ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0014

Analysis of Consensus Model based on Variable Selection in Near-infrared Spectral Data

Xueli Chen[1, a], Xinyi Xie[1, b], Yongjie Lai[2, c]

[1] School of Intelligent Manufacturing, Wenzhou Polytechnic, Wenzhou, 325035, China

[2] Wenzhou Minshang Bank, Wenzhou, 325035, China

[a]1286346780@qq.com, [b]310002650@qq.com, [c]347408872@qq.com

Abstract

This paper proposes a spectral consensus regression model based on variable selection, SOM neural network was used to cluster variables. This method uses SOM's advantages in topology preserving and anti-noise to select variables, and then uses consensus method to correct and analyze NIR spectral data.The modeling method is to cluster variables of the full spectrum data by SOM algorithm, and then establish a series of regression models by PLS algorithm based on the results of variable clustering, and select the member model with the best predictive performance. The weight coefficient of each member model is obtained by using the corresponding residuals of each member model, and a consensus regression model based on variable clustering is established.In order to observe the modeling effect of the consensus model, this paper establishes the ordinary PLS model, member model (SOM-PLS) and consensus model (C-SOM-PLS) respectively, and then uses these three models to predict unknown samples. The results show that the prediction accuracy of the modeling after variable selection is higher than that of the ordinary PLS model.Using a variety of model consensus can not only improve the prediction accuracy of the model, but also enhance the stability of the model.

Keywords

SOM Variable Clustering; Consensus Model; Member Model; Quantitative Analysis; Variable Selection.

1. Introduction

In recent years, near-infrared spectroscopy (NIR) has been widely used in agriculture, petrochemical and medical industries due to its advantages of simple operation, high sensitivity, low cost and green environmental protection.Building a stable and reliable model is crucial for studying and predicting the concentration of unknown samples.Multivariate correction analysis is the core method of spectral quantitative analysis. Its main goal is to build a mathematical relationship between the spectral matrix and the reference matrix of the model sample set, so as to predict the concentration of new samples.At present, a variety of multivariate correction methods are widely used, including multiple linear regression (MLR) [1], principal component regression (PLS) [2], partial least squares regression (PLSR) [3-4] and other linear correction methods. Nonlinear correction methods such as BP-neural network [5] and support vector machine (SVM) [6-7] are also involved.When researchers apply the above methods to build models, they typically divide the global data into two stages: the training set and the test set. Based on the training set, the data is connected with the components to build a single regression model. Then the regression model performance evaluation method is used to select the model with the best prediction effect and use it to predict the unknown sample.Literature [8] proposes a cluster-based method to select variables for multivariate correction analysis of spectral

ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0014

data.Although this method has achieved good results in predicting unknown maize NIR spectral data, its limitation is that it can only fit a single pattern in the data. There are many and complex NIR spectral modes, and it is difficult to capture these modes with a single model, which may cause information loss and reduce the prediction accuracy and stability of the regression model.

This paper proposes a consensus modeling method of multiple regression models based on clustering algorithm, including K-means, AHC and other clustering algorithms.The K-means algorithm is simple and easy to understand, but the clustering result is affected by the selection of the initial cluster center. If the initial cluster center is improperly selected, the algorithm may fall into the local optimal solution, and the best clustering result cannot be obtained. Although the clustering result of AHC algorithm is relatively fine, but the efficiency is low and it is difficult to find the best partition, so this algorithm is not selected.SOM algorithm is chosen in this paper, which can realize isotonic mapping from high-dimensional spectral data to two-dimensional plane space. It has the advantages of strong anti-noise interference ability, high clustering efficiency, good effect and meeting the requirements of processing large amounts of data.

The present research shows that the method of multi-model consensus data modeling can improve the prediction accuracy and robustness of single model. This method constructs multiple member models in a certain way and combines the model's predictions for unknown samples in an efficient way to form a more accurate and reliable consensus result. Thus, a composite model is created, which has higher prediction accuracy and stronger robustness, so as to make up for the deficiency of single model modeling.

The modeling method proposed in this paper first uses SOM clustering algorithm to screen variables, so that similar variables are gathered together, and Duplex algorithm divides the whole NIR spectral data into three stages: training set, verification set and test set.(1) Construct a series of member regression models using training sets; (2) The error corresponding to the best prediction performance of the model is selected through the verification set as the weight of the consensus model; (3) The weighted summation method is used to combine the predicted values of member models for unknown samples to form a consensus result.In order to verify the prediction ability of consensus model [9], PLS, SOM-PLS and C-SOM-PLS were respectively used in this paper to model NIR spectral data. The results show that compared with the other two models, the stability and prediction accuracy of C-SOM-PLS are greatly improved.

2. Theory and Algorithm

2.1 SOM Clustering Algorithm

SOM neural network [10] is composed of input layer and competition layer feedforward neural network, which can identify the environment and automatically cluster, and is often used in cluster analysis and data structure analysis.In the practical application of SOM network, the minimum Euclidean distance between the sample and the output neuron is selected as the winning neuron output. In the process of learning, the learning rate of weight modification and neuron neighborhood are constantly reduced, so that similar neurons gradually gather and output. After network training, neurons are divided into different regions that have their own characteristics for the input model. SOM network competitive learning algorithm is as follows:

Step1:The network weight W is initialized. Wij is a small random number.

Step2:Euclidean distance is chosen as the method of calculating distance.

For example, the distance d_j between the input vector $X=(x_1,x_2\cdots,x_n)$ and the competing layer neurons j .

$$d_j=\left|\sum_{i=1}^{m}(x_i-w_{ij})^2\right| \quad j=1,2\cdots,n$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0014

Step3: The competing layer neuron c with the smallest $\min\{d_j\}$ output vector X distance is used as the optimal neuron output.

Step4: Weight W adjustment.

Adjust node c and the node weight coefficients included in its neighborhood $N_c(t)$. Namely $N_c(t) = (t \mid find(norm(pos_t, pls_c) < r)\ t = 1, 2, \cdots, n$.

$$w_{ij} = w_{ij} + \eta(X_i - w_{ij})$$

Where, pos_t, pls_c are the positions of neurons c and t respectively; $norm$ calculates the Euclidean distance between two neurons; r is the radius of the domain; η is the learning rate.

Step5: Check whether the algorithm is complete. If not, go back to Step 2.

2.2 Partial Least Squares Regression (PLSR)

PLSR, first proposed by Wood and Abano in 1983, belongs to a new multivariate statistical data analysis method. The algorithm covers the basic skills of multiple linear regression analysis, canonical correlation analysis and principal component analysis, namely: partial least square regression ≈ multiple linear regression analysis + canonical correlation analysis + principal component analysis.The key technique to transition from ordinary least squares regression to partial least squares regression is to use principal component analysis to extract components. PLSR can not only summarize the information of independent variable system, but also explain the dependent variable well. The basic principle is as follows:

It is assumed that n sample points and p variables constitute the NIR spectral data matrix X, and the sample concentration matrix is Y. X and Y are decomposed to obtain equations (1) and (2).

$$X = TP^T + E_X \tag{1}$$

$$Y = UQ^T + E_Y \tag{2}$$

Where T and U are the score matrices of X and Y respectively, P^T and Q^T are the load matrices of X and Y respectively, and E_X E_Y are the PLS fitting residual matrices of X and Y respectively.

Then, the regression model between T and U is established:

$$U = TB \tag{3}$$

$$B = (T^T T)^{-1} T^T Y \tag{4}$$

When testing the new sample X_{new}, the score matrix T_{new} of the new sample matrix is first solved by combining P, and then the prediction is made according to equation (5).

$$Y = T_{new} BQ \tag{5}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0014

2.3 Model Consensus Rule

There are many methods for consensus modeling. The purpose of consensus modeling is to build a prediction of a multi-member model by mining information from different levels of the training set, so as to form a consensus result.Consensus modeling can reduce the impact of background noise and volatility on spectral data, and its advantage is that it has no bias to the quality and size of the data set, and it has stronger robustness to the unbalanced data set, and can make full use of redundant information.

For the consensus model analysis in this paper, we use the following method, assuming the training set $S=\{(y_n,x_n),n=1,...,N\}$, where x_n is the input vector and y_n is the output vector.Through the training set, K member models $f_1(x),f_2(x),...,f_k(x)$ are formed. In the consensus modeling, these K models need to be combined in a weighted form to obtain the required consensus model $f(x)$. The consensus strategy proposed in this paper is as follows:

$$f(x)=\sum_{k=1}^{n} w_k f_k(x) \tag{6}$$

$$w=ARG\min(\sum_{k=1}^{n} w_k^2\sigma_k^2+2\sum_{i=1}^{n}\sum_{k>i}^{n} w_i w_k r_{ik}\sigma_i\sigma_k)$$
$$s.t\begin{cases}0\le w_k\le 1\\ \sum_{k=1}^{n} w_k=1, k\in[1,n]\end{cases} \tag{7}$$

Where $f_k(x)$ is the predicted value of K member models, assuming e_k is the random error of $f_k(x)$, then σ_k is the variance of e_k, and r_{ik} is the correlation coefficient of e_i and e_k. For detailed information on the consensus model, refer to literature[12].

2.4 Consensus Model C-SOM-PLS

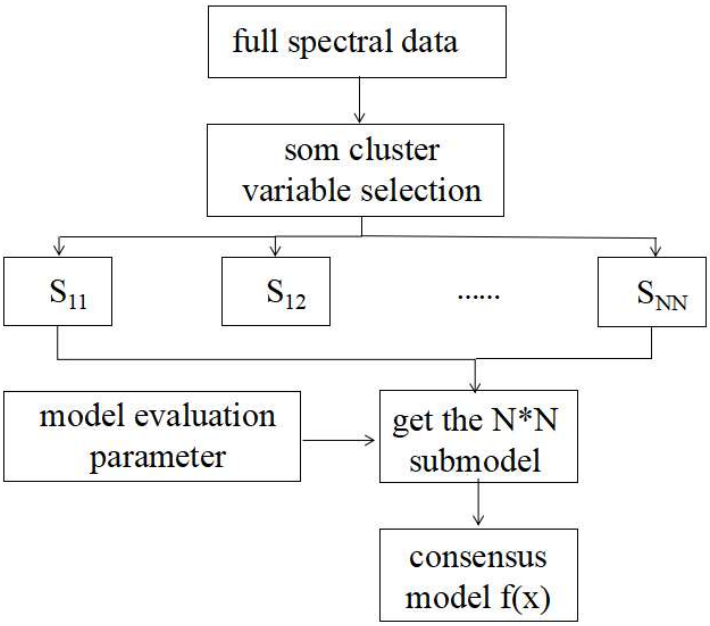

Figure 1. Implementation process of the consensus model

ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0014

The consensus model first uses SOM neural network clustering algorithm to cluster the full spectrum data into different sub-sample data according to the similarity of variables, and then establishes PLS member model according to the sub-data samples, that is, SOM-PLS.SOM clustering algorithm is the selection of sample variables in this paper. Some variables contain more useful information, while others contain more noise or useless information. Therefore, the prediction performance and stability of the model established by each subsample data are different.If only the member model with good prediction is selected, some sample information will inevitably be lost. In order to fully mine the useful information contained in the sample, this paper proposes a consensus model, which combines the member models according to the consensus rules mentioned above to form a consensus model (C-SOM-PLS).Figure 1 shows the implementation process of the algorithm.

3. Experiment

3.1 Experimental Data

The corn NIR data used in this paper are open source data provided by Eigenvector. The dataset consisted of three different NIR instruments (m5,mp5 and mp6), each used to test 80 sample datasets with a spectral sweep range of 1100 to 2498nm and a sweep interval of 2nm.The four corn reference attribute values are the content of water, oil, protein, and starch. In this paper, the spectral data measured by m5 instrument was used to correct and analyze four kinds of corn properties.

3.2 Software

All algorithms involved in this paper were implemented in the MatlabR2012a software platform, and the clustering algorithm was implemented in the kohonen_cpann_toolbox.

4. Results and Discussion

4.1 The Result of SOM Neural Networks Clustering

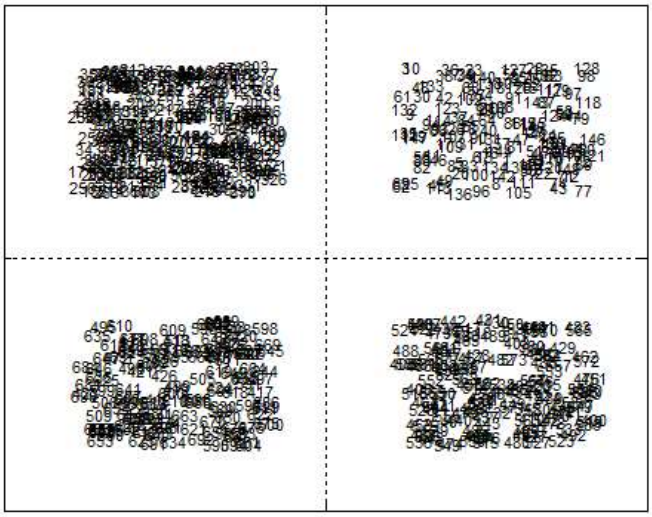

Figure 2. SOM variable clustering results

The purpose of SOM neural network clustering is to classify similar or similar variables into the (q*q) neuron matrix, so the number of variables clustering through SOM network is q, here we call the number of subsamples q.This paper studies the prediction effect of the consensus model with the number of subsamples ranging from 2 to 10, and finds that when the number of subsamples is 2, the prediction effect of the consensus model is good or there is little difference. Figure 2 shows that a 2*2 array is generated by variable clustering of data samples generated by m5 instrument. The number of each subsample is identified by s, and i and j are rows and columns of SOM clustering results respectively. For example, the first subsample is s1 and the second subsample is s2.It can be seen from the figure that s1 has the largest number of variables, s21 and s22 have several subsample

ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0014

variables, and s12 has the least number of variables. Because SOM neural network clustering is calculated according to Euclidean distance, the subsample of variables indicates that the variables are more similar, and the reverse indicates that the variables are less similar.

4.2 SOM-PLS Analysis

Figure 3 (a, b, c, d) shows the validation root mean square errors (RMSEV) and prediction root mean square errors (RMSEP) of the member models for the four attributes of corn (water, oil, protein, and starch), and the weight coefficient (Wt) obtained by the member models according to the consensus rule.Figure 3 shows the following findings: The RMSEV and RMSEP values of PLS modeling for the sub-data set obtained by SOM variable clustering are different, because SOM clustering is a cluster of similar variables, and the sub-data set with good modeling and prediction effect indicates that its variables contain more information reflecting the sample attributes; conversely, its variable information contains more noise or useless information.The same sub-data set also has different predictive effects on the modeling of the four attributes. For example, in Figure 3 (a), 3 (b), 3 (c), the fourth member model has the best predictive effect, but in 3 (d), the first member model has the best predictive effect, indicating that different variable information has different performance effects on the modeling of different attributes.Wt is the weight of the member model. It can be seen from the figure that the weight of each member is very different, and its weight represents the size of the information contained by the member model in the consensus model. The larger the weight, the more information contained, the greater the contribution of the member model to the consensus model.

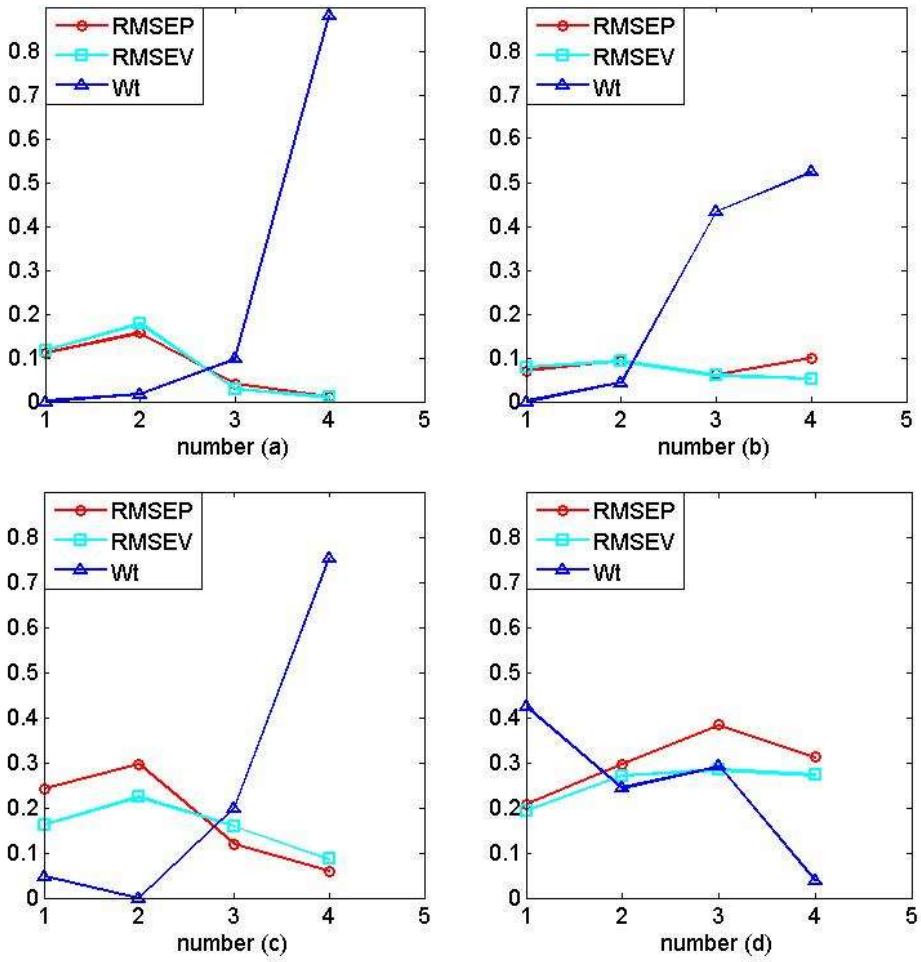

Figure 3. RMSEP, RMSEV, and Wt values of SOM-PLS

ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0014

4.3 PLS, SOM-PLS and C-SOM-PLS Analysis

In order to evaluate the effect of C-SOM-PLS model, RMSEV and RMSEP values obtained by full-spectrum PLS, optimal SOM-PLS and C-SOM-PLS algorithms for modeling corn water, oil, protein and starch are shown in Table 1. As can be seen from the table, the optimal member model (SOM-PLS) obtained by variable selection, due to the reduction of noise and useless information variables, has achieved satisfactory results in modeling water, protein and starch compared with the full-spectrum PLS algorithm, both in modeling and in predicting unknown samples,The prediction accuracy of unknown samples is increased by 49.5%, 42.1% and 19.7% respectively. For oil prediction, SOM-PLS is not as good as full-spectral PLS modeling, mainly because the sub-data set based on SOM clustering only contains part of the sample information, which is a sample composed of similar variables, and the variable information of this part is not obvious for oil prediction.C-SOM-PLS is obtained by weighted summation of each member model. Compared with PLS algorithm, the prediction accuracy of corn moisture, protein and starch is increased by 50.0%, 46.5% and 24.0% respectively. However, the prediction effect of oil is still not as good as that of PLS. The main reason is that the information variables with obvious performance for oil modeling include both similar variables and dissimilar ones. Therefore, all-optical modeling is better, but the prediction accuracy of oil is improved compared with SOM-PLS.Compared with SOM-PLS, the model prediction ability of C-SOM-PLS has been improved because the consensus model combines all the member models to extract useful information from all levels of the sample data.Compared with the member model, the consensus model not only improves the prediction accuracy of unknown samples, but also improves the stability of the model because it combines all the member models, so that one input can get a stable output.

Table 1. RMSEV and RMSEP of water, oil, protein and starch in corn

Methods	Moisture		Oil		Protein		Starch	
	RMSEV	RMSEP	RMSEV	RMSEP	RMSEV	RMSEP	RMSEV	RMSEP
PLS	0.0218	0.0208	0.0620	0.0561	0.1344	0.1035	0.2561	0.2598
SOM-PLS	0.0093	0.0105	0.0520	0.0998	0.0858	0.0599	0.1937	0.2087
C-SOP-PLS	0.0075	0.0104	0.0451	0.0704	0.0749	0.0554	0.1532	0.1974

5. Conclusion

In this paper, the consensus model strategy based on variable selection is studied for correction analysis of near-infrared spectrum, SOM clustering algorithm is used to extract sample variables, and a consensus model (C-SOM-PLS) combining SOM-PLS and consensus strategy is proposed to fully mine useful variable information in sample data.By comparing PLS, SOM-PLS and C-SOM-PLS, it is found that variable selection is beneficial to model modeling, especially the data with fewer samples and more variables, which not only reduces the high-dimensional data to low-dimensional data, but also improves the prediction accuracy of the model.Consensus modeling makes up for the shortcomings of member models, and combines all member models with different forecasting effects to extract useful information from each member model, which not only improves the prediction accuracy of the model, but also improves the stability of the model.

Acknowledgments

This work is financially supported by the 2022 project of Wenzhou Polytechnic of Zhejiang Province No.WZYzd202205;2022 Project of Wenzhou Polytechnic of Zhejiang Province No.WZY2022011; 2021 Project of Wenzhou Polytechnic of Zhejiang Province No.WZYSZKC2102.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202403_10(3).0014

References

[1] Zhang G C , Li Z , Yan X M ,et al.Rapid Analysis of Apple Leaf Nitrogen using Near Infrared Spectroscopy and Multiple Linear Regression[J].Communications in Soil Science and Plant Analysis, 2012, 43(13):1768-1772.DOI:10.1080/00103624.2012.684824.

[2] Weakley A T , Warwick P C T , Bitterwolf T E ,et al.Multivariate Analysis of Micro-Raman Spectra of Thermoplastic Polyurethane Blends Using Principal Component Analysis and Principal Component Regression[J].Applied Spectroscopy, 2012, 66(11):1269-1278.DOI:10.1366/12-06588.

[3] Wang Huiwen. Partial Least Squares Regression Method and Its Application [M]. National Defense Industry Press, 1999.

[4] Reza M T , Becker W , Sachsenheimer K ,et al.Hydrothermal carbonization (HTC): Near infrared spectroscopy and partial least-squares regression for determination of selective components in HTC solid and liquid products derived from maize silage[J].Bioresource Technology, 2014, 161(3):91-101.DOI:10.1016/j.biortech.2014.03.008.

[5] Zheng L H , Li M Z , Pan L ,et al.[Estimation of soil organic matter and soil total nitrogen based on NIR spectroscopy and BP neural network].[J].Spectroscopy & Spectral Analysis, 2008, 28(5):1160-1164. DOI:10.1016/j.sab.2008.04.001.

[6] Hong-Yan,Zou,Hai-Long,et al.Variable-weighted least-squares support vector machine for multivariate spectral analysis[J].Talanta, 2010, 80(5):1698-1701.DOI:10.1016/j.talanta.2009.10.009.

[7] Ding Shifei, Qi Bingjuan, Tan Hongyan. Support vector machine (SVM) theory and algorithm research review [J]. Journal of university of electronic science and technology, 2011, 40 (1) : 9. DOI: 10.3969 / j.i SSN. 1001-0548.2011.01.001.

[8] Luo Ke-Gang. Research on text clustering based on Self-organizing mapping [D]. Harbin Institute of Technology.

[9] Lai Yongjie. Application of consensus model based on SOM cluster variable selection method to near-infrared spectral data [D]. Wenzhou University,2018.

[10] Gomez-Carracedo,MP,Andrade,et al.Combining Kohonen neural networks and variable selection by classification trees to cluster road soil samples[J].CHEMOMETR INTELL LAB, 2010, 2010,102(1)(-): 20-34.DOI:10.1016/j.chemolab.2010.03.002.

[11] Yuan,Mingshun,Ji,et al.Using consensus interval partial least square in near infrared spectra analysis[J]. Chemometrics & Intelligent Laboratory Systems, 2015.DOI:10.1016/j.chemolab.2015.03.008.