Special Issue on
Integrating Multiple Learned Models for Improving and Scaling Machine Learning Algorithms

Most modern Machine Learning, Statistics and KDD techniques use a single model or learning algorithm at a time, or at most select one model from a set of candidate models. Recently however, there has been considerable interest in techniques that integrate the collective predictions of a set of models in some principled fashion. With such techniques often the predictive accuracy and/or the training efficiency of the overall system can be improved, since one can "mix and match" among the relative strengths of the models being combined.

By 1995 it was evident that a growing body of literature and researchers were reporting a variety of ingenious techniques and algorithms that demonstrated significant improvements over traditional machine learning techniques. Much of this work was spread out over several research communities including classical statistics, neural networks, machine learning and KDD, among others. For these reasons, the AAAI Workshop on Integrating Multiple Learned Models was organized and held in August 1996 to gather researchers actively working in this area. The twenty four papers accepted to that workshop presented techniques that generate and integrate multiple learned models: using different training data distributions or training over different partitions of the data, using different output classification schemes, or using different hyperparameters or training heuristics. A number of system architectures that implement such strategies were also reported.

In many of the cases reported, a classification system composed by the integration of a number of separately learned classifiers or models tends to improve overall accuracy achievable by any individual model. Furthermore, several of the presented methods are amenable to direct parallel or distributed computation for improved efficiency and scalability of machine learning. The latter is perhaps most important for contexts where large amounts of distributed data are available over a network of remote sites, for example, web database sites.

Buoyed by the successful IMLM Workshop, the organizing committee proposed a special issue on this topic to the Journal on Machine Learning. The outcome of this effort is the collection of papers you now see before you.

Merz and Pazzani's paper introduces the PCR* algorithm. The algorithm integrates multiple base regression models by performing a Principal Components analysis on the outputs of the models to detect highly correlated models, as well as the unique contributions of each model for specific target outcomes. The resultant eigenvectors define weights on the underlying models, based upon accuracy estimates of the models on the training data. A number of test problems are used in the evaluation with positive results reported using standard Cross Validation.

Whereas the previous paper combines all available models based upon principal component analysis, Merz' other paper considers a method for integrating only a subset of available models. The SCANN algorithm is introduced that first performs a correspondence analysis over a set of base models, and then iteratively searches for a subset of these models that are ultimately combined using stacking. It is well known that models with uncorrelated errors combine to produce a better single model than models with highly correlated errors. SCANN capitalizes on this by finding a subset of uncorrelated models. The paper presents a number of empirical tests of SCANN with significant improvements over other combining methods.

Just as with supervised learners, one would expect that adaptively combining unsupervised learners (density estimators) should often result in performance superior to that of any single one of the fixed estimators being combined. Also just as with supervised learning, one can explore a baseline case of schemes that combine unsupervised learners by forming a linear combination of their predictions, and one can form those combination coefficients by examining the correlations between out-of-sample predictions and actual out-of-sample data, i.e., by stacking (Wolpert, 1992; Breiman, 1996d; Leblanc & Tibshirani, 1993; Kim & Bartlett, 1995; Merz, 1998). In *Linearly Combining Density Estimators via Stacking* Smyth and Wolpert explore such a scheme for combining kernel density and mixture model density estimators. They find on both artificial and real world data that stacking together density estimators consistently outperforms any single one of the estimators being combined. This is true even when the (artificial) data was directly generated from a single one of the estimators being combined.

Breiman considers scalable classification from large databases (the canonical example is where the full set of data exists on disk and is far too large to fit into core memory all at once). His approach is to use adaptive resampling of the data to form successive data sets that are fed into one's (centralized) learning algorithm, and then combine the resultant estimators. This is different from integrating models trained from disjoint subsets (Chan & Stolfo, 1997). On the other hand, this is similar to boosting (Schapire, 1990). Only rather than the standard schemes used in boosting, Breiman combines the estimators using out-of-sample techniques, as in his work on arcing (Breiman, 1996b), as well as previous work on stacking, and on estimating the error of bagging (Wolpert & Macready, 1997; Tibshirani, 1996; Breiman, 1996c)

Bauer and Kohavi's article provides a large-scale empirical comparison of a number of voting-based algorithms for combining classifiers. Using fourteen data sets, they investigated variants of bagging (Breiman, 1996a) and boosting (Schapire, 1990) with decision tree (three variants) and Naive-Bayes algorithms as the base inducers. They analyzed error rates through a decomposition of bias and variance and observed their influence on misclassification. Their results provide some insights on earlier observations (Breiman, 1996a) that combining is beneficial to "unstable" learning algorithms. Additional voting variants allow tree pruning to be disabled, base inducers to report probablistic estimates, mean-square error for evaluation, estimation of leaf probabilities based on the entire training set in bagging, and others. One interesting finding is the positive correlation between increase in tree size and reduction in error rate in AdaBoost (Freund & Schapire, 1996).

We wish to thank Tom Dietterich for his support and enthusiasm for this special issue and Doug Fisher for serving as an independent editor for one of the articles in this issue. Also, this special issue would not be possible without the help from the thirty three colleagues who reviewed the twenty four submissions.

Philip K. Chan
Salvatore J. Stolfo
David Wolpert

References

Breiman, L. (1996a). Bagging predictors. *Machine Learning*, *24*, 123–140.

Breiman, L. (1996b). Bias, variance and arcing classifiers. Tech. rep. 460, University of California, Berkeley, CA.

Breiman, L. (1996c). Out-of-bag estimation. ftp.stat.berkeley.edu/pub/users/breiman/OOBestimation.ps.

Breiman, L. (1996d). Stacked regressions. *Machine Learning*, *24*, 41–48.

Chan, P., & Stolfo, S. (1997). On the accuracy of meta-learning for scalable data mining. *J. Intelligent Information Systems*, *9*, 5–28.

Freund, Y., & Schapire, R. (1996). Experiments with a new boosting algorithm. In *Proc. Thirteenth Conf. Machine Learning*, pp. 148–156.

Kim, K., & Bartlett, E. B. (1995). Error estimation by series association for neural networks. *Neural Computation*, *7*, 799–821.

Leblanc, M., & Tibshirani, R. (1993). Combining estimates in regression and classification. University of Toronto, Dept. of Statistics.

Merz, C. (1998). *Classification and Regression by Combining Models*. Ph.D. thesis, Univ. of California, Irvine, CA.

Schapire, R. (1990). The strength of weak learnability. *Machine Learning*, *5*, 197–226.

Tibshirani, R. (1996). Bias, variance and prediction error for classification rules. University of Toronto Statistics Department Technical Report.

Wolpert, D. (1992). Stacked generalization. *Neural Networks*, *5*, 241–259.

Wolpert, D. H., & Macready, W. G. (1997). An efficient method to estimate bagging's generalization error. Accepted at Machine Learning Journal.