

ORAL PRESENTATION

Open Access

Cross-validation is dead. Long live cross-validation! Model validation based on resampling

Knut Baumann

From 5th German Conference on Cheminformatics: 23. CIC-Workshop
Goslar, Germany. 8-10 November 2009

Cross-validation was originally invented to estimate the prediction error of a mathematical modelling procedure. It can be shown that cross-validation estimates the prediction error almost unbiasedly. Nonetheless, there are numerous reports in the chemoinformatic literature that cross-validated figures of merit cannot be trusted and that a so-called external test set has to be used to estimate the prediction error of a mathematical model. In most cases where cross-validation fails to estimate the prediction error correctly, this can be traced back to the fact that it was employed as an objective function for model selection. Typically each model has some meta-parameters that need to be tuned such as the choice of the actual descriptors and the number of variables in a QSAR equation, the network topology of a neural net, or the complexity of a decision tree. In this case the meta-parameter is varied and the cross-validated prediction error is determined for each setting. Finally, the parameter setting is chosen that optimizes the cross-validated prediction error in an attempt to optimize the predictivity of the model. However, in these cases cross-validation is no longer an unbiased estimator of the prediction error and may grossly deviate from the result of an external test set. It can be shown that the "amount" of model selection can directly be related to the inflation of cross-validated figures of merit. Hence, the model selection step has to be separated from the step of estimating the prediction error. If this is done correctly, cross-validation (or resampling in general) retains its property of unbiasedly estimating the prediction error. Matter of factly, it can be shown that data splitting into a training set and an external test set often estimates

the prediction error less precise than proper cross-validation. It is this variability of prediction errors, which depends on test set size, that causes seemingly paradox phenomena such as the so-called "Kubinyi's paradoxon" for small data sets.

Published: 4 May 2010

doi:10.1186/1758-2946-2-S1-O5

Cite this article as: Baumann: Cross-validation is dead. Long live cross-validation! Model validation based on resampling. *Journal of Cheminformatics* 2010 **2**(Suppl 1):O5.

Publish with **ChemistryCentral** and every scientist can read your work free of charge

"Open access provides opportunities to our colleagues in other parts of the globe, by allowing anyone to view the content free of charge."

W. Jeffery Hurst, The Hershey Company.

- available free of charge to the entire scientific community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
<http://www.chemistrycentral.com/manuscript/>



ChemistryCentral

