Improving Cross-Validation Classifier Selection Accuracy through Meta-learning

Jesse H. Krijthe J.H.KRIJTHE@LUMC.NL
Leiden University Medical Center, Einthovenweg 20, 2333 ZC Leiden, The Netherlands

Tin Kam Ho TKH@RESEARCH.BELL-LABS.COM
Bell Laboratories, Alcatel-Lucent, 600 Mountain Ave., Murray Hill, New Jersey, 07974-0636, USA

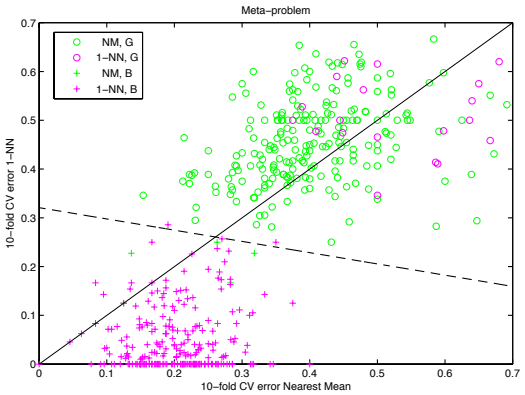
Marco Loog M.LOOG@TUDELFT.NL
Delft University of Technology, Mekelweg 4, Mekelweg 4, 2628 CD Delft, The Netherlands

Keywords: classifier selection, meta-learning, cross-validation

Given the large amount of classication algorithms available, choosing an algorithm for a given dataset is a non-trivial problem. In practice, a cross-validation procedure is often employed to estimate the true errors of a set of classifiers and the classifier with the lowest error estimate is used. However, for small sample sizes, cross-validation error estimates have been shown to become unreliable (Braga-Neto & Dougherty, 2004). Krijthe et. al. (2012) explore whether one can improve classifier selection using techniques from the field of meta-learning. This contribution recapitulates the main finding.

Meta-learning assumes a collection of datasets is given. Selecting a classifier can then be seen as a classification problem on a *meta* level where datasets are the meta-objects and the meta-features can be any measure derived from a dataset. The meta-classes are the classifiers that have the lowest true error on each dataset. One could consider as a special case of meta-features the cross-validation errors of all classifiers under consideration.

As an illustration, the figure shows the meta-problem of a simulated collection of datasets consisting of two base problems. The goal is to choose which of two classifiers would give a lower generalization error. Regular cross-validation selection corresponds to the diagonal boundary in this space. It is clear that the decision boundary of a trained meta-classifier, the dotted line, is markedly different. In fact, when using this meta-classifier the error in selecting the best classifier drops from 0.16 to 0.06. Additionally, adding other meta-features, such as the variance of the cross-validation errors, further improves the classifier selection.



These results corroborate the interesting observation that classifier selection by meta-learning techniques can outperform the de facto standard: cross-validation. Experiments on quasi-real world data suggests these effect may be present in non-simulated data as well. Secondly, the usefulness of adding additional meta-features indicates that not all information relevant in classifier selection may be present in the cross-validation estimates, suggesting improved classifier selection techniques may be possible.

References

Braga-Neto, U., & Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, *20*, 374–380.

Krijthe, J. H., Ho, T. K., & Loog, M. (2012). Improving cross-validation based classifier selection using meta-learning. *21st International Conference on Pattern Recognition* (pp. 2873–2876).