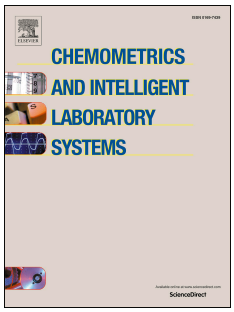
Accepted Manuscript



A novel variable selection method based on stability and variable permutation for multivariate calibration

Junming Chen, Chunhua Yang, Hongqiu Zhu, Yonggang Li, Weihua Gui

PII: S0169-7439(18)30051-0

DOI: 10.1016/j.chemolab.2018.09.009

Reference: CHEMOM 3685

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received Date: 23 January 2018

Revised Date: 2 August 2018

Accepted Date: 20 September 2018

Please cite this article as: J. Chen, C. Yang, H. Zhu, Y. Li, W. Gui, A novel variable selection method based on stability and variable permutation for multivariate calibration, *Chemometrics and Intelligent Laboratory Systems* (2018), doi: https://doi.org/10.1016/j.chemolab.2018.09.009.

A novel variable selection method based on stability and variable permutation for multivariate calibration

Junming Chen[a], Chunhua Yang[a,*], Hongqiu Zhu[a,*], Yonggang Li[a], Weihua Gui[a]

[a] *School of Information Science and Engineering, Central South University, Changsha, Hunan 410083, China*

Abstract

A novel variable selection method named stability and variable permutation (SVP) is proposed based on evolutionary principles of 'intraspecific competition' and 'survival of the fittest'. In SVP, variables are selected in an iterative and competitive manner. In each iteration, Monte Carlo sampling (MCS) runs in sample space and variable space for stability and variable permutation, respectively. Variables are divided into elite variables and normal variables according to stability by adaptive reweighted sampling (ARS). Then, combining variable permutation analysis, exponentially decreasing function (EDF) is employed to select important variables from normal variables. Elite variables and important variables construct a new variable subset for the next iteration. After the selection iterations are terminated, the root mean square error of cross validation (RMSECV) of each subset is calculated. The variable subset with the minimal RMSECV is considered as the optimal variable subset. The performance of SVP is evaluated by three near-infrared (NIR) datasets: corn oil dataset, diesel fuel total aromatics dataset and wheat protein dataset. Compared with methods of moving window PLS (MWPLS), Monte Carlo uninformative variable elimination (MCUVE), competitive adaptive reweighted sampling (CARS), stability competitive adaptive reweighted sampling (SCARS) and variable permutation population analysis (VPPA), SVP shows better prediction results.

Keywords: Variable selection; Stability; Variable permutation; Exponentially decreasing function; Adaptive reweighted sampling.

Corresponding author. Tel.: +86 13873126460; E-mail address: hqcsu@csu.edu.cn (H. Zhu). Tel.: +86 0731 88836876; E-mail address: ychh@csu.edu.cn (C. Yang).