Abstract

In Khanmohammadi M. (ed.), ***Current Applications of Chemometrics***,
Nova Science Publishers, New York, USA (2015), p. 15 - 31.

Varmuza K., Filzmoser P.: [*]

Repeated double cross validation (rdCV) - a strategy for optimizing empirical multivariate models, and for comparing their prediction performances

Repeated double cross validation (rdCV) is a resampling strategy for the development of multivariate models in calibration or classification. The basic ideas of rdCV are described, and application examples are discussed.

rdCV consists of three nested loops. The outmost loop performs repetitions with different random sequences of the objects. The second loop performs CV by splitting the objects into calibration and test sets. The third (most inner) loop uses a calibration set for estimating an optimal model complexity by CV.

Optimization of model complexity is strictly separated from the estimation of the model performance. Model performance is solely derived from test set objects. The repetition loop allows an estimation of the variability of the used performance measure and thus makes comparisons of models more reliable.

rdCV is combined with PLS regression, DPLS classification, and KNN classification.

Worked out examples with data sets from analytical chemistry demonstrate the use of rdCV for
(1) determination of the ethanol concentration in fermentation samples using NIR data;
(2) comparison of variable selection methods;
(3) classification of synthetic organic pigments (relevant for artists' paints) using IR data;
(4) classification of AMES mutagenicity (QSAR).

Results are mostly presented in plots directly interpretable by the user. Software for rdCV has been developed in the R programming environment.

[*] Vienna University of Technology, Institute of Statistics and Mathematical Methods in Economics, Vienna, Austria