CHINESE JOURNAL OF ANALYTICAL CHEMISTRY
Volume 46, Issue 1, January 2018
Online English edition of the Chinese language journal

Available online at www.sciencedirect.com
ScienceDirect

Cite this article as: Chinese J. Anal. Chem., 2018, 46(1): 136–142

RESEARCH PAPER

A Variable Selection Method of Near Infrared Spectroscopy Based on Automatic Weighting Variable Combination Population Analysis

ZHAO Huan, HUAN Ke-Wei*, SHI Xiao-Guang, ZHENG Feng, LIU Li-Ying, LIU Wei, ZHAO Chun-Ying
College of Science,Changchun University of Science and Technology, Changchun 130022, China

Abstract: Near-infrared spectroscopy (NIR) is widely used in food quantitative and qualitative analysis. Variable selection technique is a critical step of the spectrum modeling with the development of chemometrics. In this study, a novel variable selection strategy, automatic weighting variable combination population analysis (AWVCPA), is proposed. Firstly, binary matrix sampling (BMS) strategy, which provides each variable the same chance to be selected and generates different variable combinations, is used to produce a population of subsets to construct a population of sub-models. Then, the variable frequency (Fre) and partial least squares regression (Reg) , two kinds of information vector (IVs), are weighted to obtain the value of the contribution of each spectral variables, and the influence of two IVs of Rre and Reg is considered to each spectral variable. Finally, it uses the exponentially decreasing function (EDF) to remove the low contribution wavelengths so as to select the characteristic variables. In the case of near infrared spectra of beer and corn, yeast and oil concentration models based on partial least squares (PLS) of prediction are established. Compared with other variable selection methods, the research shows that AWVCPA is the best variable selection strategy in the same situation. It has 72.7% improvement comparing AWVCPA-PLS to PLS and the predicted root mean square error (RMSEP) decreases from 0.5348 to 0.1457 on beer dataset. Also it has 64.7% improvement comparing AWVCPA-PLS to PLS and the RMSEP decreases from 0.0702 to 0.0248 on corn dataset.

Key Words: Near infrared spectroscopy; Chemometrics; Variable selection; Automatic weighting variable combination population analysis; Information vector

1 Introduction

Variable selection techniques have become critical steps in the analysis of datasets with the development of near infrared spectroscopy and chemometrics, and the variable selection techniques can be used to improve the prediction performance of the prediction models, thus providing faster and more cost-effective predictions by reducing the dimensionality of spectral data, and providing a better understanding and interpretation of the underlying process that generated the data. However, in the face of the 'large p, small n' problem, that is the number of samples is much smaller than the number of variables, it becomes a non-deterministic polynomial-time (NP) hard optimization problem to find the optimal subset that satisfies the above three aspects[1–4]. The methods of variable-deletion with no information in common at home and abroad include uninformative variable elimination (UVE)[3], Monte Carlo based UVE (MC-UVE)[5], genetic algorithm (GA)[6–9] and so on. With the development of model population analysis (MPA), many new selection methods have been developed, such as random frog (RF)[10], competitive adaptive reweighted sampling (CARS)[11,12], iteratively retains informative variables (IRIV)[13], variable combination population analysis (VCPA)[14] and so on. However, researchers often use

Received 8 August 2017; accepted 8 October 2017
*Corresponding author. E-mail: huankewei@126.com
This work was supported by the Special Scientific Research Fund of Meteorological Public Welfare Profession of China (No. GYHY201406037), and the Doctoral Fund of Ministry of Education of China (No. 20112216110006).

DOI: 10.1016/S1872-2040(17)61065-X

information vectors (IVs)[15] to evaluate the importance of each variable. Common information vectors include partial least squares regression coefficient vectors (Reg), correlation coefficient vector (Cor), residual vectors (Res), vector of importance of projection variables (VIP), NET signal vector (NAS), signal-to-noise ratio vector (StN), covariance vector (Cov), select a specific vector (SR), predictive residual vectors (Ssr), variable Occurrence frequency vector (Fre), and covariance selection vectors (Covsel) and so on[16–21]. Although a large number of variable selection methods and information vectors have been proposed, each variable selection method only uses one of these information vectors as the importance of variables to judge the basis, thus ignoring other information vectors influence on the prediction models. Therefore it is easy to produce over fitting phenomenon of prediction models. In order to solve this problem, this paper presents a new method of variable selection for the first time-Automatic weighting variable combination population analysis (AWVCPA). This method combines MPA with IVs weighting for the first time. Also this variable selection method adopts the model population analysis (MPA) strategy for searching the optimal variable subset using root mean squares error of cross validation (RMSECV) as the objective function, and the results of the two information vectors of Fre and Reg are normalized and weighted. Then the final contribution value of each variable in these two information vectors are calculated considering the effect of the two information vectors on each spectral variable, as a result, the stability of the prediction model is improved. In this study, AWVCPA coupled with PLS was investigated through two real near infrared spectral datasets, beer and corn. It yielded better results and better selectivity when comparing to three well performing methods (GA, MC-UVE, RF, VCPA), and the experimental results showed that AWVCPA was a good variable selection method for multivariate calibration.

2 Experiment

2.1 Data sources

2.1.1 Beer dataset

The NIR dataset of beer[22] consisted of 60 samples, and it was recorded with a 30-mm quartz cell directly on the undiluted degassed beer, and collected at intervals of 2 nm within the wave numbers range of 1100–2250 nm. The spectra consisted of 576 data points in the NIR region 1100–2250 nm. Original extract concentration which indicated the substrate potential for the yeast to ferment alcohol was considered as property of interest. K-S method was used to divided the dataset into calibration set (40 samples) and independent test set (20 samples) by sorting the extract values. The calibration set was used to select the original variable, and the independent test set was used to verify the performance of the selected variable. NIR spectra of beer were displayed in Fig.1.

2.1.2 Corn dataset

The corn dataset was available in the website: http://www.eigenvector.com/data/Corn/index.html. The corn dataset contained 80 samples of corn measured on three different types of NIR spectrometers[14]. Each spectrum consisted of 700 data points at intervals of 2 nm within the wave numbers range of 1100–2498 nm. Because the working principle of each spectrometer was different, the NIR spectra dataset obtained by different spectrometers were different. In the present study, the NIR spectra of 80 corn samples measured on M5 instrument were considered as X and the oil value was considered as property of interest Y. We used Kennard-Stone (K-S) method to split into a calibration set (60 samples) and an independent test set (20 samples). NIR spectra of corn were displayed in Fig.2.

2.2 Model evaluation and spectral pretreatment

2.2.1 Model evaluation parameters

The function of model evaluation parameters is to evaluate the reliability of prediction model established by calibration set sample. The model evaluation parameters including prediction residual error sum of squares (PRESS), root mean square error of calibration set (RMSEC), root mean square

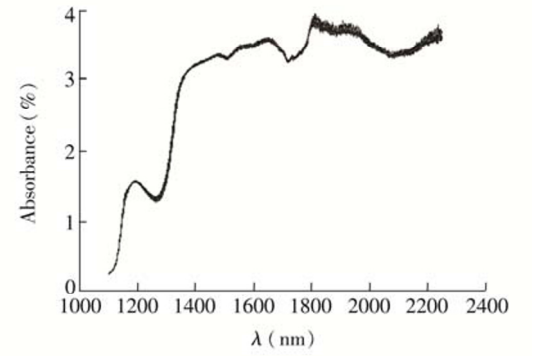

Fig.1 Raw NIR spectra of beer

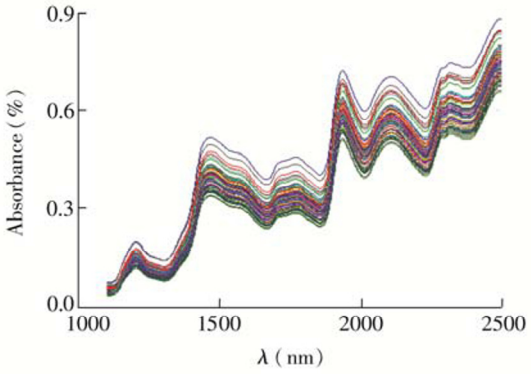

Fig.2 Raw NIR spectra of corn