On Explaining Degree of Error Reduction due to Combining Multiple Decision Trees

Kamal Ali*
IBM Almaden Research Center
San Jose, CA, 95120,
ali@almaden.ibm.com
(408)927-1354

Abstract

In multiple models research, classifications or evidence is combined from several classification models in order to form overall classifications with higher accuracy than any of the individual models. However, most previous work has concentrated on demonstrating that use of multiple models leads to statistically significant error reduction as compared to the error rate of the single model learned on the same data. In this paper we present a model of the *degree* of error reduction due to the use of multiple decision trees. In particular, we show that the degree of error reduction can be linearly modeled by the "degree to which decision trees make uncorrelated errors." These conclusions are based on results from 21 domains.

Keywords: multiple decision trees, ensembles, error correlation

*This work was done while the author was at the University of California, Irvine, under the supervision of Michael Pazzani. We acknowledge the support of NSF grant IRI 93-10413.

Introduction

Multiple models research involves learning a set of classifiers, an *ensemble*, and then combining their classifications or evidence in order to make more accurate classifications. To our knowledge, this work has been done for decision trees first by Kwok & Carter (1990) and Buntine (1990), for rules first by Smyth *et al.* (1990), for neural networks first by Hansen & Salamon (1990), for regression first by Perrone (1993) and for Bayesian networks first by Madigan & York (1993). The research typically compares the error rate of the ensemble on test examples to the error rate of a single model learned on the same training data and usually demonstrates statistically significant error reduction. However, experiments by Ali & Pazzani (in press) and those of Breiman (in press) show that the degree of error reduction varies significantly from domain to domain. In this paper we show that the degree to which models make correlated errors is one major factor in explaining the degree of error reduction. Two models are said to make a correlated error if they make the same kind of classification error on the same test example.

Error reduction and error correlation

In order to model the degree of error reduction of the ensemble compared to the single classifier it is necessary to define error reduction precisely. We use error ratio - the ratio of the error rate of the ensemble to the error rate of the single classifier - as our measure. This is measured by counting the number of errors made by the ensemble over all trials and dividing by the corresponding number of errors made by the single model algorithm. We will denote error ratio by $E_r = E_e/E_s$ where E_e is the ensemble error rate and E_s is the error rate of the single classifier. The lower the error ratio, the greater the error reduction.

Now we offer a precise definition of the degree of uncorrelatedness, $1-\phi_e$, of errors made by members of an ensemble. This will be defined in terms of the degree of correlatedness, ϕ_e of the ensemble $\mathcal{F}$. Let the ensemble $\mathcal{F}$ consist of the models $\{\hat{f}_1 ... \hat{f}_T\}$ and let the true, target function be denoted by f. Therefore, $f(x) = y$ means that example x belongs to class y. In order to define "the degree of error correlatedness," let $p(\hat{f}_i(x) = \hat{f}_j(x), \hat{f}_i(x) \neq f(x))$ denote the probability that models $\hat{f}_i$ and $\hat{f}_j$ make the same kind of error. $\phi_e(\mathcal{F})$ is then just the average of the probability of making the same kind of error taken over all pairs in the ensemble. That is,

$$\phi_e(\mathcal{F}) = \frac{1}{T(T-1)} \sum_{i=1}^{T} \sum_{j \neq i}^{T} p(\hat{f}_i(x) = \hat{f}_j(x), \hat{f}_i(x) \neq f(x)) \tag{1}$$

The higher the values of ϕ_e, the more correlated the errors made by members of the ensemble. In our experiments the values of ϕ_e for the data sets were estimated on the *test* set of examples. Therefore, they cannot be used by the learning algorithm but provide us with a post-hoc understanding of why error is reduced more in some data sets.

The standard measure of correlatedness of two variables, r (the linear correlation coefficient) cannot be used for our purposes because we want the models to be correlated when they make correct predictions but to be uncorrelated when they make errors.

Learning decision tree ensembles

We use the method of top-down induction of decision trees (ID3: Quinlan, (1986)) with 1-step lookahead with respect to entropy minimization to learn a single tree.

Stochastic search (e.g. Kononenko & Kovacic, (1992)) is used to generate multiple trees from one training set. Stochastic search is used in learning decision trees as follows. Instead of picking the decision tree split that minimizes entropy we consider all decision tree splits whose resultant entropy (Quinlan 1986) is within a factor of 1.25 of the entropy of the best split.[1] Then, a split is randomly chosen from this set according to the distribution in which the probability of picking a split is equal to 1/Entropy. That is, splits that offer lower entropy have a higher probability of being picked. To prevent zero values for Entropy, we used the Laplace approximation (Kruskal & Tanur 1978) for the probabilities involved in the Entropy expression.

[1] Other values could be used but the exploration of this parameter was not a focus of our research.

Combining evidence from decision trees

We explored four methods for combining classifications or evidence from individual decision trees in order to make an overall classification. We consider four evidence combination functions to demonstrate that our results on the relation between error reduction and the tendency to make correlated errors is not sensitive to the type of combination function.

- Uniform Voting - The classification predicted by each tree is noted and the class that is predicted most frequently is used as the prediction of the ensemble. This is the only combining function that simply combines classifications. For the combination functions below, each tree must also provide the combining function with a vector of "degrees of belief" - the i'th component in the vector denotes the degree to which the tree "believes" that the test example should be classified to class i.

- Distribution Summation (Clark & Boswell 1991) - This method assocates a C-component vector (the distribution) with each leaf. C denotes the number of classes. The vector records the numbers of training examples that reached that leaf. In order to produce a classification for the ensemble for a test example, that example is filtered to the leaf of each decision tree. Then, a component-wise summation of the vectors associated with those leaves is done. The prediction of the ensemble corresponds to the class with the greatest value in the summed vector.

- Likelihood Combination (Duda, Gaschnig, & Hart 1979) - This method associates a "degree of logical sufficiency" (LS) for each class i with each leaf j. In the context of classification, the

LS of a leaf j for $Class_i$ is defined by

$$\frac{p(x \in ext(j)|x \in Class_i)}{p(x \in ext(j)|x \notin Class_i)}$$

where $ext(j)$ denotes the set of training examples that filter to leaf j and where x is a random example. These LS's are combined using the odds form (the odds of a proposition with probability p are $p/(1-p)$) of Bayes rule:

$$O(Class_i|M) \propto O(Class_i) \times \prod_j O(Class_i|M_j)$$

where M is the set of learned decision trees and M_j is the j-th tree. $O(Class_i)$ denotes the prior odds of the i-th class. For model j, $O(Class_i|M_j)$ is set to the LS of class i stored at leaf j. Finally, the test example is assigned to the class with the highest posterior odds, $O(Class_i|M)$. Likelihood Combination only works with two classes but this is consistent with our framework because with respect to any given class, all the other classes are treated as a single "negative" class.

- Bayesian Combination (Buntine 1990) - According to Bayesian probability theory, we should assign test example x to the class c with the maximum expectation for $p(c|x,\vec{x})$ taken over $\mathcal{T}$, the hypothesis space of all possible decision trees over the chosen set of attributes:

$$E_{\mathcal{T}}(p(c|x,\vec{x})) = \sum_{T \in \mathcal{T}} p(c|x,T) \times p(T|\vec{x})$$

($\vec{x}$ denotes the set of training examples.) The posterior probability of a tree T, $p(T|\vec{x})$, is calculated as in (Buntine 1990). For the "degree of endorsement," $p(c|x,T)$, made by tree T for class c for example x, we use a Laplace estimate from the training data (see Ali & Pazzani (1995) for details).

Experimental results

Figure 1 plots error ratio as a function of ϕ_e for the 72% of the data sets for which there was significant error reduction. One point represents one data set - a combination of a data set from the UCI repository[2] and a training set size. We chose domains from the UCI repository from each of the major groups of domains (medical diagnosis, molecular biology, ...). We took each domain/training-set-size combination and made 9 additional versions of it, using fewer training examples. So, for instance, if there were N examples in the original training set, we also used 0.1N, ... 0.9N. At each of these points, we conducted 30 trials. Ensembles of 11 decision trees[3] were used.

The top-left figure in Figure 1, for instance, shows that the tendency to make correlated errors explains 49% of the variance in the error ratio variable. We used the standard method of evaluating goodness of fit of a linear model: r: the linear correlation coefficient (Kruskal & Tanur 1978). The linear correlation coefficient (r_{E_r,ϕ_e}) between error correlation (ϕ_e) and error ratio (E_r) can be used to measure how well ϕ_e linearly models error ratio. We obtain $r^2_{\phi_e,E_r} = 0.49$ indicating that 49% of the variance in error ratio can be explained by correlatedness of errors when Uniform Voting is used. The figures show two things: i) there is a relationship between correlatedness of errors (ϕ_e) and error ratio (E_r) that can be modeled well by a linear form, and ii) correlatedness of errors explains a *significant* amount of the variance in error ratio. Because there is a positive relationship between error ratio and correlatedness of errors it follows that there is a negative relationship between error ratio and *un*correlateness of errors. Therefore, the figures establish the main result of this paper: that the degree of error reduction (as encapsulated by the error ratio variable) can be modeled by the degree to which models make errors in an uncorrelated manner, $1 - \phi_e$.

Of the 260 data sets (from 21 domains), statistically significant error reduction occurred on 163 occasions when using Uniform Voting.[4] Figure 1 also shows that for the other three combination functions the degree to which error is reduced is negatively correlated with the degree to which constituents in the ensemble make individual errors in an correlated manner.

Because r is distributed normally for samples of large (greater than 30) size we can apply a significance test to see what the probability of achieving

[2]URL: http://www.ics.uci.edu/$\tilde{m}$learn/MLRepository.html

[3](Ali 1996) contains further exploration of the effect of ensemble size - it is not our focus here.

[4]On most of the remaining cases, the multiple models method did equally as well as the single model method.

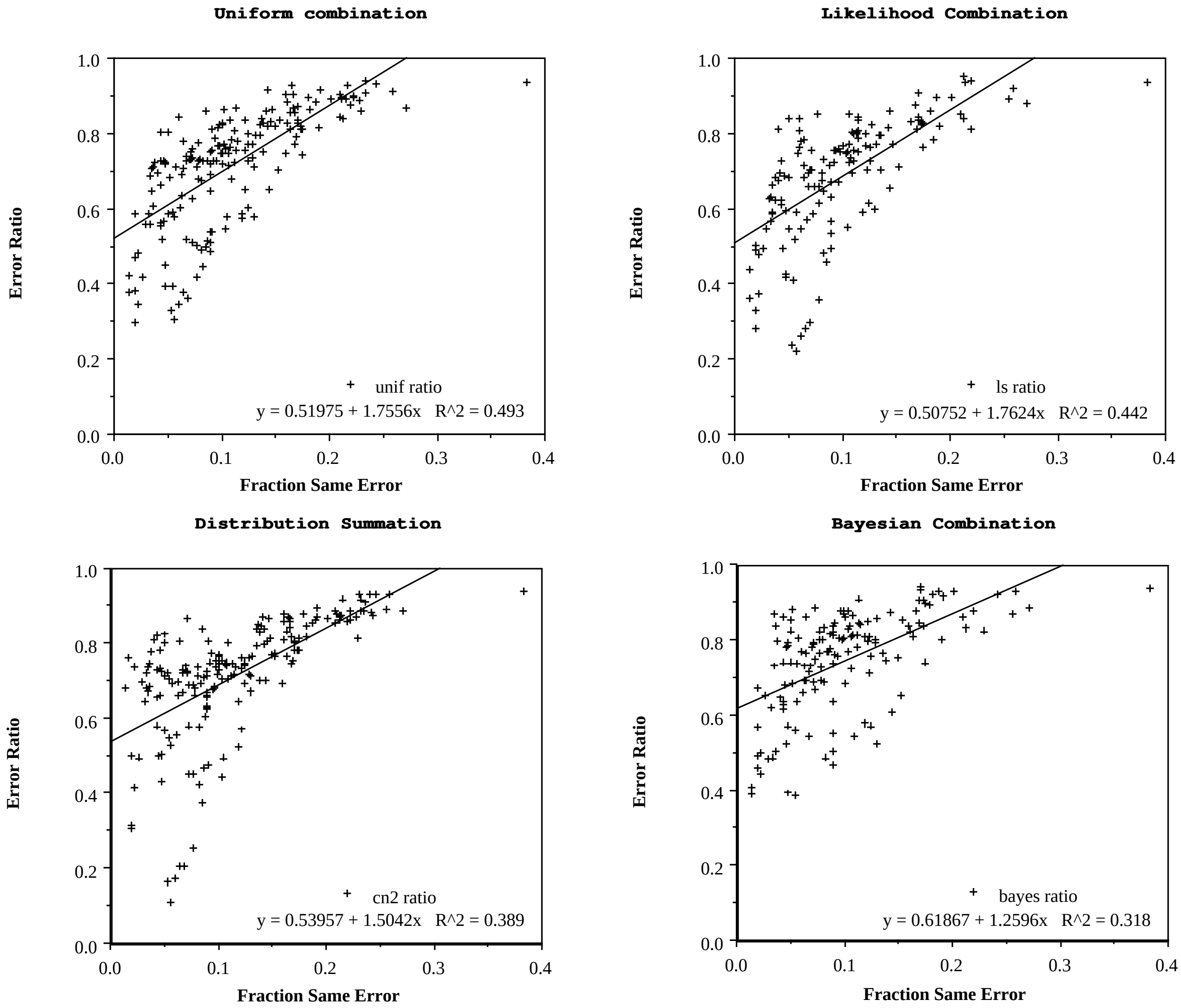

Figure 1: The figures above illustrate that greatest error reduction is obtained for ensembles which make less correlated errors (have lower values of ϕ_e). One point represents one data set - a combination of a domain and a specific training set size.

Table 1: Comparison of errors made by single decision tree and an ensemble consisting of eleven stochastically-learned decision trees combined with the Uniform Voting function. "-" indicates a statistically significant decrease in error rate. Suffixes: i: number of irrelevant attributes; e: number of training examples; a: level of attribute noise; c: level of class noise.

Domain	Base Error Rate	Number Training Examples	1 Dec. Tree Error Rate	11 Dec. Trees Uniform Voting Error Rate	
Led 8i	90.0%	30	13.1%	9.4%	
Led 17i	90.0%	30	20.9%	12.3%	–
Tic-tac-toe	34.7%	670	15.9%	5.2%	–
Krkp	48.0%	200	5.8%	5.2%	
Krk 100e	33.4%	100	3.8%	4.4%	
Krk 200e	33.4%	200	1.8%	1.7%	
Krk 160e 5a	33.4%	160	8.6%	8.6%	
Krk 320e 5a	33.4%	320	5.7%	5.7%	
Krk 160e 20c	33.4%	160	12.9%	11.8%	
Krk 320e 20c	33.4%	320	9.4%	9.6%	
Led 20a	90.0%	30	10.0%	10.0%	
Led 40a	90.0%	30	26.0%	21.7%	–
DNA	50.0%	105	17.0%	6.4%	–
Splice	46.6%	200	24.6%	12.2%	–
Mushroom	50.0%	100	1.6%	1.2%	
Hypothyroid	5.0%	200	2.3%	1.9%	
BC-Wisconsin	34.5%	200	6.5%	4.4%	–
Voting	38.0%	100	6.5%	6.4%	
Wine	60.2%	118	6.5%	2.8%	–
Iris	66.7%	50	5.4%	5.3%	
Soybean (Large)	85.4%	290	13.9%	11.9%	–
Horse-colic	36.6%	245	17.0%	14.0%	–
Hepatitis	20.4%	103	25.2%	20.4%	–
Lymph.	45.3%	110	25.0%	26.5%	
Audiology	74.7%	145	21.6%	22.3%	
Diabetes	34.9%	200	31.5%	27.0%	–
B.Cancer	29.8%	190	36.6%	35.6%	
Heart (5 class)	45.9%	200	49.9%	45.3%	–
Primary-tumor	75.3%	225	64.0%	59.8%	–

a r of 0.70 ($r^2 = 0.49$) under the null hypothesis H_0 (that there is no correlation between E_r and ϕ_e) would be for the results shown in Figure 1. For each of the four combination functions, the probability of attaining the observed r values under H_0, is less than 0.0005. Therefore, we can confidently say that the perceived linear correlation between ϕ_e and E_r is very unlikely to arise by chance under the null hypothesis. We can also compute 95% confidence intervals: $[0.38, 0.60]$ for Uniform Voting, $[0.20, 44]$ for Bayesian combination, $[0.28, 49]$ for Distribution Summation and $[0.32, 56]$ for Likelihood Combination. Therefore our results hold for many evidence combination methods.

In other experiments, we calcuated $r^2_{E_e,\phi_e}$ *within* each domain. Note that this is between ϕ_e and error *rate*, not error *ratio*. This within-domain experiment factors out the influence of optimal Bayes error rate and other domain-dependent variables. For the within-domain experiments, a separate value for ϕ_e is calculated per trial, rather than averaging over 30 trials. In these experiments, we obtained very high values for $r^2_{E_e,\phi_e}$ for most domains; up to 96.8% for tic-tac-toe.

Pruned decision tree results

It may be that the multiple models approach is able to provide such significant error reduction because non-pruned decision trees are being used. To check this, we use χ^2-pruned decision trees (Quinlan 1986) at the 99% confidence level. Figure 2 shows that under Uniform Voting, 53% of the variance in error ratio can be explained by variance in correlatedness of errors. For other evidence combination methods, the results are 39% (Likelihood Combination), 30% (Distribution Summation) and 25% (Bayesian Combination). These results are not as good as those for Uniform Voting because ϕ_e does not allow for the fact that some models may have greater voting weight than others.

Future Work

One limitation of this work is that the measure of correlatedness of errors between two models ($\phi_{ij} = p(\hat{f}_i(x) = \hat{f}_j(x), \hat{f}_i(x) \neq f(x))$) is really a function of two different factors. Abbreviating

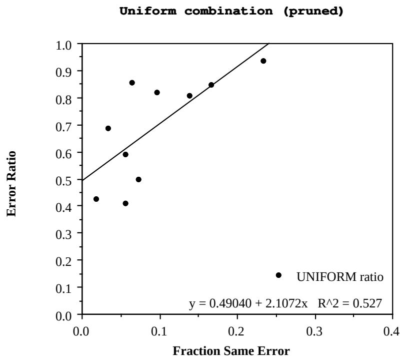

Figure 2: Tendency to make correlated errors, ϕ_e, explains 53% of the variance in degree of error reduction, E_r, for ensembles of pruned decision trees under Uniform Voting.

$\hat{f}_i(x)$ as f_i and so on, ϕ_{ij} is a product of the probability that model j makes an error $(p(f_j \neq f))$ and the conditional probability that model i makes the same error given that model j makes an error $(p(f_i = f_j, f_i \neq f \mid f_j \neq f))$. This can be seen by rewriting ϕ_{ij}:

$$\begin{aligned}\phi_{ij} &= p(f_i = f_j, f_i \neq f, f_j \neq f) \\ &= p(f_i = f_j, f_i \neq f \mid f_j \neq f) \times p(f_j \neq f)\end{aligned}$$

Therefore, future work in this area should aim to build a bivariate model of error ratio as a function of the two factors given above. However, the current work shows that even just combining these two factors in a product explains much of the variance in error reduction given the distribution of data sets in the UCI repository.

Related Work

Our work is related to the recent work of Breiman (in press) which explains that "unstable" algorithms benefit from ensemble-type combination. An algorithm is unstable if small changes in the training data lead to a great proportion of changes in classifications on test examples. The nearest neighbor algorithm is given as an example of an algorithm that is not unstable whereas decision-tree algorithms and neural-network algorithms are presumably unstable since forming ensembles for such algorithms lead to lower error rates. Breiman does not give a definition for unstability.

Our work is also related to the concept boosting work of Schapire (1990) and adaptive boosting (Freund & Schapire 1995). His boosting algorithm is the only learning algorithm which incorporates the goal of minimizing correlated errors into the learning mechanism. However, the number of training examples needed by that algorithm increases as a function of the accuracy of the learned models and could not be used on the modest sized training sets used in this paper. Adaptive boosting is constructed to require fewer training examples than boosting. Adaptive boosting works by constructing a model M_0 on the initial training set T_0. Next, another training set T_1 is constructed from T_0 by sampling from T_0 with examples that were incorrectly classified by M_0 having a greater chance of being picked. This iteration continues for a user-defined number of iterations. Adaptive boosting seems promising but has not yet been empirically tested in a thorough manner.

Finally, our work is related to that of Krogh & Vedelsby (1995). They define the ambiguity of an ensemble (on a 2-class or continuous output problem) as the variance of predictions of the ensemble: if the outputs of the models are $f_1(x), \ldots, f_T(x)$ and the mean of those values is denoted by $\bar{f}(x)$, then the ambiguity is defined as $\sum_i w_i(f_i(x) - \bar{f}(x))$ for some weights $w_1, \ldots, w_T$. This is the ambiguity for a single input x. They then extend to define the ambiguity of model j on a set of inputs (A_j) and error rate on a set of inputs (E_j). Using these concepts, Krogh & Vedelsby derive the following elegant relationship between the ensemble error (E), weighted error of the models $(\bar{E} = \sum_j w_j E_j)$ and weighted ambiguity of the models $(\bar{A} = \sum_j w_j A_j)$

$$E = \bar{E} - \bar{A}$$

This means that *the ensemble is able to reduce the weighted error rates of the members of the ensemble by the degree of ambiguity.* Therefore, their work is modeling the degree of error reduction but not with respect to the single model as is done in our work - their work models error reduction with respect to the average weighted error of the individual models.

Conclusions

The paper presents empirical evidence that for decision trees the tendency to make uncorrelated errors $(1 - \phi_e)$ in an ensemble explains much of the degree of error reduction. We show that degree of error reduction can be well modeled by the tendency to make uncorrelated errors for pruned and unpruned decision tree ensembles for various combination functions. These results provide empirical validation (for trees) of the widely held belief that the multiple models approach is able to do better than the single models approach when the learned models make uncorrelated errors.

References

Ali, K., and Pazzani, M. 1995. Learning multiple relational rule-based models. In Fisher, D., and Lenz, H., eds., *Learning from Data: Artificial Intelligence and Statistics, Vol. 5.* Fort Lauderdale, FL: Springer-Verlag.

Ali, K., and Pazzani, M. in press. Error reduction through learning multiple descriptions. *Machine Learning* ?

Ali, K. 1996. *Learning Probabilistic Relational Concept Descriptions.* Ph.D. Dissertation, University of California, Irvine.

Breiman, L. in press. Bagging predictors. *Machine Learning* ?

Buntine, W. 1990. *A Theory of Learning Classification Rules.* Ph.D. Dissertation, School of Computing Science, University of Technology.

Clark, P., and Boswell, R. 1991. Rule induction with CN2: Some recent improvements. In *Proceedings of the European Working Session on Learning, 1991.* Pitman.

Duda, R.; Gaschnig, J.; and Hart, P. 1979. Model design in the PROSPECTOR consultant system for mineral exploration. In Michie, D., ed., *Expert systems in the micro-electronic age*. Edinburgh, England: Edinburgh University Press.

Freund, Y., and Schapire, R. 1995. A decision-theoretic generalization of on-line learning and an application to boosting. In Vitanyi, P. E., ed., *Lecture Notes in Artificial Intelligence, Vol. 904.* Berlin, Germany: Springer-Verlag.

Hansen, L., and Salamon, P. 1990. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12, 10:993–1001.

Kononenko, I., and Kovacic, M. 1992. Learning as optimization: Stochastic generation of multiple knowledge. In *Machine Learning: Proceedings of the Ninth International Workshop.* Aberdeen, Scotland: Morgan Kaufmann.

Krogh, A., and Vedelsby, J. 1995. Neural network ensembles, cross validation and active learning. In *Advances in Neural Information Processing Systems 7.* Cambridge, MA: MIT Press.

Kruskal, W., and Tanur, J. 1978. *International encyclopedia of statistics.* New York, NY: Free Press.

Kwok, S., and Carter, C. 1990. Multiple decision trees. *Uncertainty in Artificial Intelligence* 4:327–335.

Madigan, D., and York, J. 1993. Bayesian graphical models for discrete data. Technical Report UW-93-259, Statistics Department, University of Washington.

Perrone, M. 1993. *Improving Regression Estimation: Averaging Methods for Variance Reduction with Extensions to General Convex Measure Optimization.* Ph.D. Dissertation, Department of Physics, Brown University.

Quinlan, R. 1986. Induction of decision trees. *Machine Learning* 1, 1:81–106.

Schapire, R. 1990. The strength of weak learnability. *Machine Learning* 5, 2:197–227.

Smyth, P.; Goodman, R.; and Higgins, C. 1990. A hybrid rule-based/Bayesian classifier. In *Proceedings of the 1990 European Conference on Artificial Intelligence.* London, UK: Pitman.