Why Does Feature Selection Work?

Rich Caruana
Microsoft Corporation
One Microsoft Way, Redmond WA 98052

Art Munson
Department of Computer Science
Cornell University, Ithaca, NY 14853

rcaruana@microsoft.com
mmunson@cs.cornell.edu

We examine the mechanism by which feature selection improves the accuracy of supervised learning. An empirical bias/variance analysis as feature selection progresses indicates that most of the benefit from feature selection results from a decrease in variance, not from other mechanisms such as a reduction in number of correlated features, elimination of harmful features, or separating relevant from irrelevant features. In particular, the results show that the optimal feature set is not the set that contains mainly relevant features while discarding irrelevant features, but is the set that best trades-off the benefits of variance reduction with the harm caused by the increased bias that occurs as relevant features are eliminated.

The discovery that the main benefit of feature selection is variance reduction led us to ask if similar benefits can be achieved without feature selection by using other variance reduction methods such as bagging. The answer is "yes". Bagging models trained on *all* features consistently yields better performance than bagging models on reduced feature sets: feature selection almost always hurts accuracy when models are bagged.

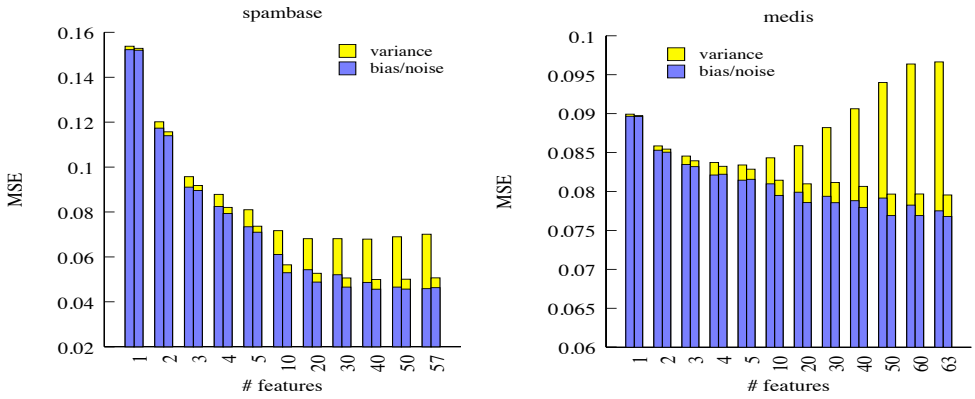

Figure 1: Bias-variance decomposition for squared loss on two data sets. In each plot, bars on the left of each pair are for learning without bagging, and bars on the right of each pair are learning with bagging. On spambase, the benefits from feature selection are modest, and bagging performs best with little or no feature selection. On medis, feature selection is critical without bagging, but bagging clearly performs best without any feature selection.

The results suggest that 1) because feature selection only finds an optimal tradeoff between the bias of having too few features and the variance of having too many features, feature selection should not be considered a method for determining which features are relevant to a problem; 2) where feature selection is not feasible, bagging is a useful alternative that yields similar or better results; and 3) when models will be bagged, feature selection usually is detrimental and it is better to train the base models using all available features. The results of our analysis hold even for learning methods such as decision trees that incorporate feature selection into the learning algorithm.