

Stratification bias in heterogeneous, small low-signal datasets

Brian J. Parker¹, Jiayu Wen²

¹Statistical and Machine Learning Group, National ICT Australia (NICTA) and Research School of Information Sciences and Engineering, Australian National University, Canberra, Australia

²Australian Research Council (ARC) Centre of Excellence for Integrative Legume Research and Bioinformatics Laboratory, Research School of Biological Sciences, Australian National University, Canberra, Australia

Email: Brian J. Parker* - brian.parker@nicta.com.au;

*Corresponding author

Abstract

We show that when estimating classifier performance on the heterogeneous (i.e. a mixture of subtypes), small low-signal datasets common in biomedical datasets such as microarray or expressed sequence tag (EST) datasets, common sample-reuse validation schemes such as repeated holdout or cross-validation can lead to very high variance and inaccurate, pessimistic estimates. This effect is due to inadequate stratification when sampling from the (unidentified) subclasses of heterogeneous datasets. We demonstrate a new form of sample-reuse scheme that correctly stratifies for these subclasses, leading to dramatically lower variances in error rate and AUC estimates on such datasets.
