Chemometrics and Intelligent Laboratory Systems 144 (2015) 56–62

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Using consensus interval partial least square in near infrared spectra analysis

Guoli Ji [a,d], Guangzao Huang [a,b], Zijiang Yang [b], Xiaohui Wu [a], Xiaojing Chen [c,*], Mingshun Yuan [a,b]

[a] *Department of Automation, Xiamen University, Xiamen 361005, Fujian, China*
[b] *School of Information Technology, York University, Toronto M3J 1P3, Canada*
[c] *College of Physics and Electronic Engineering Information, Wenzhou University, Wenzhou 325035, Zhejiang, China*
[d] *Innovation Center for Cell Biology, Xiamen University, Xiamen 361102, Fujian, China*

ARTICLE INFO

Article history:
Received 26 December 2014
Received in revised form 18 March 2015
Accepted 21 March 2015
Available online 28 March 2015

Keywords:
Consensus modeling
Error correlation
Partial least squares
Interval partial least squares
Consensus interval partial least squares
Near infrared spectra

ABSTRACT

This paper proposes a novel consensus modeling method for regression, which optimizes the weight coefficients of member models considering both error and error correlation of member models. Thus, the optimized objective function has clear physical significance. Furthermore, the root-mean-square error of cross-validation (RMSECV) and root-mean-square error of prediction (RMSEP) of the consensus model are better than any member model. Integrating this method with interval partial least squares algorithm (iPLS), the novel consensus interval partial least squares algorithm (CPLS) is achieved. The typical near infrared spectroscopy datasets are used to validate the effectiveness of CPLS. Compared to the commonly used partial least squares (PLS), iPLS and staked interval partial least squares algorithm (SPLS), CPLS produces better prediction performance.

1. Introduction

Near-infrared spectrometry (NIRS) is a simple, rapid, low cost, pollution-free, and non-destructive technique and has been widely applied in many fields such as farm produce, food, Chinese herbs, tobacco, and so on in recent years [1–5]. It is based on molecular overtone and combination vibrations. In addition, there are various disturbances and the influence of physical factors. Thus, NIRS has complicated background with peak overlapping and weak signal. Therefore, using chemometrics to extract the useful information from complicated, overlapping and changing analytical signals to build calibration models is the key in NIRS analysis.

Partial least squares algorithm (PLS) [6–15] is the most commonly used calibration model, which reflects the relationship between the independent variable (spectra) and the dependent variable (attribute information). Thus, the attribute information of unknown samples can be predicted by the established regression model. PLS applies principal component analysis' (PCA) idea of extracting components from the independent variables [16–18]. Moreover, it takes the interpretation of independent variables on dependent variables into consideration when extracting components. Therefore, PLS is able to interpret dependent and independent variables well and eliminate noise in the system to some extent.

Generally, NIRS has hundreds of variables and there exists serious multicollinearity among the variables. PLS, as a type of full spectrum analysis method, resolves multicollinearity problem efficiently. However, not all the variables make positive contributions to build PLS model. Some variables are not helpful, which makes PLS model more sophisticated and reduces the prediction accuracy. Theory and a large number of experimental results show that variable selection methods [19,20], which select characteristic wavelengths before modeling improve the effectiveness of PLS. Variable selection methods discard the irrelevant and nonlinear variables, which make the built model easier, more accurate and robust.

Nowadays the commonly used variable selection methods include correlation coefficient method (RC) [21], uninformative variable elimination (UVE) [22,23], successive projection algorithm (SPA) [24,25], genetic algorithm (GA) [26], interval partial least squares algorithm (iPLS) [27–30] etc. Among them, iPLS has the advantages of simplicity, visualization and can quickly access the feature spectral bands. Therefore, it is very widely used in NIRS. iPLS splits the full spectrum into several disjoint intervals with equal width. Then the optimal sub-interval is chosen to build PLS model. Obviously it is not an efficient way to utilize the spectral information and may lose useful information in other sub-intervals.

Consensus modeling introduces a new way to modeling. It establishes multiple member models and then combines them to form a consensus predicted result, which is different from traditional modeling approaches [31–37]. Consensus modeling aims to get a composition

[*] Corresponding author. Tel./fax: +86 577 86689027.
E-mail address: chenxj@wzu.edu.cn (X. Chen).

http://dx.doi.org/10.1016/j.chemolab.2015.03.008
0169-7439/© 2015 Elsevier B.V. All rights reserved.

model, which is more generalized and can make a more reliable prediction of unknown samples than single model. Consensus modeling has recently won popularity in many fields. Many studies have shown that consensus modeling does improve the applicability and precision of single models.

This paper presents a consensus modeling method for regression. A new consensus interval partial least squares algorithm (CPLS) is obtained after the consensus modeling method is combined with iPLS. This study also found that the staked interval partial least squares algorithm (SPLS) [35] is equivalent to our proposed CPLS when CPLS ignores the effects of error correlation among member models. The performance of CPLS is evaluated using the classical near infrared spectroscopy datasets and the results show that the proposed methods yield superior performance compared to PLS, iPLS and SPLS.

2. Theory

2.1. Interval partial least squares algorithm (iPLS)

iPLS is a band selection method proposed by Norgaard [29]. This approach splits the full spectrum response matrix (X for *s* samples measured at p spectral wavelengths) into n disjoint intervals (X_1,X_2,…, and X_n) with equal width (p/n channels). For each interval, a local PLS is established. Since each sub-interval may contain different spectral information, the prediction ability of corresponding local PLS is also different. RMSECV values are utilized to focus on the important spectral regions and eliminate the other regions. The best regression model based on sub-intervals should produce the lowest RMSECV value. iPLS can extract the spectral channels highly relevant to the property, thus achieving the objective to improve the stability of the prediction model and increase the interpretability of the relationship between the response and property. Selecting suitable regions or channels in the spectrum, iPLS could get a lower root mean square error of prediction (RMSEP) than PLS.

2.2. Stacked interval partial least squares algorithm (SPLS)

Similarly to iPLS, SPLS also splits the full spectrum response matrix into n disjoint intervals with equal width and establishes a local PLS regression for each interval. The main difference between these two approaches lies in the way to deal with the local PLS models. iPLS chooses the single best local PLS model while SPLS integrates the local PLS models using different weights into a whole model in order to minimize the RMSECV value. SPLS is illustrated below [35]:

$$
\begin{aligned}
&f(x) = \sum_{k=1}^{n} w_k \hat{y}_k \\
&\mathbf{w} = \text{ARG min}\left(y - \sum_{k=1}^{\text{n}} w_k \hat{y}_k\right)^2
\end{aligned}
\tag{1}
$$

where y is the true value of the property, $\hat{y}_k$ is the property prediction from PLS model developed on the kth interval, w_k is the weight of PLS model developed for the kth interval and n is the number of intervals. w_k can be obtained by cross-validation performed on individual interval model using the calibration set based on the following equation [35]:

$$
w_k = \frac{s_k^2}{\sum_{k=1}^{n} s_k^2}
\tag{2}
$$

where s_k is the reciprocal of the cross-validation error of PLS model developed on the kth interval.

2.3. Consensus modeling for regression

Consensus modeling combines multiple member models to produce a consensus predicted result. Consensus modeling includes two steps: one is how to establish member models and the other is how to design the consensus strategy. There does not exist any universal consensus strategy valid for all types of datasets. How to design appropriate consensus strategy for a specific dataset is the key problem.

The consensus strategy presented in this paper is for regression models and the corresponding consensus modeling method is presented below:

$$
\begin{aligned}
&h^*(h_1(x), h_2(x), \cdots, h_n(x)) = \sum_{k=1}^{n} w_k h_k(x) \\
&\mathbf{w} = \text{ARG min}\left(\sum_{k=1}^{n} w_k^2 \sigma_k^2 + 2\sum_{i=1}^{n}\sum_{k>i}^{n} w_i w_k r_{ik} \sigma_i \sigma_k\right) \\
&s.t\begin{cases} 0 \le w_k \le 1 \\ \sum_{k=1}^{n} w_k = 1, k \in [1, n] \end{cases}
\end{aligned}
\tag{3}
$$

where $\mathrm{h}_k(x)$ is the kth member model that is a regress model, e_k is the random error of $\mathrm{h}_k(x)$, σ_k is variance of e_k, r_{ik} is the correlation coefficient between e_i and e_k and n is the number of member models.

Suppose e_k follows the normal distribution $N(0, \sigma_k)$. e_k represents the ignored random factors in the kth regression model. When these random factors are independent of each other, the overall error e_k is approximate to normal distribution based on central limit theorem. Similarly, e, the random error of consensus model, also follows normal distribution $N(0, \sigma)$. Thus, we can reach the following conclusion:

$$
E\left(e^2\right) = \sum_{k=1}^{n} w_k^2 \sigma_k^2 + 2\sum_{i=1}^{n}\sum_{k>i}^{n} w_i w_k r_{ik} \sigma_i \sigma_k.
\tag{4}
$$

The proof of Eq. (4) is provided in the Appendix A. According to Eqs. (3) and (4), it can be seen that the physical interpretation of optimizing weight coefficient is to minimize the value of $\mathrm{E}(e^2)$, which is the theoretical consensus model error under the constraint of $0 \le w_k \le 1$. The constraint can enhance the generalization ability of consensus model [36]. With limited samples, the parameters in Eq. (3) are estimates.

If the impacts of error correlation in member models are not considered ($r_{ik}\sigma_i\sigma_k = 0$), then the weight coefficient optimization can be reached using the below:

$$
\begin{aligned}
&\mathbf{w} = \text{ARG min}\left(\sum_{k=1}^{n} w_k^2 \sigma_k^2\right) \\
&s.t\begin{cases} 0 \le w_k \le 1 \\ \sum_{k=1}^{n} w_k = 1. \end{cases}
\end{aligned}
\tag{5}
$$

The solution of Eq. (5) is given by

$$
w_k = \frac{1 \big/ \sigma_k^2}{\sum_{i=1}^{n} 1 \big/ \sigma_i^2}
\tag{6}
$$

and the proof is given in the Appendix A.