Titel:

Optimal classifier selection and negative bias in error rate estimation: An empirical study on high-dimensional prediction

Abstract:

In biometrical practice, researchers often apply a large number of different methods in a "trial-and-error" strategy to get as much as possible out of their data and, due to publication pressure or pressure from the consulting customer, present only the most favorable results. This strategy may induce a substantial optimistic bias in prediction error estimation, which is quantitatively assessed in the present study.

The focus of this talk is on class prediction based on high-dimensional data (e.g. microarray data), since such analyzes are particularly exposed to this kind of bias. In this study I consider a total of 124 correct variants of classifiers (possibly including variable selection or tuning steps) within a cross-validation evaluation scheme. The classifiers are applied to original and modified real microarray data sets, some of which are obtained by randomly permuting the class labels to mimic non-informative predictors while preserving their correlation structure. I then asseess the minimal misclassification rate over the results of the 124 variants of classifiers in order to quantify the bias arising when the optimal classifier is selected a posteriori in a data-driven manner.

I conclude that the strategy to present only the optimal result is not acceptable, and suggest an alternative approach for properly reporting classification accuracy.