlo outlier detection mean/standard deviation plot of prediction errors for (A) Dataset 1 and (B) Dataset 2.

Dataset 1, the threshold of average of prediction errors was set to 3.0, while the one of STD value was 1.2. From Fig. S1, whatever the threshold was set, normal samples appear at the same region of determinate normal samples in the MV/STD plot, even though they are regarded as dubious samples. The results indicate that the threshold does not influence the outlier detection. The result is shown in Fig. 2A, which illustrates that IMCOD has a better result since the outliers have correctly been detected.

Dataset 2 represents the Hawkins–Bradu–Kass data. As shown on the right of Fig. 2b, the M/SD plot indicates that 14 samples (Nos. 1–14) are outliers. 52 samples with the lowest standard deviations of prediction errors (<0.5) were selected as normal samples. Other 23 samples were then detected and tested one by one by the MC prediction models of the dataset established with the 52 samples. As shown on the left hand of Fig. 2B, the prediction errors of 9 normal samples decrease and the prediction errors of 14 outliers greatly increase. The distances between normal samples and outliers significantly increase.

Compared with EMCOD, all dubious samples could be predicted together but not one by one. Therefore, the run time significantly decreases.

3.2 Method validation

To validate our method, Dataset 3 was used, which consisted of 54 soy flour samples measured on NIR spectrometers. The oil contents of

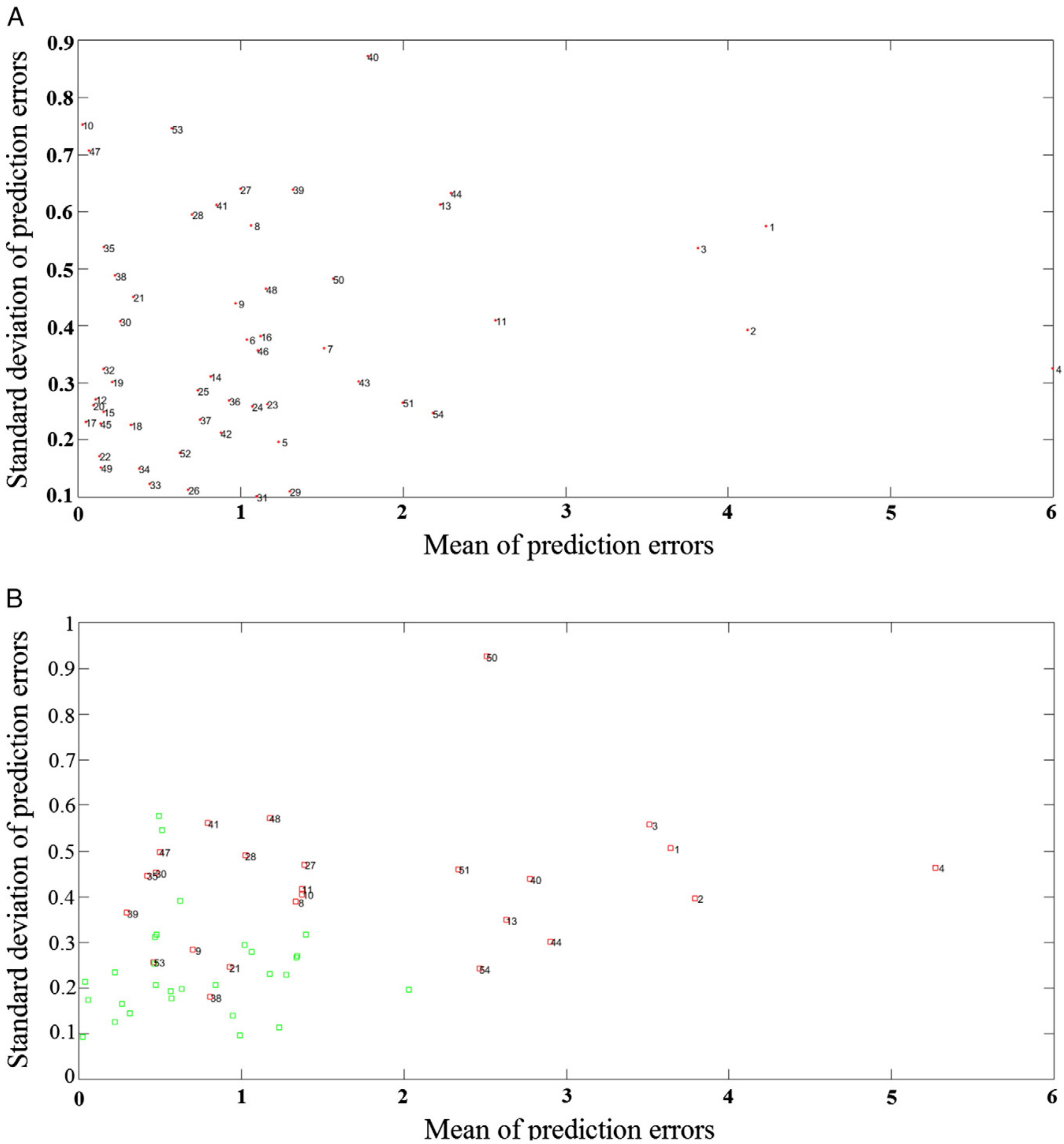

Fig. 3. Mean/standard deviation plot of prediction errors for Dataset 3: (A) Monte-Carlo outlier detection (left) and (B) improved Monte-Carlo outlier detection.

first 4 samples were deliberately changed to make them as outlier. MCOD was initially conducted to detect the outliers.

As shown in Fig. 3A, 4 outliers have a clear tendency to separate from the majority of training set. Theoretically, the y outliers have large prediction errors, while X outliers (good leverage point) have large STD values [13–15]. In IMCOD, the MVs and STDs of prediction errors were used to determine 38 samples with the smallest MVs (<2.0) and STDs (<0.6). When the number (N) of Monte-Carlo models and sampling ratio are respectively set to 5000 and 0.8, the MVs and STDs of prediction errors could be used to diagnose outliers. From Fig. 3B, except the first 4 samples, 6 samples (Nos. 13, 40, 44, 50, 51 and 54) were also separated from the majority of training set. The PLS models built by all samples were compared with those built by normal samples. The 5 fold cross validation was used to evaluate the performance of PLS models. The results showed that, when the first 4 samples were removed, the Root Mean Square Error of Prediction (RMSEP) decreased from 1.4811 to 0.8397. Obviously, after outliers were removed, the accuracy of the model significantly improved.

Dataset 4 is boiling point of diesel fuels. In MCOD, the number (N) of Monte-Carlo models and sampling ratio is set to 5000 and 0.8, respectively. 157 samples with the smallest MVs (<10.0) and STDs (<1.5) were selected and employed to diagnose potential outliers. We established MC prediction models using these 157 samples and used these models to predict other samples. The number (N) of Monte-Carlo models and sampling ratio are also set to 5 000 and 0.8, respectively. The MV/STD plot of the prediction errors for 246 samples

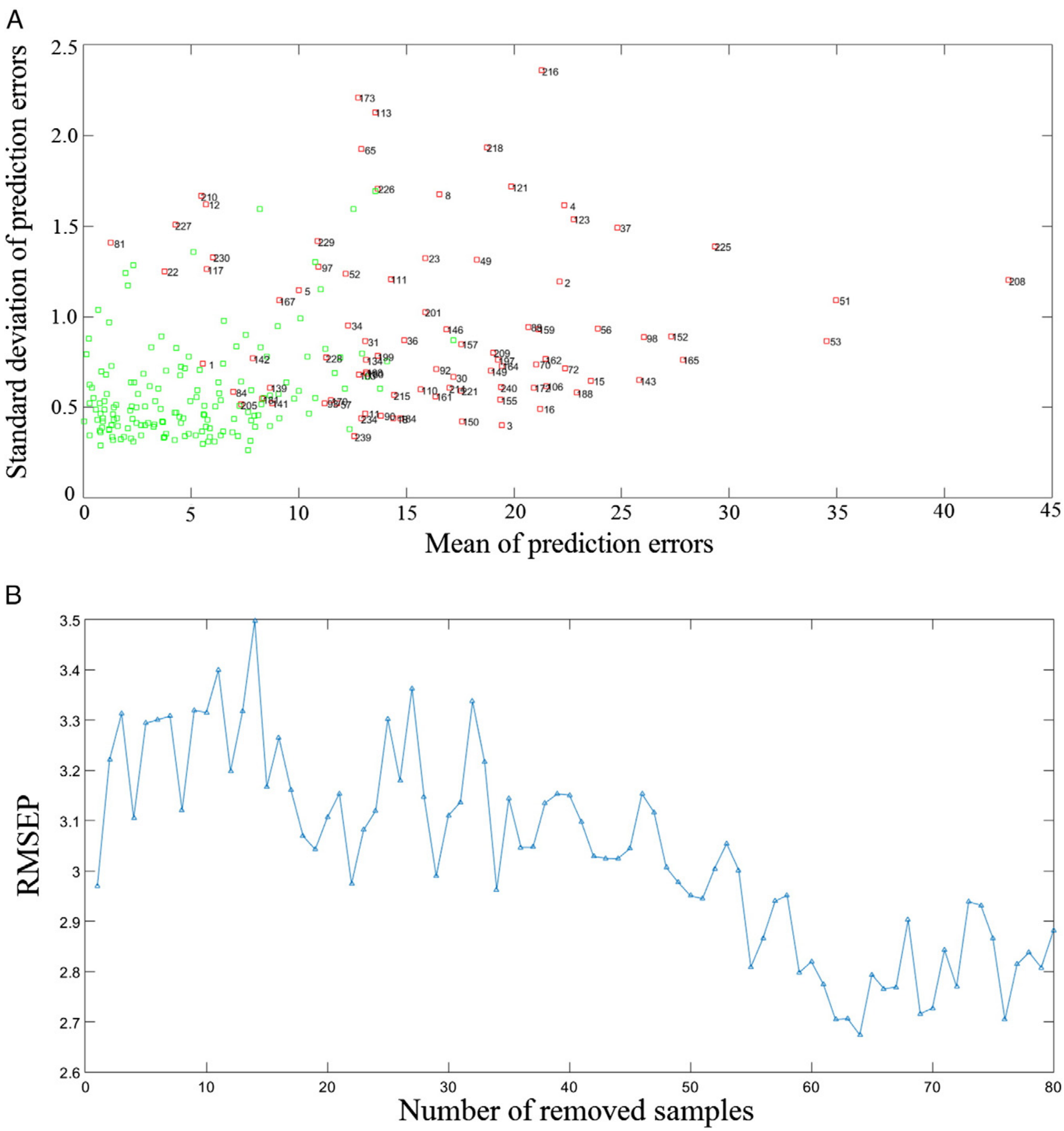

Fig. 4. (A) Mean/standard deviation plot of prediction errors for Dataset 4: improved Monte-Carlo outlier detection; (B) RMSEP of multivariate model after removing the different number of samples.

was shown on Fig. 4. As shown in Fig. 4A, since the number of samples is large, it is still hard to obtain clear separation between normal samples and potential outliers. In this case, we investigate the number of removed samples to RMSEP of the multivariate model after potential samples were removed. According to the mean value of predictive errors, one more potential outlier was removed in each cycle. Monte-Carlo cross validation was used to evaluate the multivariate model [20]. As shown in Fig. 4B, when the number of removed samples reaches to 64, the lowest RMSEP was achieved.

Compared with EMCOD, IMCOD just needs 6.4 min, which is significantly less than EMCOD of more than 3 h. With the help of IMCOD, 182 samples with the smallest MV and STD of prediction errors could be selected and employed to build PLS model for boiling point of diesel fuels. The Monte-Carlo cross validation indicates that the RMSEP decreases to 2.674 and Q^2 increases to 0.968.

4. Conclusion

In this study, we improved EMCOD method for short run time. In IMCOD, normal samples were employed to build multiple multivariate models, while the dubious samples were taken as test set. Therefore, since the MCOD was just conducted to the whole dataset and normal samples, the run time of IMCOD significantly decreases. Moreover, IMCOD is not susceptible to the threshold for selecting the determinate normal samples. Four datasets were employed to illustrate and validate our method. The results indicated that IMCOD could save computation time and improve the performance of multivariate model by diagnosing and removing outliers.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.chemolab.2015.12.006.

Conflict of interests

No authors declared any potential conflicts of interest.

Acknowledgments

This work was supported by the Project of National Science & Technology Pillar Plan (2012BAK08B03), the National Major Project for Agro-product Quality & Safety Risk Assessment (GJFP2015007), the National Natural Science Foundation of China (21205118), and the earmarked fund for China Agriculture research system (CARS-13).

References

[1] Å. Rinnan, F. van den Berg, S.B. Engelsen, Review of the most common preprocessing techniques for near-infrared spectra, TrAC-Trend. Anal. Chem. 28 (2009) 1201–1222.
[2] E.M. Knorr, R.T. Ng, V. Tucakov, Distance-based outliers: algorithms and applications, VLDB J. 8 (2000) 237–253.
[3] W.J. Egan, S.L. Morgan, Outlier detection in multivariate analytical chemical data, Anal. Chem. 79 (1998) 2372–2379.
[4] R. Gnanadesikan, J.R. Kettenring, S. Maloor, Better alternatives to current methods of scaling and weighting data for cluster analysis, J. Stat. Plan. Infer. 137 (2007) 3483–3496.
[5] D.M. Rocke, D.L. Woodruff, Identification of outliers in multivariate data, J. Am. Stat. Assoc. 91 (1996) 1047–1061.
[6] P.J. Rousseeuw, V.D. Katrien, A fast algorithm for the minimum covariance determinant estimator, Technometrics 41 (1999) 212–223.
[7] L. Davies, Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices, Ann. Stat. 15 (1987) 1269–1292.
[8] J.T. Kent, D.E. Tyler, Constrained M-estimation for multivariate location and scatter, Ann. Stat. 24 (1996) 1346–1370.
[9] H.P. Lopuhaä, Multivariate τ-estimators for location and scatter, Can. J. Stat. 19 (1991) 307–321.
[10] K.S. Tatsuoka, D.E. Tyler, On the uniqueness of S-functionals and M-functionals under nonelliptical distributions, Ann. Stat. 28 (2000) 1219–1243.
[11] S. Visuri, V. Koivunen, H. Oja, Sign and rank covariance matrices, J. Stat. Plan. Infer. 91 (2000) 557–575.
[12] D.L. Donoho, M. Gasko, Breakdown properties of location estimates based on halfspace depth and projected outlyingness, Ann. Stat. 20 (1992) 1803–1827.
[13] L. Zhang, P. Li, J. Mao, F. Ma, X. Ding, Q. Zhang, An enhanced Monte–Carlo outlier detection method, J. Comput. Chem. 36 (2015) 1902–1906.
[14] D.S. Cao, Y.Z. Liang, Q.S. Xu, H.D. Li, X. Chen, A new strategy of outlier detection for QSAR/QSPR, J. Comput. Chem. 31 (2010) 592–602.
[15] H.D. Li, Y.Z. Liang, Q.S. Xu, D.S. Cao, Model-population analysis and its applications in chemical and biological modeling, TrAC-Trend. Anal. Chem. 38 (2012) 154–162.
[16] K.A. Brownlee, Statistical Theory, Methodology in Science and Engineering, Academic, New York, 1965 pp. 491-500.
[17] R.A. Becker, J.M. Chambers, A.R. Wilks, The New S LanguageWadsworth & Brooks/Cole 1988.
[18] D. Bradu, D.M. Hawkins, An anscombe type robust regression statistic, Comput. Stat. Data Anal. 20 (1995) 355–386.
[19] 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, Transfer of calibration function in near-infrared spectroscopy, Chemometr. Intell. Lab. Syst. 27 (1995) 189–203.
[20] Q.S. Xu, Y.Z. Liang, Monte Carlo cross validation, Chemometr. Intell. Lab. Syst. 56 (2001) 1–11.