Analytica Chimica Acta 1009 (2018) 20–26

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Robust boosting neural networks with random weights for multivariate calibration of complex samples

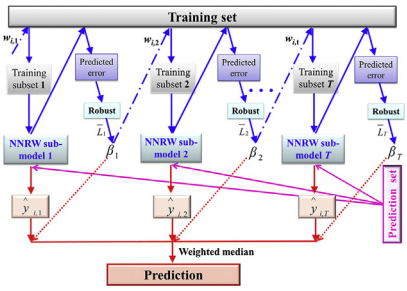
Xihui Bian [a, b, *], Pengyao Diwu [a, b], Caixia Zhang [a, b], Ligang Lin [a], Guohui Chen [b], Xiaoyao Tan [a, b], Yugao Guo [b], Bowen Cheng [a]

[a] State Key Laboratory of Separation Membranes and Membrane Processes, Tianjin Polytechnic University, Tianjin, 300387, PR China
[b] School of Environmental and Chemical Engineering, Tianjin Polytechnic University, Tianjin, 300387, PR China

HIGHLIGHTS

- A novel ensemble method named as robust boosting neural networks with random weights (RBNNRW) is proposed.
- Hampel robust step is introduced for the method.
- The method has marked superiorities in predictive accuracy and stability especially when outliers exist.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 21 March 2017
Received in revised form
10 December 2017
Accepted 5 January 2018
Available online 3 February 2018

Keywords:
Ensemble modeling
Boosting
Neural networks with random weights
Extreme learning machine
Outlier
Complex samples

ABSTRACT

Neural networks with random weights (NNRW) has been used for regression due to its excellent performance. However, NNRW is sensitive to outliers and unstable to some extent in dealing with the real-world complex samples. To overcome these drawbacks, a new method called robust boosting NNRW (RBNNRW) is proposed by integrating a robust version of boosting with NNRW. The method builds a large number of NNRW sub-models sequentially by robustly reweighted sampling from the original training set and then aggregates these predictions by weighted median. The performance of RBNNRW is tested with three spectral datasets of wheat, light gas oil and diesel fuel samples. As comparisons to RBNNRW, the conventional PLS, NNRW and boosting NNRW (BNNRW) have also been investigated. The results demonstrate that the introduction of robust boosting greatly enhances the stability and accuracy of NNRW. Moreover, RBNNRW is superior to BNNRW particularly when outliers exist.

1. Introduction

Analysis of complex samples is a great challenge in laboratories and industries due to their complex composition and the unavoidable outliers [1–3]. Therefore, rapid and robust analytical techniques for reliable quantification of components in complex samples are needed. Spectroscopic methods coupled with multivariate calibration [4–6] have been widely used in the analysis of agricultural, petrochemical, medical and other products. Construction of high quality multivariate calibration models is the key for quantitative analysis of spectroscopy. Among multivariate

* Corresponding author. State Key Laboratory of Separation Membranes and Membrane Processes, School of Environmental and Chemical Engineering, Tianjin Polytechnic University, Tianjin, 300387, PR China.
E-mail address: bianxihui@mail.nankai.edu.cn (X. Bian).

https://doi.org/10.1016/j.aca.2018.01.013
0003-2670/

calibration methods, partial least squares (PLS) [4,7] is one of the most popular methods to model a target when there are a large number of variables, and those variables are highly correlated or even collinear. Noting that sample spectra and targets may not follow a linear relationship, nonlinear modeling techniques such as artificial neural network (ANN) [8–10] should be in place to produce more accurate prediction results. Although feed-forward neural networks with error back-propagation have been widely used in resolving nonlinear regression problems, it suffers from local minima, sensitivity of parameter option and slow learning rate since all the parameters of the networks need to be tuned iteratively [11].

A special method for single-hidden layer feed forward neural networks (SLFNs) named as feed forward neural networks with random weights (NNRW) was proposed by Schmidt et al. [12]. They gave some experimental evidence showing that the weights of neural networks are of less important and could be generated randomly without adjusting iteratively [12]. Huang et al. [13] proved theoretically that the input weights and hidden layer biases can be generated randomly and popularized [14] this method as extreme learning machine (ELM). We mentioned the method as NNRW in this study since it is the origin of ELM. NNRW has attracted an increasing attention in quantitative modeling of complex samples [8,15] due to its advantages of simple structure, high learning speed and good generalization performance.

The random initialization of weights and hidden layer biases increases the learning speed and reduces optimization parameters of NNRW. However, this random initialization also makes NNRW unstable in practice [16–18], which means different runs of the NNRW model will lead to fluctuation in the predictive results. In addition, the output layer weights of NNRW are computed by a simple batch learning scheme, such as the standard ordinary least squares (OLS) method [19]. This makes NNRW tend to suffer from outliers in the training set [17,20], which originate from the recording mistakes or exceptional circumstances and are usually unavoidable in the actual complex samples [21–23]. In brief, NNRW is unstable and sensitive to outliers. Therefore, it is highly demanding to develop new approaches for improving the stability and robustness of NNRW.

In recent years, many efforts have been devoted to improving the performance of NNRW. Some robust versions of NNRW have been developed [17,24] to improve the predictive accuracy with the presence of outliers in the training set. Although the robust ability of NNRW is enhanced greatly, the problem of instability still exists. It has been discovered that an ensemble of sub-models is one of the best ways to improve both the accuracy and stability of a single model [4,25]. The most popular ensemble strategies are bagging, boosting and random forest [26]. These ensemble strategies have been widely used to improve the performance of single multivariate calibration models such as PLS [27–32], ANN [9,10], support vector regression (SVR) [33–35]. Recently, ensemble NNRW models have also been developed to improve the stability of NNRW [15,18]. However, few studies improve the stability and robustness of NNRW simultaneously [36].

As one of the most prominent ensemble strategies [26,37], boosting has attracted increasing interest in chemometrics [27–32,34,35]. By combining a series of rough and inaccurate sub-models, boosting can obtain an accurate prediction. Such a series of sub-models is developed by using the training subsets selected from the original training set according to the distribution of the sampling weights. For the first cycle, all the samples in the training set are given the same sampling weights. In the following cycles, the samples with larger predictive errors are given higher weights, implying that the worse predicted samples are more likely to be picked up into the training subset for the subsequent iteration. Such a sampling strategy will degrade or even ruin the performance of boosting, especially when outliers are present [28]. Recently, Shao et al. [3] and Zhou et al. [26,28] have designed robust version of boosting by introducing a robust step before renovating the weights to improve the performance of PLS and regression tree, respectively. The robust step is carried out by depressing the weights of the samples larger than a certain value to prevent the samples with large errors from being selected in the following training subset.

To improve the stability and robustness of NNRW, a novel method named as robust boosting NNRW (RBNNRW) is proposed for multivariate calibration of complex samples. To assess the predictive ability of RBNNRW model, the conventional PLS, NNRW and boosting NNRW (BNNRW) have also been investigated with two datasets with priori known outliers and one neat dataset. Results show that robust boosting can improve the predictive accuracy and stability of NNRW greatly while keep most of the appealing properties of NNRW with the presence of outliers.

2. Experimental

Three spectral datasets were used to evaluate the proposed method. Dataset 1 consists of visible-near infrared (Vis-NIR) spectra and six properties of 884 wheat samples. The Vis-NIR spectra and the protein contents are used in this study. The spectra were scanned on a Foss Model 6500 over 1050 channels recorded in the wavelength range of 400–2498 nm with the digitization interval 2 nm. The reference values of protein contents were determined at the Grain Research Laboratory, Winnipeg. The dataset was contributed by P.C. Williams and can be downloaded freely from http://www.idrc-chambersburg.org/shootout2008.html. According to the description in the website, the samples Nos. 680 and 681 are two outliers.

Dataset 2 consists of ultraviolet (UV) spectra and four hydrocarbon contents of 115 light gas oil and diesel fuel samples. The UV spectra and monoaromatics contents are used in this study. The spectra were measured with Cary 3 UV-visible spectrophotometer (Varian Instruments, San Fernando, Calif.) over 572 channels recorded in the wavelength range 200–400 nm with the digitization interval 0.35 nm. The reference values of monoaromatics contents were measured with HP model G1205A supercritical fluid chromatography (Hewlett-Packard, Palo Alto, Calif.). The dataset was supplied by Wentzell et al. [38] and can be downloaded freely from http://myweb.dal.ca/pdwentze/downloads.html. According to the description in the website, the sample No. 115 is an outlier.

Dataset 3 consists of NIR spectra and six physical properties of 256 diesel fuel samples [39]. The NIR spectra and total aromatics contents are used in this study. The spectra were measured at Southwest Research Institute (SWRI) on a project sponsored by the U.S. Army. Each spectrum is composed of 401 variables recorded in the wavelength range 750–1550 nm with the digitization interval 2 nm. The reference values of the total aromatics contents were measured by the American Society of Testing and Materials (ASTM) standard method. The dataset was provided by SWRI, San Antonio, TX through Eigenvector Research, Inc. (Manson, Washington) and can be downloaded freely from http://www.eigenvector.com/Data/SWRI. From the description in the website, the dataset has been thoroughly vetted and no outlier exists.

Before calculation, the three datasets were divided into training, validation and prediction sets for model building, parameter optimization and performance validation, respectively. For the three datasets, the training sets described on the websites were used as the training sets and the original prediction sets on the websites were divided into validation sets and prediction sets by KS algorithm for this study. Furthermore, the reported or artificial outliers