Joint variable selection and preprocessing optimization approach to the multivariate calibration

Vladislav Galyanin[1], Andrey Bogomolov[1,2,3]

[1] Samara State Technical University, Samara, Russia

[2] art photonics GmbH, Berlin, Germany

[3] Global Modelling, Aalen, Germany

Variable selection is a one of the most common and important treatments preceding the multivariate calibration, particularly, of spectral data. Irrelevant and noisy data parts are commonly removed from the calibration set in order to simplify multivariate models and improve their performance.

Nowadays, data scientists usually determine an optimal preprocessing method and perform variable selection separately, as two independent data pretreatment steps preceding the calibration. In general, however, the preprocessing method, the optimal variables and even the model complexity (the number of latent variables, LVs) are related to each other. Sequential optimization may result in the selection of variables, which are only optimal for the chosen preprocessing method for a given number of LVs.

Variable selection is usually performed using the binary variable coding, as in conventional genetic algorithm (GA) or using a sequential way, like interval partial least-squares (iPLS) regression. Conventional GA optimization with the binary coding leads to an excessive complexity of the optimization problem, and sequential interval selection cannot guarantee finding the global optimal solution.

In the present work a generalized optimization approach for simultaneous determination of an optimal preprocessing method, variable intervals (including their widths) and the number of LVs is presented. The proposed approach is “time-dependent” but well generalized, has lower optimization complexity than common variable selection algorithms and converges to a global optimal solution in many cases.

Here, GA has been used as a parametric optimization routine in order to obtain optimal values of the merit function. Various calibration statistics are considered for quantification and discrimination cases. Problems of chromosome coding, genetic algorithm adaptation and

constraints, interval width and spectral resolution, performance issues and merit function selection are discussed.

Code for Interval selection was written in GNU Octave/MATLAB and used in TPT cloud software (www.tptcloud.com) for online optimization of variable selection “on the cloud”. All m-files are designed in a way to be optionally used as an independent standalone GNU Octave/MATLAB toolbox. The software (Interval Selection toolbox) has been registered in Russian Federal Service on Intellectual Properties.