Abstract

***J. Chemometrics,* 23**, 160-171 (2009)

Filzmoser P., Liebmann B., Varmuza K.:

Repeated double cross validation.

Repeated double cross validation (rdCV) is a strategy for

(a) optimizing the complexity of regression models and (b) for a realistic estimation of prediction errors when the model is applied to new cases (that are within the population of the data used).

This strategy is suited for small data sets and is a complementary method to bootstrap methods. rdCV is a formal, partly new combination of known procedures and methods, and has been implemented in a function for the programming environment R, providing several types of plots for model evaluation. The current version of the software is dedicated to regression models obtained by partial least-squares (PLS).

The applied methods for repeated splits of the data into test sets and calibration sets, as well as for estimation of the optimum number of PLS components, are described. The relevance of some parameters (number of segments in CV, number of repetitions) is investigated.

rdCV is applied to two data sets from chemistry:

(1) determination of glucose concentrations from near infrared (NIR) data in mash samples from bioethanol production;

(2) modeling the gas chromatographic retention indices of polycyclic aromatic compounds from molecular descriptors.

Models using all original variables and models using a small subset of the variables, selected by a genetic algorithm (GA), are compared by rdCV.