Predicting Classification Accuracy When Adding New Unobserved Classes

Yuli Slavutsky, Yuval Benjamini
Department of Statistics and Data Science, The Hebrew University of Jerusalem

Abstract: Multiclass classifiers are often designed and evaluated only on a sample from the classes on which they will eventually be applied. Hence, their final accuracy remains unknown. We study how a classifier's performance over the initial class sample can be used to extrapolate its expected accuracy on another, unobserved set of classes. For this, we define a measure of separation between correct and incorrect classes that is independent of the number of classes: the reversed ROC (rROC). We show that classification accuracy is a function of the rROC in multiclass classifiers, for which the learned representation of data from the initial classes remains unchanged when new classes are added. We formulate a robust neural-network-based algorithm, CleaneX, which learns to estimate the accuracy of such classifiers on arbitrarily large sets of classes, and show that our method achieves remarkably better predictions than current state-of-the-art methods on both simulations and real datasets.

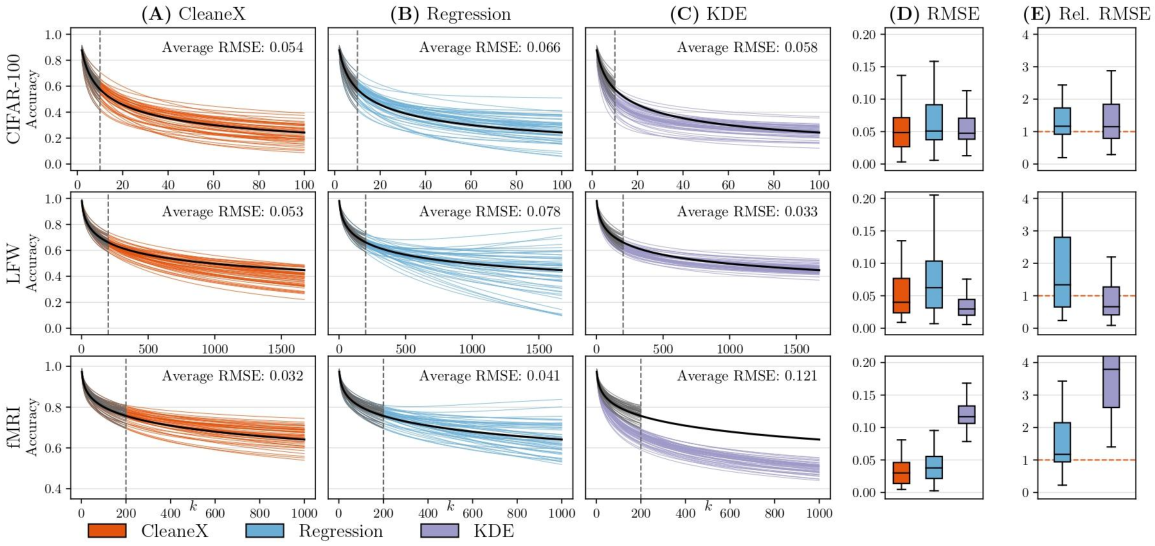

Figure 1: Experimental results. A, B, C: accuracy curves of three tasks. The colored lines show predicted accuracies by CleaneX (Slavutsky & Benjamini, 2021), regression (Zheng et al., 2018) and KDE (Kay et al., 2008). Dotted vertical lines denote the number of classes of which accuracy was observed, grey curves correspond to the observed accuracy at each repetition, black curves correspond to the true accuracy curve. D: distribution of root mean squared errors (RMSE) over 50 repetitions. E: distribution of the ratio between RMSE values of regression/KDE and CleaneX. CleaneX outperforms the regression and KDE methods in all cases except for KDE on the LFW dataset.

References

Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. Nature, 452(7185):352–355, 2008.

Yuli Slavutsky and Yuval Benjamini. Predicting classification accuracy when adding new unobserved classes. In International Conference on Learning Representations (ICLR), 2021.

Charles Zheng, Rakesh Achanta, and Yuval Benjamini. Extrapolating expected accuracies for large multi-class problems. Journal of Machine Learning Research, 19(65):1–30, 2018.