Ensemble Methods in Machine Learning

Thomas G. Dietterich

Oregon State University, Corvallis, Oregon, USA,
tgd@cs.orst.edu,
WWW home page: http://www.cs.orst.edu/~tgd

Abstract. Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions. The original ensemble method is Bayesian averaging, but more recent algorithms include error-correcting output coding, Bagging, and boosting. This paper reviews these methods and explains why ensembles can often perform better than any single classifier. Some previous studies comparing ensemble methods are reviewed, and some new experiments are presented to uncover the reasons that Adaboost does not overfit rapidly.

1 Introduction

Consider the standard supervised learning problem. A learning program is given training examples of the form $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ for some unknown function $y = f(\mathbf{x})$. The $\mathbf{x}_i$ values are typically vectors of the form $\langle x_{i,1}, x_{i,2}, \ldots, x_{i,n}\rangle$ whose components are discrete- or real-valued such as height, weight, color, age, and so on. These are also called the *features* of $\mathbf{x}_i$. Let us use the notation x_{ij} to refer to the j-th feature of $\mathbf{x}_i$. In some situations, we will drop the i subscript when it is implied by the context.

The y values are typically drawn from a discrete set of classes $\{1, \ldots, K\}$ in the case of *classification* or from the real line in the case of *regression*. In this chapter, we will consider only classification. The training examples may be corrupted by some random noise.

Given a set S of training examples, a learning algorithm outputs a *classifier*. The classifier is an hypothesis about the true function f. Given new $\mathbf{x}$ values, it predicts the corresponding y values. I will denote classifiers by $h_1, \ldots, h_L$.

An ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way (typically by weighted or unweighted voting) to classify new examples. One of the most active areas of research in supervised learning has been to study methods for constructing good ensembles of classifiers. The main discovery is that ensembles are often much more accurate than the individual classifiers that make them up.

A necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of its individual members is if the classifiers are accurate and diverse (Hansen & Salamon, 1990). An accurate classifier is one that has an error rate of better than random guessing on new $\mathbf{x}$ values. Two classifiers are

J. Kittler and F. Roli (Eds.): MCS 2000, LNCS 1857, pp. 1–15, 2000.

diverse if they make different errors on new data points. To see why accuracy and diversity are good, imagine that we have an ensemble of three classifiers: $\{h_1, h_2, h_3\}$ and consider a new case $\mathbf{x}$. If the three classifiers are identical (i.e., not diverse), then when $h_1(\mathbf{x})$ is wrong, $h_2(\mathbf{x})$ and $h_3(\mathbf{x})$ will also be wrong. However, if the errors made by the classifiers are uncorrelated, then when $h_1(\mathbf{x})$ is wrong, $h_2(\mathbf{x})$ and $h_3(\mathbf{x})$ may be correct, so that a majority vote will correctly classify $\mathbf{x}$. More precisely, if the error rates of L hypotheses h_ℓ are all equal to $p < 1/2$ and if the errors are independent, then the probability that the majority vote will be wrong will be the area under the binomial distribution where more than $L/2$ hypotheses are wrong. Figure 1 shows this for a simulated ensemble of 21 hypotheses, each having an error rate of 0.3. The area under the curve for 11 or more hypotheses being simultaneously wrong is 0.026, which is much less than the error rate of the individual hypotheses.

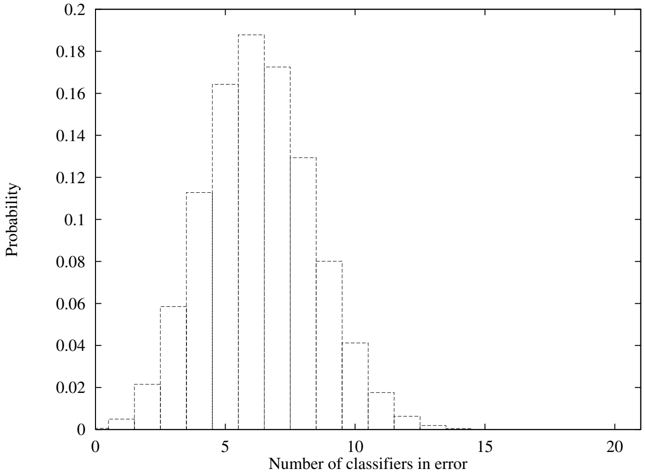

Fig. 1. The probability that exactly ℓ (of 21) hypotheses will make an error, assuming each hypothesis has an error rate of 0.3 and makes its errors independently of the other hypotheses.

Of course, if the individual hypotheses make uncorrelated errors at rates exceeding 0.5, then the error rate of the voted ensemble will *increase* as a result of the voting. Hence, one key to successful ensemble methods is to construct individual classifiers with error rates below 0.5 whose errors are at least somewhat uncorrelated.

This formal characterization of the problem is intriguing, but it does not address the question of whether it is possible in practice to construct good ensembles. Fortunately, it is often possible to construct very good ensembles. There are three fundamental reasons for this.