Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 219 (2019) 274–280

Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

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

A new strategy of applying modeling indicator determined method to high-level fusion for quantitative analysis

Qianqian Li [a,c], Gaowei Li [b], Jixiong Zhang [c], Hong Yan [c], Wei Liu [d], Shungeng Min [c,*]

[a] School of Marine Sciences, China University of Geosciences, Beijing 100083, China
[b] Beijing Haiguang Instrument Co., Ltd., Beijing 100015, China
[c] College of Science, China Agricultural University, Beijing 100193, China
[d] Chongqing Grain and Oil Quality Supervision and Inspection Station, Chongqing 400026, China

ARTICLE INFO

Article history:
Received 5 February 2019
Received in revised form 11 April 2019
Accepted 13 April 2019
Available online 15 April 2019

Keywords:
Modeling indicator determined method
Root mean square error of prediction weighted
Ratio performance deviation weighted
High-level fusion
Quantitative analysis

ABSTRACT

A novel method, named as modeling indicator determined (MID) method, based on two model evaluation parameters i.e., root mean square error of prediction (RMSEP) and ratio performance deviation (RPD), is proposed to employ high-level fusion for quantitative analysis. The two MID methods of root mean square error of prediction weighted (RMSEPW) method and ratio performance deviation weighted (RPDW) method are put forward on the basis of the model evaluation indicators from the individual models. Performance of RMSEPW method and RPDW method are evaluated in terms of the predictive ability of root mean square error of prediction for fusion ($RMSEP_f$) through the fused models. The two MID methods are applied to UV–visible (UV–vis), near infrared (NIR) and mid-infrared (MIR) spectral data of active ingredient in pesticide, and gas chromatography-mass spectrometer (GC–MS) and NIR spectral data of *n*-heptane in chemical complex for high-level fusion. Moreover, the results are compared with the individual methods. As a result, the overall results show that the two MID methods are promising with significant improvement of predictive performance for high-level fusion when executing quantitative analysis.

1. Introduction

Data fusion [1,2], as an emerging branch in chemometrics, is a procedure of integrating data obtained from different analytical techniques into a single global model. It has become a vital important tool to combine the data of different analytical techniques, and makes a comprehensive use of these data for qualitative and quantitative analysis. Data fusion techniques are commonly fallen into three categories [3,4]: low-level data fusion, mid-level data fusion and high-level data fusion. Low-level data fusion is a cohesion of the original data matrix from different sensors, however, the final array contains a great number of variables, which in return takes a long time for analysis. Mid-level data fusion is to concatenate the features extracted from individual sensors by variable selection methods. High-level data fusion is executed to merge the results of each sensor and thus produce a final response via the individual results.

High-level data fusion is an approach to combine the predictive results from two or more models with the consideration of all the individual methods [5]. As far as we know, the high-level fusion is mainly focused on classification issues [6–12] and rare research is focused on quantitative analysis. There are multiple approaches for data fusion in classification issues such as majority vote [13], naive Bayes approach [14], Dempstere-Shafer's [15] method and so on. Generally, the classification results of the fusion method might perform better than individual models. As has been stated, high-level fusion is performed to produce the results by combining all individual sensors. Therefore, it is essential to assign each sensor a weight according to their own contributions. Under this way, the results of individual methods are fused by different weighted coefficients, and high-level fusion is employed to execute quantitative analysis with the modeling indicator determined (MID) method, which is relying on the original models of partial least squares (PLS). PLS regression has been extensively used for developing models owing to its outstanding ability in overcoming deviations caused by effects such as spectral bands overlapping and components interacting [16]. PLS regression is a commonly used multivariate method by decomposing the spectral and concentration arrays simultaneously so as to make the model suitable to extract the maximum information from the spectra [17,18]. Root mean square error of prediction (RMSEP) and ratio performance deviation (RPD) are two model evaluation indicators derived from PLS models. Specifically, RMSEP [19,20] is a parameter that utilized to evaluate the predictive ability of the model, whereas RPD [21,22] is an indicator for evaluating the established model. As a result, two MID methods of root mean square error of prediction weighted method (RMSEPW) and ratio performance deviation weighted method (RPDW) are proposed to employ high-level fusion

* Corresponding author.
E-mail address: minsg@cau.edu.cn (S. Min).

https://doi.org/10.1016/j.saa.2019.04.022
1386-1425/

for quantitative analysis. The weight of each method is calculated reasonably and scientifically by the proposed quantitative analysis methods. On these circumstances, the comprehensive features of each model are fully taken advantage by the proposed quantitative analysis methods. Since the original individual models have direct impacts on the corresponding methods, the results of MID methods are depended on the effectiveness of the existed models. When one method is assigned to a high weight, which indicating the corresponding method is playing a vital role for the fusion approach.

In views of the above, a weighted procedure is applied to fuse each sensor by the proposed method which is summarized as follows. Firstly, the average outputs of each method are obtained through PLS algorithm after fifty times of Monte-Carlo (MC) sampling approach [23]. Secondly, each method is assigned a weight according to its own regression model. Finally, the ensemble outputs of all the methods are obtained with each method offering its own contributions to the fusion approach. Theoretically, the fusion models give better performance than individual models.

In fusion schemes, the feasibility of the two MID methods are investigated by two cases study of pesticide data set and chemical data set. In order to prove the widespread applicability of MID method, different systems formed with different analytical methods are used to execute high-level fusion for quantitative analysis. Therefore, the feasibility of applying MID method to high-level fusion are verified by various techniques, that is, UV–vis, NIR and MIR spectral data of active ingredient in pesticide, GC–MS and NIR spectral data of *n*-heptane in chemical complex. The Performance of RMSEPW method and RPDW method are evaluated in terms of the predictive ability of the models characterized by root mean square error of prediction for fusion ($RMSEP_f$), and the results were compared with the individual methods as well.

2. Theory and algorithms

2.1. High-level fusion

In order to obtain more competitive results of high-level data fusion, the model responses are combined to produce a fused decision with each individual sensor. In the fusion approach, the results are acquired by combining the outcome of various methods via their weights. The outputs can be represented as degrees of support for a sensor through its own weight in Eq. 1

$$y_p = \sum_{i=1}^{L} y_i w_i \tag{1}$$

where L is the number of sensors, w_i is the weight for the ith sensor, y_i is the predict result of the ith sensor, and y_p is the fusion result. The output y_p is an ensemble of all individual results and their own weights.

2.2. The framework of modeling indicator determined (MID) method

The weights are assigned to each sensor according to the MID method. In this study, two model evaluation indexes i.e., root mean square error of prediction weighted (RMSEPW) method and ratio performance deviation weighted (RPDW) method are applied for quantitative analysis. The flowchart of the high-level fusion for quantitative analysis is represented in Fig. 1. As shown in the schematization of the fusion framework, the high-level approach is assembled by different weights via the two MID methods. Some individual methods are given larger weights whereas others are assigned to smaller weights, which are relying on the roles it played. Generally, the larger the weight, the greater contributions it makes to the fusion model. What is more, the high-level fusion procedure can be summarized in the following steps (Fig. 2):

According to the principle that the smaller $RMSEP_f$, the better predictive ability of the model, the one with smaller $RMSEP_f$ is regarded as the optimal MID method. Moreover, it is essential to search an effective MID method for each data set. In the following sections, the characteristics and behaviors of the two MID methods are discussed detailedly by the pesticide and the chemical data sets.

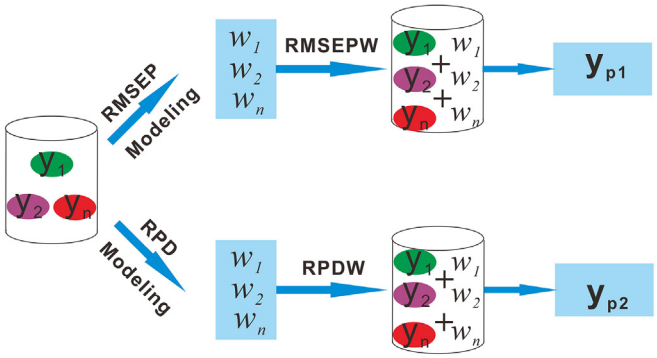

Fig. 1. Scheme for explanation of high-level fusion for quantitative analysis.

2.3. Modeling indicator determined method

2.3.1. Root mean square of prediction weighted (RMSEPW)

The spectral data are optimized at the stage of calibration set and then evaluated by RMSEP. In calibration set, the optimal number of latent variables (LVs) is determined by five-fold cross-validation method with the lowest root mean square error of cross-validation (RMSECV).

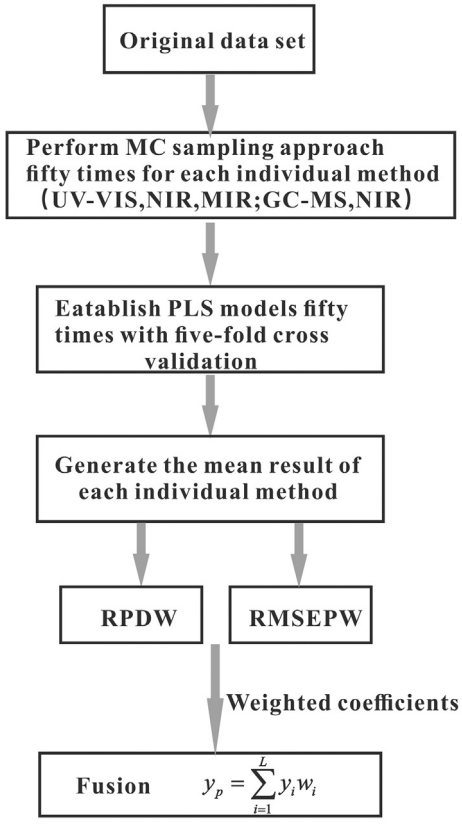

Fig. 2. Scheme for explanation of RMSEPW and RPDW methods for high-level fusion to perform quantitative analysis. (1) Perform Monte-Carlo (MC) sampling approach fifty times to acquire the calibration and validation sets. (2) Generate the mean results of each individual technique by running partial least squares (PLS) with five-fold cross validation fifty times. (3) Combine the individual results to implement high-level fusion by two proposed MID methods (RMSEPW and RPDW). (4) Obtain the final results of $RMSEP_f$ (RMSEP for the fusion result) generated by the two MID methods.

In addition, the performance of the model is assessed by RMSEP which is mathematically expressed as Eq. 2.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y)}{n}} \quad (2)$$

where n is the number of samples in the calibration matrix, $\hat{y}_i$ and y_i are the predicted and measured values, respectively. In RMSEPW approach, the inverse value of RMSEP is assigned to the weight of the associated method. In other word, a high RMSEP value results in a small weigh of the corresponding method for fusion.

2.3.2. Ratio performance deviation weighted (RPDW)

RPD is the ratio of the standard deviation (STD) to RMSECV [24,25]. RPD is an indicator for qualifying the model quality and is calculated by Eq. 3. As acknowledged, the higher the RPD values, the better the model's quality is. According to the literature [26], a RPD higher than 1.5 is considered as an acceptable index for initial screenings, a preliminary prediction of 2.0–2.5 indicates a satisfactory model, and a RPD higher than 3 indicates that the model predicts overwhelmingly efficiency of the model. In RPDW method, the RPD value is in proportion to the weight of the individual method. As a matter of fact, a high RPD brings about a large weight of the fusion method.

$$RPD = \frac{SD}{RMSECV} \quad (3)$$

2.4. Partial least-square regression modeling (PLS)

PLS includes a $(n \times m)$ spectral matrix $\boldsymbol{X}$ with p predictor variables, and a $(n \times p)$ concentration vector $\boldsymbol{y}$. The PLS algorithm is based on the relationship of the signal intensity ($\boldsymbol{X}$) and the sample characteristics ($\boldsymbol{y}$) [27]. In prediction process, the predictive result $\boldsymbol{y_n}$ is obtained from Eq. 4, where the $\boldsymbol{T_n}$ (score matrix of the unknown samples) is calculated from $\boldsymbol{X_n}$, $\boldsymbol{Q}$ is the loading matrix of $\boldsymbol{y}$. In order to obtain a good estimate of $\boldsymbol{b}$, the PLS model needs to be calibrated on samples that span the variation in $\boldsymbol{Y}$.

$$y_n = T_n b Q \quad (4)$$

2.5. Model evaluation

In this study, the results of spectral data are optimized at the calibration stage and then evaluated by root mean squared error of prediction for fusion ($RMSEP_f$), which is the model evaluation indicators for the two weighted methods. The models are all established by five-fold cross validation and the maximum number of LVs is limited to ten.

2.6. Software

The algorithms involved in this study are programmed by Matlab (Version 2016a, the MathWorks, Inc.). The coding scripts used in this study are available upon request.

3. Data description

3.1. Pesticide samples

3.1.1. Samples

Eighty deltamethrin samples were prepared with technical deltamethrin (98.1%, obtained from Jiangsu Huangma Agrochemicals, China), dimethylbenzene (99.0%, Beijing Chemical Works, China) and deltamethrin formulation (25 g/L, Bayer Crop Science, China). The concentration of deltamethrin was ranged from 0.1% to 4.98% (w/w) with the mean value of 2.55%. The exact concentration of deltamethrin in the commercial formulation was determined by high performance liquid chromatography (HPLC).

During the calibration procedure, Monte-Carlo (MC) outlier approach was carried out by running 1000 times to pick out samples exhibited the largest minimum distance. It was essential to identify the outliers and remove them, as they had significant large effect on the model. After kicking out two outliers, the remaining samples were divided into calibration set (48 samples), validation set (15 samples) and test set (15 samples). The test set was sequentially chosen according to the concentration from high to low, while the calibration and validation sets were obtained through MC sampling approach by operating fifty times. In fact, fifty different calibration and validation sets were obtained by MC approach.

3.1.2. UV–visible (UV–vis) spectroscopy

The UV–vis spectral data were acquired by a spectrophotometer (Lambda 35, Perkin Elmer, USA) over the range of 350 to 800 nm. A quartz cuvette with a 1.0 cm path length was employed. The spectral bandwidth and data point interval were both 1.0 nm, and totally 451 data points for each spectrum.

3.1.3. Near infrared (NIR) spectroscopy

The NIR spectra were measured by an FT-NIR spectrometer (Spectrum One NTS, Perkin Elmer, USA) and were recorded from 800 to 2500 nm at a resolution of 4 cm^{-1} with 64 accumulations co-added. Carbon tetrachloride was taken as a reference of the background. The NIR spectra included 2125 data points.

3.1.4. Mid-infrared (MIR) spectroscopy

The MIR spectra were measured by an FT-IR spectrometer (Cary 630, Agilent, USA) with ATR accessory. The spectra were collected over the range of 2500 nm to 15,000 nm (resolution of 4 cm^{-1}, 64 scans) and totally generated 869 data points for each spectrum.

3.2. Chemical samples

3.2.1. Samples

Forty samples were prepared with a mixture of butyl acetate (AR, Beijing Chemical Reagent Company), toluene (AR, Beijing Chemical Reagent Company), acetophenone (AR, Beijing Chemical Reagent Company), cyclohexane (AR, Beijing Chemical Reagent Company) and *n*-heptane (AR, Beijing Chemical Reagent Company), the concentration of *n*-heptane was ranged from 0.01% to 0.38% (w/w) with the mean value of 0.18%. MC outlier approach was carried out to detect outliers after 1000 times running. After identifying two outliers, the remaining samples were divided into calibration set (23 samples), validation set (7 samples) and test set (8 samples). The test set was obtained sequentially according to the concentration, and the calibration and validation set were requires by fifty runs of MC sampling approach. Actually, fifty calibration and validation sets were yield by MC sampling approach.

3.2.2. NIR analysis

The spectral data were acquired by an NIR spectrophotometer (Spectrum One NTS, Perkin Elmer, USA) over the wavelength range of 12,000 to 4000 cm^{-1}. The averaged spectra were obtained with a resolution of 2 cm^{-1} after scanned 32 times. A quartz cell with a 1.0 mm path length was employed. Carbon tetrachloride solution was used as the reference.

3.2.3. Gas chromatography-mass spectrometer (GC–MS) analysis

Chromatographic analysis was performed on a Clarus 500 GC–MS (Perkin Elmer, USA) with HP-5MS column (30 m × 0.25 mm × 0.25 μm). Initially, the GC oven was set as 60 °C, and it rose to 200 °C at rate of 5 °C/min gradually. Sample injection volume was 1 μL with the split ratio of 10:1. Helium (99.999%) was served as the carrier gas with the flowing rate of 1.0 ml/min. In addition, the source temperature

was kept at 180 °C, and the injector and the MS transfer line were 250 °C and 280 °C, respectively. The MS scan range was from 50 to 300 amu.

3.2.4. GC-MS data process

All mass spectral data were stacking together along the time axis. Afterwards, the three-dimensional spectra were converted into a two-dimensional chart with *m/z* as horizontal axis and abundance as vertical axis by the accumulation algorithms.

4. Result and discussion

4.1. Influence of different parameters for high-level fusion quantitative analysis

There were two methods to simplify the GC–MS data matrix and convert the tri-dimensional data cube into a two-dimensional one for PLS analysis. One method was to sum the time-elution profiles over the mass dimension, the other was to sum the mass spectra over the retention time dimension. The former compressed method was to accumulate the retention time at a fixed *m/z*. Since a compound owned its specific *m/z*, the cumulated peak intensity was linearly associated with the chemical concentration. It should be noted that, even if one compound was totally overlapped by others and remaining no characteristic fragment ion peaks, the cumulated mass spectra was still satisfied with the principle of linear addition method. Thus, it was still feasible for quantitative analysis after summing all the retention time along the mass spectrum dimension. With regard to the other method of stacking the *m/z* along the retention time axis, the accumulated peaks were not linearly correlated with the ion concentration, ascribing to that one chemical produced various ion peaks but the ions were acquired with different ionization efficiency. As a consequence, it was failed to execute quantitative analysis. In summary, adding the retention time up over the mass dimension was applied to the GC–MS spectra for further analysis.

After choosing the accumulation method, the parameters of model evaluation indicators were needed to be discussed. It was worth noting that, RMSEP, the prediction error of the test set, was the evaluation index of the model. As acknowledged, a smaller RMSEP gave a better predictability, thereby a smaller RMSEP accounted for a larger weight for the fusion result. It was firmly convinced that RMSEP was inversely proportional to the weight. Therefore, the reciprocal of RMSEP was assigned as the weight for each sensor.

As for RPD, it was an indicator to qualify the ability of the establishing a model. A higher RPD value conduced to a better modeling capability. Generally speaking, if the establishment ability of the model was satisfied, the corresponding predictive ability of the test set would be acceptable. As a result, the weight of each sensor was in proportion to the RPD value.

RMSEP represented the consistency between the measured value and the predicted value. What was more, it was applied to evaluate the accuracy of the model as the error of the developed model. Consequently, RMSEP was generally exploited as the main parameter to evaluate the performance of the model. Accordingly, $RMSEP_f$ was employed to assess the performance of each weighted index. $RMSEP_f$ was obtained by each individual method associated with its own weight, i.e., each sensor accounted for w_i for fusion. Besides, the result of $RMSEP_f$ was compared with the individual RMSEP aiming to assess the performance and effectiveness of different MID methods.

4.2. Pesticide data

The pesticide data matrix was implemented for the high-level fusion. It was coupled with the individual approaches of UV–vis, NIR and MIR for quantitative analysis. In pesticide system, the deltamethrin data set was pre-processed by center treatment standardization before establishing individual regression models. As was known, cross validation was an effective and widely used technique, five-fold cross validation technique was explored and the individual approaches was performed with the maximum LVs number of ten. The two MID methods of RMSEPW and RPDW were carried out for fusion separately.

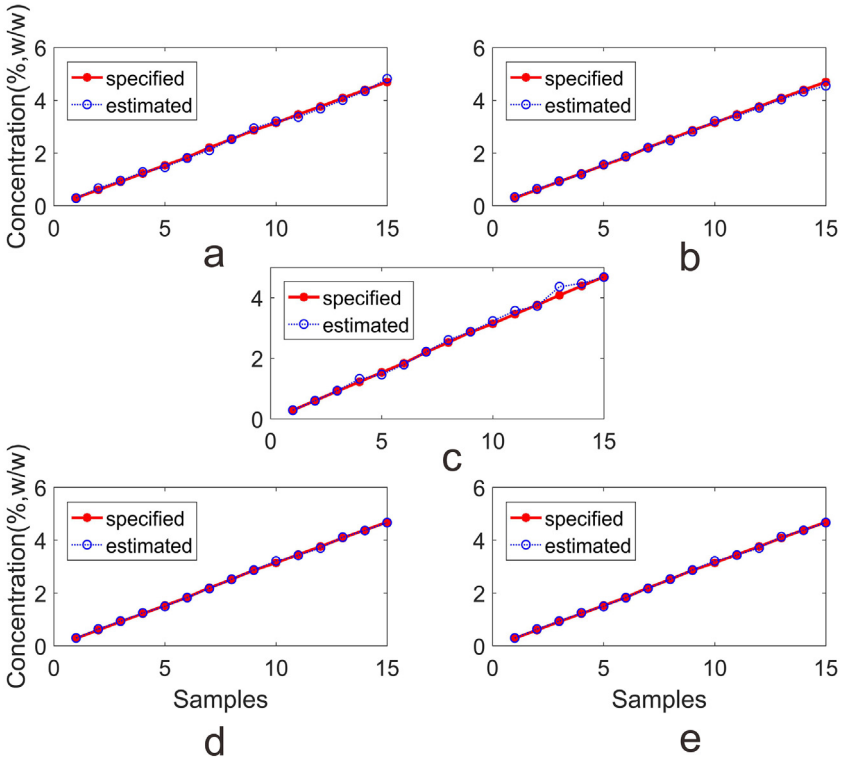

Fig. 3. The plot of predicted and actual value in test set. (a) individual UV–vis; (b) individual NIR; (c) individual MIR; (d) RMSEPW for fusion; (e) RPDW for fusion.

Table 1
Results of RMSEPW, RPDW and individual methods on pesticide data set and chemical data set; STD represents the standard deviation in fifty runs; the value of individual RMSEP and individual RPD are the mean value of fifty runs; the bold font represents the method with better predictive ability.

Data set	Methods	Fusion $RMSEP_f$ (%)		Individual RMSEP (%)	Individual RPD	STD of individual RMSEP (%)	STD of individual RPD
		RMSEP weighted	RPD weighted				
Pesticide (deltamethrin)	UV-vis	**0.0326**	0.0335	0.0862	9.5056	0.0118	1.5573
	NIR			0.0698	11.8119	0.0226	1.5296
	MIR			0.1041	9.8882	0.0107	1.1282
Chemical (*n*-heptane)	NIR	**0.0151**	0.0161	0.0211	3.1008	0.0065	0.7424
	GC–MS			0.0374	2.4729	0.0105	0.4166

To perform quantitative analysis, the individual model of each sensor was employed to calculate its own weight. The individual results of UV–vis, NIR and MIR were shown in Fig. 3a-c. In detail, the models of UV–vis and MIR models were not as efficient as the NIR model. Namely, the NIR method played a key rule during RMSEPW fusion approach. It was explicitly displayed in Table 1 that a smaller RMSEP gave a larger weight, since the weight of the associated method had an inverse relationship with the RMSEP. Fig. 4a was the pie chart of RMSEPW method, wherein the blue portion represented the weight of UV–vis, the yellow and green portions revealed the weights of NIR and MIR. As explained earlier, NIR held a higher contribution rate (40.33%) on account of a better predictive performance, whereas UV–vis and MIR account for a little smaller proportion i.e., 32.64% and 27.03% of the fusion approach.

Based on the weights obtained by the RMSEP evaluation criterion, the high-level fusion was performed for quantitative analysis through Eq. 1. After RMSEPW fusion, the results of predicted and actual value in test set were shown in Fig. 3d-e. Moreover, the evaluation results of different methods on pesticide data set were illustrated in Table 1. It could be drawn from Table 1 that RMSEPW method exhibited better predictive results, i.e., $RMSEP_f$ 0.0326 % compared with the individual methods of UV–vis, NIR and MIR, which revealed that RMSEPW method had improved the stability and predictive performance on pesticide data set. The overall results indicated that the predictive ability was enhanced significantly and thus it was indispensable to perform the weighted strategy on pesticide data set for fusion.

Depending upon the model developing ability, RPD was utilized as an indicator to calculate the weight for each individual method. As seen in Table 1, the three individual models were all well established with the RPD values larger than 3. Specifically, the RPD of UV–vis appeared less satisfactory as that obtained from NIR and MIR models. It was obvious that the weight was linearly associated with the RPD value of corresponding method. As a result, the UV–vis data occupied a smaller proportion (30.13%), whereas NIR and MIR accounted for a larger proportion of 37.45% and 32.42% for high-level fusion, respectively (Fig. 4b). On the basis of the weights gained from the RPD evaluation criterion, the three individual methods were executed for high-level fusion approach. As outlined in Table 1, the RPDW method ($RMSEP_f$ 0.0335 %) had achieved a more impressive performance than individual methods, verifying the obvious advantage of the RPDW method upon the pesticide data. Furthermore, it was also manifested that RPDW method was feasible for high-level fusion to execute quantitative analysis.

As the RMSEPW method revealed the predictive result directly, a higher predictive ability were usually accompanied by RMSEPW method (Fig. 5a). On the whole, RMSEPW method demonstrated a better predictive ability than the RPDW method in terms of $RMSEP_f$. The predictive performance of the five methods followed the order: RMSEPW > RPDW > NIR > UV–vis > MIR. As a matter of fact, RMSEPW method and RPDW method performed superior results in comparison with the individual methods by virtue of taking advantage of each individual method.

4.3. Chemical data

The individual approaches of NIR and GC–MS of chemical data set were implemented for high-level fusion. In chemical system, the *n*-heptane data set was pre-processed by center treatment standardization before establishing individual regression models. Besides, the maximum number of LVs was set to ten according to five-fold cross validation technique. After spectral pre-treatment and modeling parameters well set, the proposed methods of RMSEPW and RPDW were executed on the chemical data set for fusion respectively.

In order to perform the chemical data set of NIR and GC–MS spectra for fusion, the RMSEP and RPD of each model were obtained at first. The individual results of NIR and GC–MS were shown in Fig. 6a-b. As illustrated in Table 1, the individual NIR yielded better results than individual GC–MS. It had been illustrated that the weight of the corresponding method was inversely proportional to the RMSEP. Accordingly, the blue and yellow portion demonstrated the weights of NIR and GC–MS from the pie chart of Fig. 4c. Obviously, the weight proportion for NIR was 63.87%, whereas GC–MS contributed much smaller i.e., 36.13%, clarifying the NIR method was comparatively important than GC–MS method in the RMSEPW method for fusion. The plots of predicted and actual value in test set after fusion were shown in Fig. 6c-d. Results of RMSEPW method and individual methods on chemical data were summarized in Table 1. Compared with the individual methods, RMSEPW method was well predicted on the test set and had

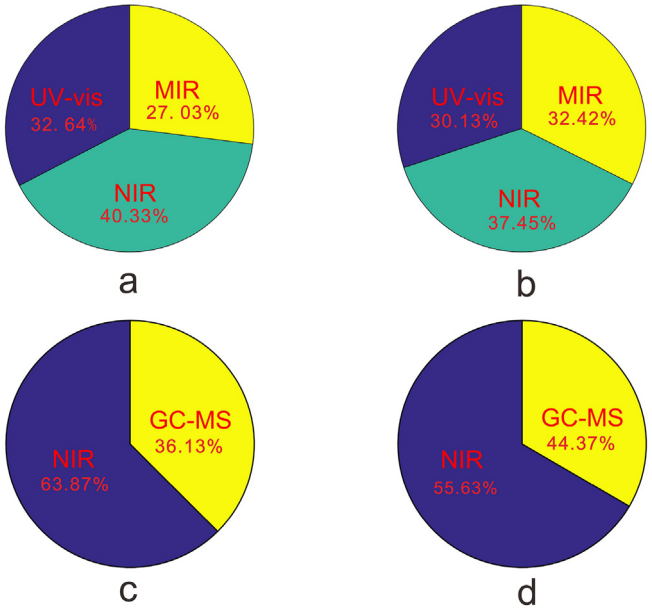

Fig. 4. The pie chart of the weight of each individual method. (a) RMSEPW method for pesticide data set; (b) RPDW method for pesticide data set; (c) RMSEPW method for chemical data set; (d) RPDW for chemical data set.

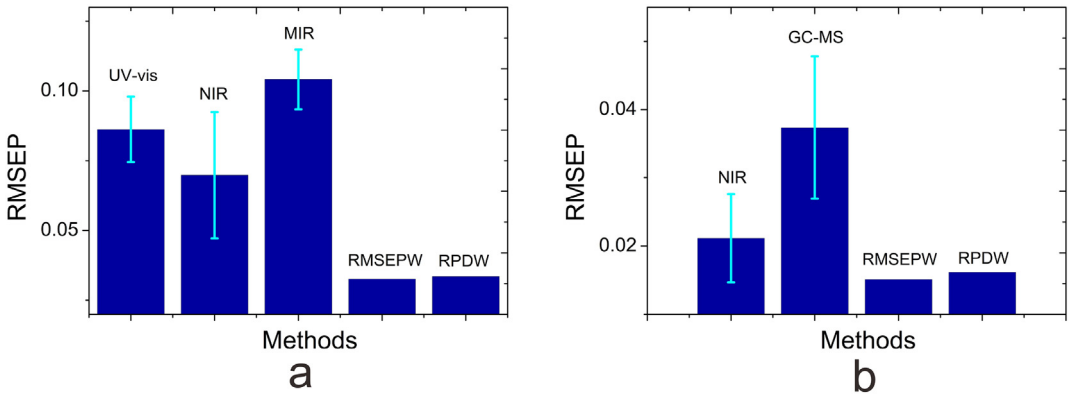

Fig. 5. The RMSEP of individual and fusion methods. (a) RMSEP for pesticide data set; (b) RMSEP for chemical data set.

improved the predictive performance, revealing that it was essential to perform the RMSEPW method for fusion. Namely, the RMSEPW method was more promising than individual methods.

For the RPDW method, RPD was utilized to obtain the weight of the individual methods. In RPDW method, the weight of the corresponding method was linear dependent on the RPD value.

Fig. 4d was the weight of each individual method for RPDW fusion. As concluded from Fig. 4d, GC–MS held a slightly smaller proportion (44.37%), and NIR attributed 55.63% for high-level fusion analysis. Table 1 listed the results of RPDW method and individual methods on chemical data set. As displayed in Table 1, the NIR method with higher RPD than GC–MS method appeared more outstanding than GC–MS method. In particular, RPDW method achieved more impressive performance than individual methods on pesticide data i.e. $RMSEP_f$ 0.0161 %, manifesting the obvious advantage of the RPDW method upon the chemical data. In addition, it indicated that RPDW method was practicable for fusion with smaller $RMSEP_f$ compared with the individual methods.

As demonstrated in Fig. 5b, the RMSEPW method performed better than RPDW method and individual methods which was predominantly owing to that RMSEP was a parameter straightly indicating the predictive performance. Conclusively, a clear ranking for the four methods was displayed: RMSEPW > RPDW > NIR > GC–MS. Actually, in contrast with the performance of the global model, the two weighted methods performed imperative results by means of fusing the individual method for quantitative analysis.

5. Conclusion

In order to perform the high-level fusion for quantitative analysis, we proposed two MID methods of RMSEPW and RPDW to proceed high-level fusion. The weights were calculated by the individual modeling indicator for the fusion models. A low $RMSEPW_f$ directly revealed the predictive performance and consequently gave optimal results. Actually, a better predictive ability was usually coming along with RMSEPW method, whereas RPDW method was a little worse than RMSEPW method for not directly reflecting the predicted results. Only when the original model was acceptable without overfitting and outliers, the outcome of RPDW method might be comparable with RMSEPW method.

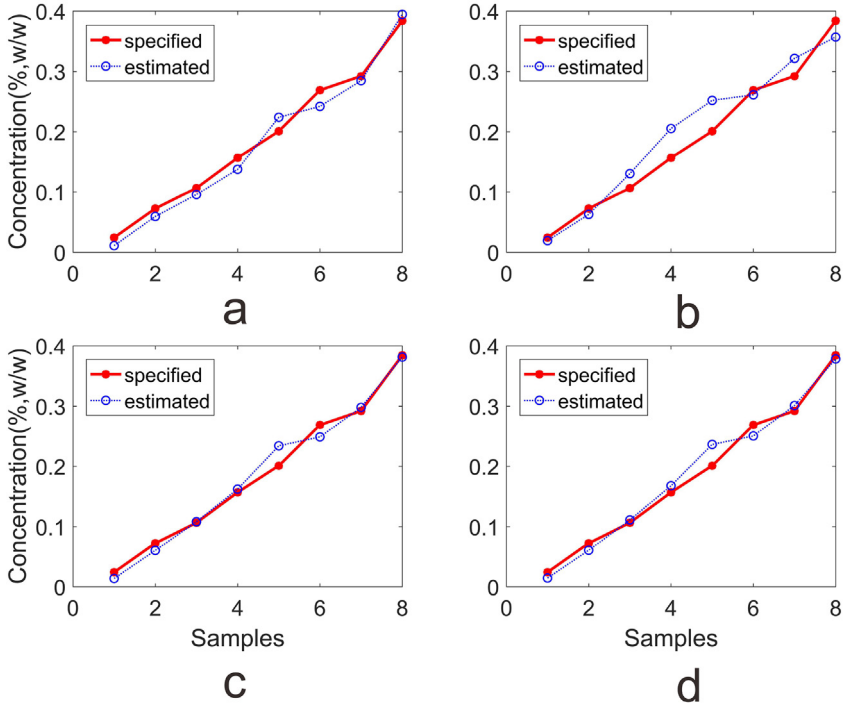

Fig. 6. The plot of predicted and actual value in test set. (a) individual NIR; (b) individual GC–MS; (c) RMSEPW for fusion; (d) RPDW for fusion.

The proposed RMSEPW and RPDW methods were successfully applied to fusion methods of UV–vis, NIR and MIR spectral data of active ingredient and GC–MS and NIR spectral data of *n*-heptane. In a word, the two MID methods were promising with substantial improvement of predictive ability compared with individual methods.

Notes

The authors declare no competing financial interest.

Acknowledgments

This study was funded by the National Natural Science Foundation of China (NSFC) (No. 31301685), and Fundamental Research Funds for the Central Universities (No. 2652015164).

References

[1] I.V. Mechelen, A.K. Smilde, A generic linked-mode decomposition model for data fusion, Chemom. Intell. Lab. Syst. 104 (2010) 83–94.
[2] C. Fernández, M.P. Callao, M.S. Larrechi, UV-visible-DAD and 1H-NMR spectroscopy data fusion for studying the photodegradation process of azo-dyes using MCR-ALS, Talanta 117 (2013) 75–80.
[3] J. Forshed, H. Idborg, S.P. Jacobsson, Evaluation of different techniques for data fusion of LC/MS and 1H-NMR, Chemom. Intell. Lab. Syst. 85 (2007) 102–109.
[4] L. Vera, L. Aceña, J. Guasch, R. Boqué, M. Mestres, O. Busto, Discrimination and sensory description of beers through data fusion, Talanta 87 (2011) 136–142.
[5] B.B. Madhavan, T. Sasagawa, K. Tachibana, K.K. Mishra, A high-level data fusion and spatial modelling system for change-detection analysis using high-resolution airborne digital sensor data, In. J. Remote Sensing 27 (2006) 3571–3591.
[6] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment-a review, Anal. Chim. Acta 891 (2015) 1–14.
[7] J.I. Arandasanchez, A. Baltazar, G. Gonzálezaguilar, Implementation of a Bayesian classifier using repeated measurements for discrimination of tomato fruit ripening stages, Biosyst. Eng. 102 (2009) 274–284.
[8] W. Sun, X. Zhang, Z. Zhang, R. Zhu, Data fusion of near-infrared and mid-infrared spectra for identification of rhubarb, Spectrochim. Acta, Part A 171 (2017) 72–79.
[9] Y. Li, J. Zhang, T. Li, H. Liu, J. Li, Y. Wang, Geographical traceability of wild boletus edulis based on data fusion of FT-MIR and ICP-AES coupled with data mining methods (SVM), Spectrochim. Acta, Part A 177 (2017) 20–27.
[10] A. Baltazar, J.I. Aranda, G. González-Aguilar, Bayesian classification of ripening stages of tomato fruit using acoustic impact and colorimeter sensor data, Comput. Electron. Agric. 60 (2008) 113–121.
[11] X. Zou, J. Zhao, Apple quality assessment by fusion three sensors, IEEE Sens 4 (2005) 389–392.
[12] R. Li, P. Wang, W. Hu, A novel method for wine analysis based on sensor fusion technique, Sens. Actuators B Chem. 66 (2000) 246–250.
[13] L.I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, John Wiley and Sons, Inc, Hoboken, New Jersey, 2004 106–107 Chapter 3.
[14] T.G. Doeswijk, A.K. Smilde, J.A. Hageman, J.A. Westerhuis, F.A. van Eeuwijk, On the increase of predictive performance with high-level data fusion, Anal. Chim. Acta 705 (2011) 0–47.
[15] I. Ruthven, M. Lalmas, Using Dempster-Shafer's theory of evidence to combine aspects of information use, J. Intell. Inf. Syst. 19 (2002) 267–301.
[16] J. Wang, J. Cheng, H. Liu, Z. Tang, D. Han, Optimization of informative spectral regions in FT-NIR spectroscopy for measuring the soluble solids content of apple, Intell. Automation Soft Comput. 21 (2015) 355–370.
[17] J.C. Tewari, V. Dixit, B.K. Cho, K.A. Malik, Determination of origin and sugars of citrus fruits using genetic algorithm, correspondence analysis and partial least square combined with fiber optic NIR spectroscopy, Spectrochim. Acta, Part A 71 (2009) 1119–1127.
[18] C. Tan, X. Qin, M. Li, An ensemble method based on a self-organizing map for near-infrared spectral calibration of complex beverage samples, Anal. Bioanal. Chem. 392 (2008) 515–521.
[19] X. Wu, Z. Liu, T. Zhang, H. Li, A method based on double models combination to further reduce root-mean-square error and relative error of prediction, Chin. J. Anal. Chem. 43 (2015) 754–758.
[20] A.O. Aptula, N.G. Jeliazkova, T.W. Schultz, M.T.D. Cronin, The better predictive model: high q^2 for the training set or low root mean square error of prediction for the test set? QSAR Comb. Sci. 24 (2005) 385–396.
[21] R.A.V. Rossel, R.N. Mcglynn, A.B. Mcbratney, Determining the composition of mineral-organic mixes using UV-vis-NIR diffuse reflectance spectroscopy, Geoderma 137 (2006) 70–82.
[22] L. Karlinasari, M. Sabed, I.N.J. Wistara, Y.A.J. Purwanto, Near infrared (NIR) spectroscopy for estimating the chemical composition of (Acacia mangium, Willd.) wood, Indian Acad. Wood Sci. 11 (2014) 162–167.
[23] A. Decelle, F. Krzakala, Belief-propagation-guided Monte-Carlo sampling, Phys. Rev. B 89 (2014) 817–824.
[24] B. Kuang, A.M. Mouazen, Non-biased prediction of soil organic carbon and total nitrogen with vis-NIR spectroscopy, as affected by soil moisture content and texture, Biosyst. Eng. 114 (2013) 249–258.
[25] L.R. Schimleck, R. Evans, J. Ilic, Estimation of Eucalyptus delegatensis wood properties by near infrared spectroscopy, Can. J. For. Res. 31 (2001) 1671–1675.
[26] A.M. Mouazen, W. Saeys, J. Xing, J. de Baerdemaeker, H. Ramon, Near infrared spectroscopy for agricultural materials: an instrument comparison, J. Near Infrared Spectrosc. 13 (2005) 87–98.
[27] V. Gaydou, J. Kister, N. Dupuy, Evaluation of multiblock NIR/MIR PLS predictive models to detect adulteration of diesel/biodiesel blends by vegetal oil, Chemom. Intell. Lab. Syst. 106 (2011) 190–197.