Comparison of Ensemble Strategies in Online NIR for Monitoring the Extraction Process of Pericarpium Citri Reticulatae Based on Different Variable Selections

Planta Med

This electronic reprint is provided for non-commercial and personal use only: this reprint may be forwarded to individual colleagues or may be used on the author’s homepage. This reprint is not provided for distribution in repositories, including social and scientific networks and platforms.

Publishing House and Copyright:
© 2015 by
Georg Thieme Verlag KG
Rüdigerstraße 14
70469 Stuttgart
ISSN 0032-0943

Any further use only by permission of the Publishing House
Comparison of Ensemble Strategies in Online NIR for Monitoring the Extraction Process of Pericarpium Citri Reticulatae Based on Different Variable Selections

Zheng Zhou¹, ⁴, Yang Li¹, Qiao Zhang¹, Xinyuan Shi¹, ², ³, Zhisheng Wu¹, ², ³, Yanjiang Qiao¹, ², ³

¹ Beijing University of Chinese Medicine, Beijing, China
² Beijing Key Laboratory for Basic and Development Research on Chinese Medicine, Beijing, China
³ Beijing Key Laboratory of TCM-information Engineering of State Administration of TCM, Beijing, China
⁴ Fujian University of Traditional Chinese Medicine, Fujian, China

Abstract

Different ensemble strategies were compared in online near-infrared models for monitoring active pharmaceutical ingredients of Traditional Chinese Medicine. Bagging partial least square regression and boosting partial least square regression were adopted into the quantiative analysis. The results demonstrated that the established approach could be applied for rapid determination and real-time monitoring of hesperidin and nobiletin in Pericarpium Citri Reticulatae (Citrus reticulata) during the extraction process. Comparing the results, the boosting partial least square regression provided a slightly better accuracy than the bagging partial least square regression. Finally, this paper provides a promising ensemble strategy on online near-infrared models in Chinese medicine.

Introduction

The development of a calibration model is a critical procedure for near-infrared (NIR) quantitative analysis. However, in the application process, lots of factors could have a great impact on the calibration model performance, such as overlapping absorption bands, high detection limit, low sensitivity, etc. [1, 2]. It requires great efforts to improve the model performance, including spectral pretreatment methods, variable selection, and selection of latent variables. Most of those efforts are carried out on a single model, nevertheless the accuracy and the robustness of established models are not always persuasive enough [3].

Ensemble learning methods combing the results of multiple regression models have attracted more and more attention. This can lead to an increased accuracy of the model and also to a better stability [4]. Bagging and boosting are two of the best known ensemble learning methods due to their theoretical performance and strong experimental results. They are both established based on re-sampling algorithms that are applied on the training data.

Bagging, short for “bootstrap aggregating”, was originally introduced by Breiman [5]. It trains different classifiers with bootstrapped replicas of the original training data. Therefore, a new data sample is generated out of the original data by random sampling with replacement (usually maintaining the original size of the data) [6]. Diversity models are obtained with this resampling procedure by using different data subsets, which could lead to a good prediction performance towards an unknown instance. The bagging partial least square (bagging-PLS) method has been applied in the field of chemometrics. Viscarra Rossel [7] has established a partial least square (PLS) method coupled with a bagging algorithm to predict the soil organic carbon. These results showed a better performance than single PLS models.

Boosting, based on an ensemble of individual models, was initially proposed by Schapire [8]. The basic principle of boosting is to combine the
outputs of many “weak” learners to produce a powerful “committee”. Compared to the random sampling with replacement, which is applied in bagging, the predictive accuracy of boosting is higher, because of its sampling rule is according to the sample weight [9]. Samples in calibration sets with a larger prediction error would be assigned to a larger weight and then be selected in a higher probability. This can undoubtedly reduce the overall risk of making a final decision and increase the accuracy and robustness of the calibration [10]. Massart [11] proposed a method called boosting PLS, and the results demonstrated that it is more resistant to overfitting than classical PLS, but without reducing the accuracy. Therefore, both bagging and boosting algorithms are worthwhile, as they can improve the results that are obtained by the usage of data preprocessing methods and training of a single model [6].

Pericarpium Citri Reticulatae (Chenpi in Chinese), the dried ripe pericarp of Citrus reticulata Blanco (Rutaceae) or its cultivars, is one of the most commonly used medicinal herbs listed in the Chinese Pharmacopoeia [12]. Chenpi extract is rich in flavonoids, which are proved to possess a variety of biological activities, such as antispasmodic, anti-inflammatory, or antimicrobial activities [13]. Among these flavonoids, hesperidin and nobiletin are two important active components. Thus, the hesperidin and nobiletin contents are important indicators for the quality control of Chenpi.

Since the “Process Analysis Technology (PAT) Industry Guide” was issued by the U.S. Food and Drug Administration in September 2004 [14], methods based on inline and realtime measurements and a rapid, reliable, and non-invasive technique gain increasing interest in the pharmaceutical industry. With development of chemometrics, the application of NIR spectroscopy as an online and realtime monitoring technique has grown rapidly in recent years. A method for the determination of rutin concentration was established for a pilot scale extraction process of Sophora japonica L. (Fabaceae) using NIR technology [15]. The chlorogenic acid content in Lonicera japonica Thunb. (Caprifoliaceae) was determined by NIR spectroscopy [16]. Major bioactive isoflavonoid compounds obtained through the extraction process of kudzu [Pueraria lobata (Willd.) Ohwi] (Fabaceae)] were rapidly determined by NIR transmission spectroscopy [17]. Extraction is one of the most important manufacturing steps and the initial process to obtain active pharmaceutical ingredients. To improve the efficiency of the process and to guarantee the final product’s quality, PAT should be utilized in the manufacturing process of Traditional Chinese Medicine, especially in the extraction unit. The aim of this paper was to use NIR spectroscopy for online and realtime monitoring of the extraction process of Chenpi in a pilot scale system. NIR spectroscopy, performed with a flow cell and two fiber optic probes connected to a NIR instrument, was tested to simultaneously monitor the concentration variation of hesperidin and nobiletin during the Chenpi extraction process. PLS was adopted to establish quantitative models between the NIR spectra and HPLC reference data. Two variable selection methods, including synergy interval partial least squares algorithm (SiPLS) and backward interval partial least squares (BiPLS), were applied to select the efficient spectral regions and to optimize the model performance. Ensemble strategy, bagging-PLS and boosting-PLS were applied to compare the accuracy and the robustness of the model. The performance of the models were evaluated using common chemometric indicators, such as root-mean-square error of cross-validation (RMSECV), root-mean-square error of calibration (RMSEC), root-mean-square error of prediction (RMSEP), and the coefficient of determination ($R^2$).

**Results and Discussion**

The raw NIR spectra of the sample solutions are shown in Fig. 1. As displayed there, no significant differences can be observed between the raw spectra of the samples due to high overlapping. The NIR spectra show intense absorption bands around 1400 nm from the first O–H overtone and around 1900 nm from the combination of stretching and deformation of the O–H group in water. These two bands, which are typical for a NIR spectrum of an aqueous solution, mask any other band present in this spectral range [18]. Therefore, it is necessary to identify the usable spectral regions before the development of calibration models, in order to maximize the investigated components in the PLS model, and to minimize the influence of noise or other useless signals at the same time.

PLS regression was employed to investigate the correlation between the spectral data and the concentration variables measured by HPLC assays. It is generally known that the spectral preprocessing treatments and the number of latent factors are critical parameters in PLS models. To test the robustness of PLS models, several types of preprocessing methods were tested on the spectral data set, including Savitzky-Golay smoothing, derivatives, multiplicative scatter correction, standard normal variate, normalize and their combinations. The influence of different preprocessing methods on the performance of PLS models is shown in Table 1. The optimum number of latent factors is determined by the predicted residual sum of squares (PRESS). Fig. 2 shows PRESS as a function of latent factors to determine the contents of two components using different spectral preprocessing methods. The raw spectra of hesperidin and nobiletin were superior to other spectral preprocessing methods for the PLS model. Due to redundant information gained through NIR spectroscopy, it is necessary to optimize the wavelength regions before modeling calibration, and to find optimum modeling bands for Citrus hesperidin and nobiletin. A few consecutive wavelengths can be combined to cause a decrease of variables. In this study, SiPLS and BiPLS were applied to select the efficient spectral regions that provide the lowest prediction error, in comparison to the full-spectral model.

The whole spectral range from 800 nm to 2200 nm was divided into 20 equal subintervals, from which three were selected to establish the PLS models; the model with 10 factors was taken into account. For the comparison of these models and the full spectra model, RMSECV was used. For hesperidin and nobiletin, the optical subinterval combinations were 1080–1150 nm, 1290–1360 nm, and 2130–2200 nm, or 1010–1080 nm, 1570–1640 nm, and 1710–1780 nm, respectively, as described by the dash area in Fig. 3. The RMSEC, RMSECV, and RMSEP values, as well as the coefficient of determination ($R^2$) of SiPLS are shown in Table 2, in order to compare the PLS models with each other.

The results show that the model performance was slightly enhanced for both compounds of Pericarpium Citri Reticulatae when SiPLS was used. The correlations between the NIR predicted concentration values and the reference values determined by HPLC are shown as diagrams in Fig. 4. It can be seen that the values predicted from NIR spectra were close to the values determined by HPLC. This indicates that the established online models
could be used for quantitative analysis of the Pericarpium Citri Reticulatae extract solutions.

For the BiPLS algorithm, the optical subinterval combinations of hesperidin and nobiletin were 1010–1080 nm, 1290–1430 nm, 1500–1570 nm, and 1780–1850 nm, or 940–1500 nm, 1570–1850 nm, and 2130–2200 nm, respectively, as described by the dash area in Fig. 5. It can be seen from Table 2 that the BiPLS model delivered a good performance for both compounds compared with the PLS model, with satisfying values of the chemometric indicators RMSEC, RMSECV, and coefficient of determination ($R^2$). The correlation between the NIR predicted concentration values and the reference values determined by HPLC reached a good agreement, which can be obtained from Fig. 6. This indicates that the established BiPLS models can be used for online quantitative analysis of hesperidin and nobiletin during Pericarpium Citri Reticulatae extraction.

Recently, an ensemble strategy, which is based on the concept of building a series of models rather than a single model, has shown a good performance for the model [19]. In this section, ensemble strategy combined with PLSR was employed to build online NIR quantitative models during the Pericarpium Citri Reticulatae extraction process. The ensemble modeling steps are illustrated in Fig. 7.

After selecting efficient wavelength regions, bagging-PLS and boosting-PLS were established to improve the predictive accuracy of PLS models. The number of calibration samples used in each ensemble set was the same as in the original calibration set, and the iteration number was set 1000. For hesperidin, when the number of latent variables (LVs) was 9, a stable model could be obtained. For nobiletin, a stable model could be obtained when the number of LVs was 6. Variation of RMSEP with the iteration number is shown in Fig. 8. In order to compare the performance of the ensemble models of bagging-PLS and boosting-PLS in this paper, RMSEP and RPD were used as indexes to evaluate the model performance [20], where RPD is a non-dimensional statistic for calibration model. The ratio of (the standard error of) prediction to (standard) deviation (RPD), is a simple character that enables the evaluation of a standard error of prediction (SEP) in terms of the standard deviation (SD) of the reference data. It is calculated by dividing the SD of the reference values used in the validation or in the prediction sample. It is ideal when the RPD value is higher than 5 [21, 22]. Table 3 shows that both of the two ensemble strategies produce a lower RMSEP value and a satisfying RPD. For hesperidin, the RPD values of bagging-PLS and boosting-PLS were both higher than 5, what means that the ensemble models perform well and could be used for quality con-

<table>
<thead>
<tr>
<th>Pretreatment</th>
<th>Hesperidin RMSEC (mg/mL)</th>
<th>$R^2$</th>
<th>RMSECV (mg/mL)</th>
<th>$R^2$</th>
<th>Nobiletin RMSEC (mg/mL)</th>
<th>$R^2$</th>
<th>RMSECV (mg/mL)</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAW</td>
<td>0.0202</td>
<td>0.9764</td>
<td>0.0274</td>
<td>0.9585</td>
<td>0.0005</td>
<td>0.9872</td>
<td>0.0006</td>
<td>0.9776</td>
</tr>
<tr>
<td>SG9</td>
<td>0.0210</td>
<td>0.9744</td>
<td>0.0270</td>
<td>0.9595</td>
<td>0.0005</td>
<td>0.9867</td>
<td>0.0006</td>
<td>0.9792</td>
</tr>
<tr>
<td>SG11</td>
<td>0.0212</td>
<td>0.9739</td>
<td>0.0271</td>
<td>0.9594</td>
<td>0.0005</td>
<td>0.9866</td>
<td>0.0006</td>
<td>0.9795</td>
</tr>
<tr>
<td>SG11+1D</td>
<td>0.0219</td>
<td>0.9723</td>
<td>0.0869</td>
<td>0.5833</td>
<td>0.0005</td>
<td>0.9863</td>
<td>0.0025</td>
<td>0.6454</td>
</tr>
<tr>
<td>SG11+2D</td>
<td>0.0936</td>
<td>0.4922</td>
<td>0.1390</td>
<td>0.0667</td>
<td>0.0005</td>
<td>0.9863</td>
<td>0.0025</td>
<td>0.6454</td>
</tr>
<tr>
<td>Normalize</td>
<td>0.0263</td>
<td>0.9599</td>
<td>0.0331</td>
<td>0.9395</td>
<td>0.0007</td>
<td>0.9735</td>
<td>0.0008</td>
<td>0.9595</td>
</tr>
<tr>
<td>MSC</td>
<td>0.0255</td>
<td>0.9622</td>
<td>0.0332</td>
<td>0.9392</td>
<td>0.0006</td>
<td>0.9775</td>
<td>0.0008</td>
<td>0.9621</td>
</tr>
<tr>
<td>SNV</td>
<td>0.0253</td>
<td>0.9630</td>
<td>0.0329</td>
<td>0.9402</td>
<td>0.0007</td>
<td>0.9732</td>
<td>0.0009</td>
<td>0.9527</td>
</tr>
</tbody>
</table>

Table 1 Effects of different pretreatment methods on calibration performance.
For nobiletin, the performance of the ensemble models was excellent as RPD values were higher than 9, which is very suitable for quality and process control.

As described above, we conclude that bagging-PLS and boosting-PLS could produce a better model performance than the SiPLS model for online quantitative NIR models of hesperidin and nobiletin during the Pericarpium Citri Reticulatae extraction process. It can be found that the model performance of boosting-PLS is slightly better, compared with the performance of bagging-PLS (Fig. 8), which is consistent with the literature reported [9, 23]. So, boosting-PLS were put forward to apply to an ensemble strategy in online quantitative NIR models of hesperidin and nobiletin.

**Table 2** Effects of different variable selection methods on the model performance.

<table>
<thead>
<tr>
<th>Component</th>
<th>Method</th>
<th>Calibration</th>
<th>Validation</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RMSEC (mg/mL)</td>
<td>R²</td>
<td>RMSECV (mg/mL)</td>
</tr>
<tr>
<td>Hesperidin</td>
<td>PLS</td>
<td>0.0202 0.9764</td>
<td>0.0274</td>
<td>0.9585</td>
</tr>
<tr>
<td></td>
<td>SiPLS</td>
<td>0.0169 0.9835</td>
<td>0.0219</td>
<td>0.9735</td>
</tr>
<tr>
<td></td>
<td>BiPLS</td>
<td>0.0150 0.9870</td>
<td>0.0192</td>
<td>0.9798</td>
</tr>
<tr>
<td>Nobiletin</td>
<td>PLS</td>
<td>0.0046 0.9872</td>
<td>0.0062</td>
<td>0.9776</td>
</tr>
<tr>
<td></td>
<td>SiPLS</td>
<td>0.0046 0.9872</td>
<td>0.0053</td>
<td>0.9837</td>
</tr>
<tr>
<td></td>
<td>BiPLS</td>
<td>0.0048 0.9860</td>
<td>0.0057</td>
<td>0.9813</td>
</tr>
</tbody>
</table>

---

**Fig. 2** PRESS plot of a PLS model using different pretreatments: A hesperidin, B nobiletin. (Color figure available online only.)

**Fig. 3** Optimum subinterval combinations selected by SiPLS for the quantitative determination of hesperidin (A) and nobiletin (B). (Color figure available online only.)

---

Zhou Z et al. Comparison of Ensemble... Planta Med
biletin during the Pericarpium Citri Reticulatae extraction process.
It is the aim of this study to establish a NIR model coupled with bagging-PLS and boosting-PLS, in order to improve the prediction accuracy and robustness. Hesperidin and nobiletin were simultaneously determined during the water extraction process of Pericarpium Citri Reticulatae with HPLC serving as reference. In addition, SiPLS and BiPLS are both recommended for the practical implementation in selecting spectral variables. Finally, the results showed that bagging-PLS and boosting-PLS could lead to an improvement in the prediction accuracy of online NIR models. Boosting-PLS was put forward to apply to an ensemble strategy for online NIR quantitative models of hesperidin and nobiletin during the Pericarpium Citri Reticulatae extraction process.

Materials and Methods

Materials
Pericarpium Citri Reticulatae were purchased from Bei Jing Ben Cao Fang Yuan. Standards of hesperidin (purity > 95.3%; lot number 110721–201115) and nobiletin (purity > 98%; lot number C35–111012) were supplied by the National Institute for Food and Drug Control and the National Engineering Research Center for Solid Preparation of Chinese Medicine, respectively. HPLC grade acetonitrile was purchased from Fisher Chemicals (Fisher Scientific). Deionized water was purified by using a Milli-Q water system (Millipore Corp.). All other reagents were of analytical grade.
Extraction of Pericarpium Citri Reticulatae was carried out on a 100 L multi-function extractor (Longye Medicine Equipment Co., Ltd.). 6 kg of the Pericarpium Citri Reticulatae material was extracted twice with deionized water, the first extraction process with 78 L water lasted 4 hours, the second with 72 L water 2 hours. During the extraction processes the online NIR spectra were collected. Samples of about 5 mL were taken with a sample cup at regular intervals and then analyzed by HPLC. In order to obtain a similar prediction accuracy and a trend of various concentrations, it was necessary to ensure a uniform distribution of the samples [24]. Hence, in the first extraction process, samples were collected at intervals of 4 min during the heating process, 5 min during the first 2 hours and 10 min during the subsequent 2 hours of the boiling process. In the second extraction process, samples were collected every 5 min during the heating process, 5 min during the first hour and 10 min during the remaining hour of the boiling process. In this study, 63 samples were collected, which were split into 42 calibration samples and 21 validation samples using the Kennard-Stone (KS) algorithm.

Spectra acquisition
The online NIR spectra were collected by two fiber optic probes designed to transmit NIR radiation through a 2 mm optical path flange, which was connected to a XDS process analyzer (Foss NIR Systems). Transmission spectra of samples were collected from 800 nm to 2200 nm, and for each spectrum an average of 32 scans with a resolution of 0.5 nm in air was taken as reference standard. The extract was filtered with 80 µm and 100 µm inline strainers before entering the optical path flange, in order to avoid the influence of solid impurities. Multivariate data analysis was performed with Unscrambler 9.7 (CAMO), VISION software (Foss NIR System), and Matlab 7.8 software (The Mathworks). The algorithms used in this study were developed by Norgaard et al. and downloaded from http://www.models.kvl.dk/.

High performance liquid chromatography analysis
A HPLC analysis was performed for the quantitative determination of hesperidin and nobiletin in the samples. The results were used as reference data for the NIR analysis. Chromatographic analysis was performed on a Waters 2695 apparatus, which was comprised of an auto-sampler, a column temperature controller, and a diode-array detector (DAD) (Waters). The extract solution was analyzed by using a DIAKMA Diamonsil C18 column (150 mm × 4.6 mm; 5 µm particles; Dikma). The mobile phase was composed of acetonitrile (A) and water containing 0.1% acetic acid (B). The gradient elution had the following profile: 12% (A) for 0–8 min, 12–22% (A) for 8–20 min, 22% (A) for 20–30 min, 22–60% (A) for 30–40 min, 60–95% (A) for 40–55 min, 95–12% (A) for 55–57 min, and 12% (A) for 57–65 min, at a flow rate of 1.0 mL/min [25]. The wavelength of the UV detector was set to 283 nm and the injection volume was 10 µL.

---

Fig. 6 Correlation diagrams between the NIR predicted values and the reference values of hesperidin (A) and nobiletin (B) (BiPLS). (Color figure available online only.)

Fig. 7 Illustration of ensemble strategy based re-sampling algorithms. (Color figure available online only.)
Fig. 9 shows a typical HPLC chromatogram of mixed standards and Pericarpium Citri Reticulatae solution sample. Baseline separation for the two analytes can be achieved. The main methodology parameters of the validation were consistent with the description in Chinese Pharmacopoeia [12] and as previously reported [26]. Linear correlation analysis for the two analytes was determined using six different concentrations of standard solutions. The calculated results are shown in Table 4.

Fig. 10 shows the dynamic curves of the hesperidin and nobiletin concentrations in the first and second extraction processes. In the initial 60 min of the first extraction processes, the concentration of hesperidin and nobiletin increased quickly. Then, in the second process, the concentrations of the two analytes grew slowly for 40 min and then stayed constant. In the whole process of extraction, the minimum and maximum concentration of hesperidin was 28.45 µg/mL and 391.30 µg/mL, respectively; the concentrations of nobiletin ranged from 1.89 µg/mL to 12.68 µg/mL and were included to the linear range shown in Table 4.
### Acknowledgements

This work was financially supported by National Natural Science Foundation of China (81203218), the Doctoral Fund of Ministry of Education of China (20130013120006), and the BUCM Found for Excellent Young Scholars, which are gratefully acknowledged.

### Conflict of Interest

The authors declare that there is no conflict of interest with any financial organization regarding the material discussed in this manuscript.

### References


### Table 4 Calibration curve data for the reference compounds.

<table>
<thead>
<tr>
<th>Compound</th>
<th>Regression equation*</th>
<th>$R^2$</th>
<th>Linear range (µg/mL)</th>
<th>Min (µg/mL)</th>
<th>Max (µg/mL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hesperidin</td>
<td>$y = 14419x + 227.77$</td>
<td>0.9991</td>
<td>27.84–417.60</td>
<td>28.45</td>
<td>391.30</td>
</tr>
<tr>
<td>Nobiletin</td>
<td>$y = 17118x – 79.809$</td>
<td>0.9994</td>
<td>1.78–16.70</td>
<td>1.89</td>
<td>12.68</td>
</tr>
</tbody>
</table>

* $y$ and $x$ denote the peak area and injection concentration (µg/mL)

Fig. 10 Extraction kinetic curves for hesperidin (A) and nobiletin (B).