Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 196 (2018) 131–140

Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

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

Using an optimal CC-PLSR-RBFNN model and NIR spectroscopy for the starch content determination in corn

Hao Jiang, Jiangang Lu *

State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 29 March 2017
Received in revised form 17 October 2017
Accepted 5 February 2018
Available online 6 February 2018

Keywords:
NIR spectroscopy
Corn starch
Correlation coefficient method
PLSR
RBF
Neural network

ABSTRACT

Corn starch is an important material which has been traditionally used in the fields of food and chemical industry. In order to enhance the rapidness and reliability of the determination for starch content in corn, a methodology is proposed in this work, using an optimal CC-PLSR-RBFNN calibration model and near-infrared (NIR) spectroscopy. The proposed model was developed based on the optimal selection of crucial parameters and the combination of correlation coefficient method (CC), partial least squares regression (PLSR) and radial basis function neural network (RBFNN). To test the performance of the model, a standard NIR spectroscopy data set was introduced, containing spectral information and chemical reference measurements of 80 corn samples. For comparison, several other models based on the identical data set were also briefly discussed. In this process, the root mean square error of prediction (RMSEP) and coefficient of determination (Rp^2) in the prediction set were used to make evaluations. As a result, the proposed model presented the best predictive performance with the smallest RMSEP (0.0497%) and the highest Rp^2 (0.9968). Therefore, the proposed method combining NIR spectroscopy with the optimal CC-PLSR-RBFNN model can be helpful to determine starch content in corn.

1. Introduction

In recent years, near-infrared (NIR) spectroscopy has been developing rapidly with the improvement of computer performance, NIR instruments and chemometrics. As a fast, accurate and non-destructive method for measuring, NIR spectroscopy is widely applied in the process of both off-line and on-line analysis. Nowadays, it has covered various fields including agriculture, food, chemical industry and biologic science, generating remarkable social and economic benefit [1,2].

Corn starch is an important food and chemical material. To facilitate the breeding of high-starch corn, increase corn yields and promote the added value, it's significant to establish rapid and reliable analytical methods for the determination of starch content in corn [3,4]. Generally, traditional chemical methods are time-consuming, expensive and not environment-friendly, which cannot meet the requirement of the large-scale applications of online measurements in agriculture and industry [5]. To avoid the shortcomings above, NIR spectroscopy is an alternative method. Since the NIR spectrum is formed from the signals due to the molecular vibrations of hydrogen-containing functional groups [6], it's quite suitable to use NIR spectroscopy to determinate starch content.

On the other hand, interpretation on the NIR spectra is difficult, because the bands of spectra are typically broad and spectra of different samples overlap closely. A feasible solution is to take methods of chemometrics and establish multivariate statistical calibration models, correlating spectral data with the content values obtained by chemical reference methods [7]. With the help of calibration models, characteristics of unknown samples can be exactly predicted.

In the procedure above, the robustness and accuracy of the prediction strongly depend on the accuracy of the used data and performance of calibration methods. The former part is related to the sampling process, spectral quality and accuracy of reference methods. The later part mainly includes the selection and optimization of calibration models.

Typically, linear calibration models include multiple linear regression, principal component regression (PCR) and partial least squares regression (PLSR). Nonlinear calibration models mainly include nonlinear PLSR, support vector machine (SVM) and neural network (NN) [8–11]. Among all these models, PLSR possesses especially good performance. It retains the ability of PCR to denoise and eliminate the collinearity of variables. Meanwhile, it takes the relation between dependent and independent variables into account, which improves its predictive capability significantly [12].

However, PLSR is a linear calibration method essentially. Although the number of principal components extracted can be increased to enable models partly approximating to the nonlinear portion, it is ultimately restricted by the requirement of robustness. As a result, nonlinear regression residual of PLSR models cannot be removed. But at the same time, nonlinear models are commonly less effective in the modeling of the linear parts. Thus, some models integrating PLSR with

* Corresponding author.
E-mail address: lujg@zju.edu.cn (J. Lu).

https://doi.org/10.1016/j.saa.2018.02.017
1386-1425/

Table 1
Statistical details of samples' chemical reference values.

Data sets	Range (%)	Mean (%)	Median (%)	Standard deviation (%)
Calibration set	62.884–65.903	64.678	64.772	0.808
Prediction set	62.826–66.472	64.749	64.889	0.876

nonlinear calibration methods were developed in former researches to synthesize both of the merits [13–15]. To be mentioned, radial basis function neural network (RBFNN) is a promising nonlinear calibration method. As an emerging local approximation network, RBFNN can approximate arbitrary continuous function with arbitrary precision and avoid the problem of local minimum in some other networks, such as backpropagation neural network (BPNN) [16,17].

In this research, a methodology is proposed for the starch content determination in corn, using an optimal CC-PLSR-RBFNN model and NIR spectroscopy. The proposed model was formed based on the combination of PLSR and RBFNN. Meanwhile, correlation coefficient method (CC) was used here to select informative wavelengths, and crucial parameters in the process of modeling were optimized to improve the predictive performance. For comparison, several other models were also briefly discussed in this paper. The performance of all the mentioned models was evaluated through the root mean square error of prediction (RMSEP) and coefficient of determination (Rp^2) in the prediction data set. In general, smaller RMSEP values and larger Rp^2s indicate the superior ability of calibration models for accurate determinations.

2. Materials and Methods

2.1. Data Description

A published data set (http://www.eigenvector.com/Data/Corn/index.html) was introduced here, aiming to evaluate the predictive ability of the proposed method more comprehensively by comparing with previous researches. This data set consists of NIR spectra obtained from 80 corn samples, and all the spectra were measured from 1100 to 2498 nm at 2 nm intervals (700 channels) on three different spectrometers. Besides, the reference content of samples was also involved, including moisture, oil, protein and starch.

In this work, only the NIR spectra measured on the m5 spectrometer and the corresponding starch content were used. 80 samples in the data set were randomly divided into calibration and prediction sets at a ratio of 3:1. Statistical details about the samples' chemical reference values are descripted in Table 1.

2.2. Data Preprocessing

In the raw NIR spectra, except for the abundant information about the functional groups in samples, irrelevant information is also contained, such as background noise and baseline drift [18,19]. Therefore, preprocessing spectra properly to extract effective information is meaningful for the development of accurate and reliable calibration models. Typical preprocessing methods of NIR spectra contain background and baseline corrections, normalization and smoothing [20]. Particularly, Savitzky-Golay (S-G) smoothing algorithm combined with derivative calculations shows great performance in the elimination of irrelevant information in the raw NIR spectra, and has been applied widely [21–24].

In this research, all the raw spectra were preprocessed before calibrations. Specifically, S-G smoothing with a third degree polynomial was used for denoising, and the first-order numerical derivative was made to correct the baseline drift.

2.3. Parameters Optimization on the Basic PLSR Model

After the spectral preprocessing, PLSR was applied in the calibration set to form a basic model, correlating data of the full NIR spectra with the reference starch content. In the process of preprocessing and calibration, the window width of S-G smoothing and the number of principal components extracted are especially crucial parameters. For instance, a large window of S-G smoothing may cause the spectra to distort, while a narrow window often results in a weak effect of denoising [25]. Meanwhile, extracting too many principal components leads to overfitting and bad robustness in the calibration models, while an insufficient extraction leads a greater prediction error due to the loss of effective information [26]. Moreover, these two parameters are mutual coupling in the effect on the predictive performance of models. Therefore, they need to be optimized together in the parameter space, according to the cost values in the calibration set. Cost values were calculated based on the root mean square error of cross-validation (RMSECV) and extra penalty terms:

$$\mathrm{Cost}(p,q) = \mathrm{RMSECV}(p,q) + \alpha \times p + \beta \times q \tag{1}$$

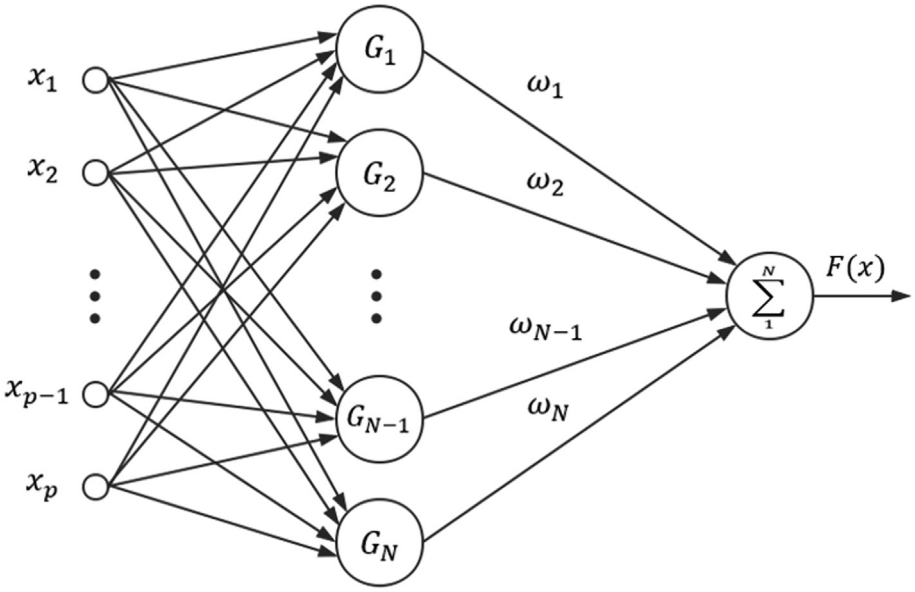

Fig. 1. Typical structure of RBFNN.