Improving model performance estimation in high-dimensional data settings by using learning curves.

Goedhart, J.M.[1], Klausch T.L.T.[1], van de Wiel, M.A.[1]

[1]Department of Epidemiology and Biostatistics, Amsterdam University Medical Centers, Amsterdam, the Netherlands

E-mail for correspondence: j.m.goedhart@amsterdamumc.nl

Abstract: In high-dimensional prediction settings, i.e. when p > n, it remains challenging to estimate the test performance (e.g. AUC). Conventional K-fold cross-validation and subsampling methods aim to balance between enough samples to reliably learn the model and estimate its performance. We show that combining estimates from a trajectory of subsample sizes, rendering a learning curve [1], leads to several benefits. Firstly, use of a smoothed curve can improve the performance estimate. Secondly, a still growing- or saturating learning curve indicates whether or not additional samples will boost the prediction accuracy. Thirdly, comparing the trajectories of different learners results in a more complete picture than doing so at one sample size only. Fourthly, the learning curve allows computation of a lower confidence bound for the performance. Standard cross-validation suffers from a limited amount of test samples, whereas the learning curve finds a better trade-off between training- and test sample sizes. This confidence bound is proven to be valid. We show coverage results from a simulation, and compare those to a state-of-the-art technique based on asymptotics [2]. Finally, we demonstrate the benefits of our approach by applying it to several classifiers of tumor location from blood platelet RNAseq data.

Key words: High-Dimensional Data; Classification; Learning Curve; Confidence Interval; Omics;

References:

[1] **Mukherjee et al** (2003). Estimating dataset size requirements for classifying DNA microarray data. Journal of computational biology, 10, 119-142

[2] **LeDell, E. et al.** (2015). Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. Electronic journal of statistics, 91, 1583-1607