The University of Chicago

Department of Statistics

Seminar

Abraham Wyner
Department of Statistics
Wharton School of the University of Pennsylvania

"Bagging and Boosting Noisy Data"

****

Monday, April 29, 2002 at 4:00 pm
133 Eckhart Hall, 5734 S. University Avenue

ABSTRACT

While the overall success of AdaBoost (Freund and Schapire 1996) in particular and boosting in general, is indisputable, there is increasing evidence that boosting algorithms are not quite as immune from overfitting as indicated in early reports. Dietterich (2000) and Opitz and Maclin (1999) among others report that boosting is quite susceptible to data corrupted by the introduction of independent label noise. Friedman et. al.(2000) provide another example of a data set for which boosting overfits: concentric spheres with significant overlap. In short, research to date point to two classes of problems where boosting may overfit: 1) independent label noise 2) overlapping regions. To be fair, classification in such situations is hardest to accomplish.

In this talk we will discuss the application of bagging and boosting algorithms to noisy datasets. We will discuss a new variation on the obvious idea of combining both bagging and boosting to produce a new algorithm (which we will call the "BB" algorithm). This algorithm is particularly designed to outperform boosting in noisy problems (mainly by smoothing in order to mitigate the overfit), while performing as well or nearly as well in non-noisy settings. In short the BB algorithm offers improvements over bagging and boosting when applied to tough, hard-to-classify noisy datasets, without sacrifices any loss in performance on those datasets where boosting algorithms have demonstrated their greatest successes. We present both empirical results (on real and simulated datasets) and theoretical results.

Finally, we take a hard look at resistance to overfitting again in the context of noisy models. We offer a new explanation for how boosting algorithms resist overfitting. What we observe is that by continuing to iterate the boosting procedure even after the training set error rate has reached zero, you allow boosting to self-smooth in a manner similar to what is obtained by applying bagging and boosting together.

4/26/02