Decimated Input Ensembles for Improved Generalization

Kagan Tumer, NASA Ames Research Center
Nikunj C. Oza, The University of California, Berkeley

Purpose: Recently, many researchers have demonstrated that using classifier ensembles (e.g., averaging the outputs of multiple classifiers before reaching a classification decision) leads to improved performance for many difficult generalization problems. However, in many domains there are serious impediments to such "turnkey" classification accuracy improvements. Most notable among these is the deleterious effect of highly correlated classifiers on the ensemble performance. One particular solution to this problem is generating "new" training sets by sampling the original one. However, with finite number of patterns, this causes a reduction in the training patterns each classifier sees, often resulting in considerably worsened generalization performance (particularly for high dimensional data domains) for each individual classifier. Generally, this drop in the accuracy of the individual classifier performance more than offsets any potential gains due to combining, unless diversity among classifiers is actively promoted. In this work, we introduce a method that (i) reduces the correlation among the classifiers; (ii) reduces the dimensionality of the datà, thus lessening the impact of the "curse of dimensionality"; and (iii) improves the classification performance of the ensemble.

Method: *Input Decimation* is primarily a dimensionality reduction method. For each output class, we form a new training set where the input features are a subset of the full feature set. By varying the features that are selected for the different classifiers, this method ensures that the correlation among the trained classifiers remains low. Intuitively, input decimation decouples the classifiers by exposing them to different aspects of the same data. In order to avoid a significant drop in the overall performance, the features for each new training set are chosen to maximize the correlation between the input and the output for one particular class. More precisely, for each output class, we select those input features that show the highest correlation to that output. For an N class problem, this generates N training sets, each consisting of slightly different features.

Results: We use three data sets for this study: The Gene1 database from the PROBEN1 benchmarks, the splice junction gene sequences and satellite image database (Statlog version) from the UCI Machine Learning Repository. As a standard against which to compare the results of our input decimation tests, we trained a one hidden-layer Multi-Layered Perceptron (MLP), which we will refer to as the "original MLP", and separately trained L copies of the same MLP which we incorporated into an averaging ensemble, which we will call the "original ensemble". We only present a summary of the Gene1 results here: input decimation on the original features reduces the dimensionality of the Gene1 data from 120 to 20. This operation of course removes "information" from the inputs. Accordingly, the error rate of a single classifier trained on this decimated data was 17.8%, approximately 10% higher than the 16.3% error rate of the original MLP. However, while multiple runs of the original MLP have an average correlation of .79, the input decimated classifiers only have an average correlation of .72. The net effect of the input decimation followed by an ensemble average yields an error rate of 10.4% instead of the 13.6% error for the original ensemble. So although input decimation may increase the individual classifier error rates, it provides better "raw material" for the ensembles (less correlated classifiers). The net result is lower ensemble error rates.

New or Breakthrough Aspects of Work: In conventional dimensionality reduction methods such as Principal Component Analysis (PCA), the focus is on extracting the axes on which the data shows the highest variability. Although this approach "spreads" out the data in the new basis, and therefore is of great help in regression problems, there are no guarantees that the new axes are consistent with the discriminatory features in a classification problem. The input decimation method on the other hand explicitly seeks out discriminating features and eliminates input features that are least correlated with the outputs. The strength of the method is that only those features that have the most "explanatory" information pertinent to the discrimination task at hand are retained.

Conclusions: In this paper we investigate a novel method that reduces the correlation among individual classifiers in an ensemble. Furthermore, as a side benefit, this method reduces the dimensionality of the input space for classification problems. When used in conjunction with an ensemble of classifiers, input decimation provides better generalization performance than either single classifiers or ensembles of classifiers trained on the full data. In the public domain data sets we used for this experiment, we reduced the dimensionality of the initial feature sets by a factor of 6 to 10, while simultaneously cutting the misclassification error by 30 – 38%.