Overfitting is a Many Splendored Thing

Rich Caruana
Microsoft Corporation
Redmond WA 98052
rcaruana@microsoft.com

We tend to think of overfitting as a global phenomenon that can be controlled by a few regularization parameters. Nothing is further from the truth. In complex data not all train cases train, and then begin to overtrain, at the same rate. The more powerful the learning algorithm, the more training cases diverge from the norm, even when regularization parameters are carefully tuned: some kinds of training cases fit and then begin to overfit early in training, while others may never adequately be fit when learning is terminated.

Figure 1 shows an example of differential overfitting in boosted decision trees. The train set contains more than 1 million labeled documents, and about 1000 features per document. The task is to predict document quality. Overfitting is measured on a large test set containing more than 200,000 labeled documents. Figure 1 (left) shows the aggregate accuracy across the full test set vs. the number of trees in a gradient boosted tree model. The graph is the average of ten runs. Regularization parameters such as tree size, minimum number of training cases per leaf, and learning rate have been adjusted so that generalization accuracy peaks at about 1000 trees. The x axis is a log scale running from 100 trees to 10,000 trees. Figure 1 (right) shows accuracy vs. number of trees for the same models trained on the same data, but in this graph the test set has been split into three subgroups. In group A generalization peaks at 200 trees and the model overfits dramatically on these cases as more trees are added to the model. In group B generalization peaks at about 1000 trees. In group C generalization does not peak until 2000 trees have been added to the model. If training is stopped after 1000 trees as the aggregate generalization curve suggests is best, cases in group C will remain underfit.

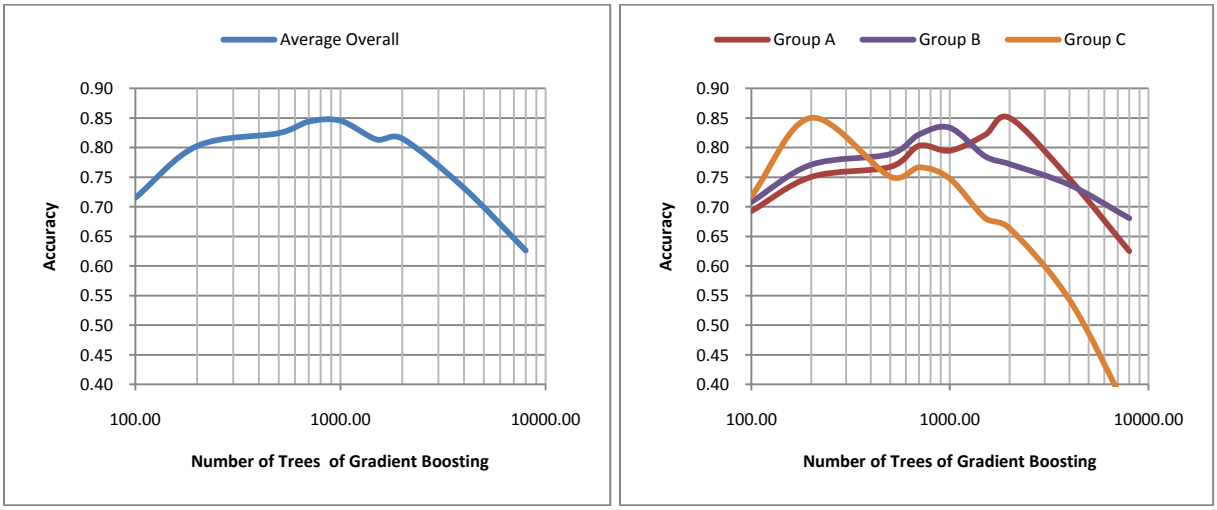

Figure 1 Generalization Overall (left) and by Group (Right) vs. # of Trees in Gradient Boosting

There are a variety of mechanisms that might be employed to prevent differential overfitting and thus increase overall accuracy. Simple methods such as splitting training data into different (possibly overlapping) train sets and tuning regularization parameters separately for each set typically does not work because too much accuracy is lost by reducing train set size. It is important to prevent overfitting without reducing effective sample size.

The presentation will focus on three issues:

1) What are the main causes of differential overfitting: local sample size? local manifold dimensionality? local mismatch between model bias and data?
2) How to find groups of training cases that suffer from differential overfitting.
3) How to increase accuracy by preventing differential overfitting without hurting learning.

Topic: supervised learning
Preference: oral