Accepted Manuscript

Title: Big data driven outlier detection for soybean straw Near Infrared Spectroscopy

Authors: Yan Wang, Qingfen Liu, Han-dan Hou, Seungmin Rho, Brij Gupta, Ying-xin Mu, Wei-zheng Shen

PII: S1877-7503(17)30370-8
DOI: http://dx.doi.org/doi:10.1016/j.jocs.2017.06.008
Reference: JOCS 707

To appear in:

Received date: 1-4-2017
Revised date: 24-5-2017
Accepted date: 13-6-2017

Please cite this article as: Yan Wang, Qingfen Liu, Han-dan Hou, Seungmin Rho, Brij Gupta, Ying-xin Mu, Wei-zheng Shen, Big data driven outlier detection for soybean straw Near Infrared Spectroscopy, Journal of Computational Sciencehttp://dx.doi.org/10.1016/j.jocs.2017.06.008

Big Data Driven Outlier Detection for Soybean Straw Near Infrared Spectroscopy

Yan Wang [1] Qingfen Liu[2] Han-dan Hou [2] Seungmin Rho[3] Brij Gupta[4]
Ying-xin Mu [1] Wei-zheng Shen [1※]

1. *College of Electrical and Information, Northeast Agricultural University, Harbin 150030, China*
2. *Harbin University of Finance, Harbin 150076, China*
3. *Department of Multimedia, Sungkyul University, Gyeonggi-do 430724, Korea*
4. *Department of Computer Engineering, National Institute of Technology Kurukshetra, Haryana 136119, India*

E-mail: wangyan_neau@126.com

High Lights

In processing of near infrared spectrum(NIRS) big data, outliers commonly occur due to the inevitable spectrum or chemical mistakes, which will seriously affect the final modeling accuracy. In this paper, an big data driven outlier detection method is explored based on NIRS big data of soybean straw. Firstly, aiming at effectively identify the representative samples, a novel re-sampling method IRHM is provided; then by collaborate IRHM with Cook's distance measurement, we further propose the IRHM-COOK method to detect outlier samples for NIRS analysis data. Specifically, the confidence interval of IRHM-COOK is optimized by balancing the best confidence intervals of IRHM and Cook's distance, which makes the latter recognition process of spectrum and chemical outliers relatively independent. The experimental results show that IRHM-COOK method is superior to traditional methods and effectively improves the performance of spectrum and chemical outliers detection for Soybean Straw Near Infrared Spectroscopy big data.

Abstract

In near infrared spectroscopy (NIRS) analysis, the prediction ability of the model is seriously affected by outliers that may be the result of errors related to the spectral measurements, the chemical analysis, or a combination of both. In this paper, an outlier detection method is described based on the NIRS analysis data of soybean straw. We improved the resampling by half-mean (RHM) method by including a confidence interval (IRHM) and combined the IRHM and Cook's distance methods (IRHM-COOK) to detect outlier samples in the NIRS data. The confidence interval is an important parameter in the IRHM-COOK method and the optimal confidence intervals for the IRHM and Cook's distance methods are combined and used as the confidence interval for the IRHM-COOK method. The selection process for the confidence interval is aimed at relative independence between the detection of the spectrum outliers and the chemical outliers. The experimental results show that the IRHM-COOK method is superior to the traditional Mahalanobis distance method, the IRHM method, and the Cook's distance method using a partial least squares regression (PLS) model. The determination coefficient (R^2) of a hemicellulose PLS calibration model increased from 0.4397918 to 0.5333039 and the root mean square error (RMSE) decreased from 0.7926415 to 0.7287254. The PLS models for lignin and cellulose performed better using the IRHM-COOK method than the original model. The results show that the IRHM-COOK