On the Link between Error Correlation and Error Reduction in Decision Tree Ensembles

Kamal Ali
Department of Information and Computer Science
University of California, Irvine, CA, 92717
ali@ics.uci.edu, pazzani@ics.uci.edu
(714)824-3491, (714)824-5888

December 4, 1995

Abstract

Recent work has shown that learning an ensemble consisting of multiple models and then making classifications by combining the classifications of the models often leads to more accurate classifications then those based on a single model learned from the same data. However, the amount of error reduction achieved varies from data set to data set. This paper provides empirical evidence that there is a linear relationship between the degree of error reduction and the degree to which patterns of errors made by individual models are uncorrelated. Ensemble error rate is most reduced in ensembles whose constituents make individual errors in a less correlated manner. The second result of the work is that some of the greatest error reductions occur on domains for which many ties in information gain occur during learning. The third result is that ensembles consisting of models that make errors in a dependent but "negatively correlated" manner will have lower ensemble error rates than ensembles whose constituents make errors in an uncorrelated manner. Previous work has aimed at learning models that make errors in a uncorrelated manner rather than those that make errors in an "negatively correlated" manner. Taken together, these results help provide an understanding of why the multiple models approach yields great error reduction in some domains but little in others.

Keywords: Multiple models, Decision Trees, Combining classifiers.

1 Introduction

Recent years have seen much work in learning multiple models for the purpose of reducing classification error.[1] Studies involving the use of multiple models typically learn a set of models from one set of training examples. This ensemble makes classifications by combining the classifications of its constituents. The error rate of this ensemble is usually compared to that of a special, single model that results from using a deterministic learning procedure on the same training examples. Most of the empirical work on multiple models has shown that the ensemble is able to achieve more accurate classifications than the single model.

Besides the impressive empirical evidence that shows that classification error rates can be reduced by learning and using multiple models, there are also relevant theoretical results. Breiman (submitted) shows for regression that the expected mean square error (MSE) of the ensemble must be lower than the average MSE of its constitutents. Hansen & Salamon (1990) show that if the models make errors independently, and if they all have the same error rate and if that error rate is less than 0.5, then the expected error rate of the ensemble will decrease monotonically as a function of the number of classifiers in the ensemble. Perrone & Cooper (1993) have the strongest result: they show for regression, that if the models all are unbiased[2] and all make errors completely independently, then the mean square error (MSE) of the ensemble will equal the average MSE of the constitutents divided by the number of models! Buntine (1990) applies Bayesian probability theory (e.g. Bernardo & Smith, 1994) to classification and shows that the expected posterior probability of a class given a test example can be computed by combining the posterior probabilities of all the hypotheses in the hypothesis space. Empirical results (Buntine, 1990; Ali & Pazzani, 1995; Oliver & Hand, 1995) use a small set of highly probable models to approximate the result of combining the posterior probabilities over the entire hypothesis space and show that the ensemble achieves a lower error rate than the special single model.

In our earlier work (Ali & Pazzani, submitted) we demonstrated that using an ensemble it was possible to achieve an error-rate just one-seventh that of the single model error-rate. This was accomplished on the wine domain from the UCI repository (Murphy & Aha, 1992) using an ensemble consisting of eleven rule-set models. However, our results and those of Breiman (submitted) indicate that the amount of error reduction varies greatly. For some domains (e.g. Iris, Breast Cancer) the multiple models approach does not lead to any reduction in error.

[1] Some examples are:
Decision trees: Kwok & Carter, 1990; Buntine, 1990; Oliver & Hand, 1995; Breiman, 1994.
Rules: Kononenko & Kovacic, 1992; Kovacic, 1994; Smyth *et al.*, 1990.
Rule sets: Gams, 1989; Ali & Pazzani, 1995.
Neural networks: Hansen & Salamon, 1990; Perrone & Cooper, 1993; and many others.
Bayesian networks: Madigan & York, 1993.
Regression: Perrone & Cooper, 1993; Breiman, submitted.

[2] In regression, this has a precise definition. It means that the expected value predicted by the classifier should equal the true expected value.

This leads to the main question addressed in this paper: "What influences the *amount* of error reduction?" In particular, we are interested in exploring the widely held belief (articulated by Hansen & Salamon (1990)) that error is most reduced for domains for which the errors made by the models are made in an independent manner. This is also echoed in Kong & Dietterich (1995) in which they hypothesize that error-correcting output codes are able to reduce error because they rely on learning several functions (models) that vote to make a classification and that those functions make errors in an uncorrelated manner. The validation of this hypothesized link between independence (uncorrelatedness) of individual model errors and overall error reduction is the main goal of this paper.

The rest of the paper is organized as follows: Section 2 defines our measures of correlatedness and degree of error reduction. The next section presents our method for learning ensembles and the following section presents our methods for combining ensembles. Section 5 presents our three main results.

2 Error reduction and error correlation

Now we present precise definitions of the degree of error reduction (E_r) and the degree of error correlatedness (ϕ_e). Two obvious measures comparing the error of the ensemble (E_e) to the error of the single model (E_s) are error difference ($E_s - E_e$) and error ratio ($E_r = E_e/E_s$). We use error ratio because it reflects the fact that it becomes increasingly difficult to obtain reductions in error as the error of the single model approaches zero. Error ratios less than 1 indicate that multiple models approach was able to obtain a lower error rate than the single model approach. The lower the error ratio, the greater the error reduction and the better the situation.

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. ϕ_e is then just the average of the probability of making the same kind of error taken over all pairs in the ensemble. That is,

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

The higher the values of ϕ_e, the more correlated the errors made by members of the ensemble. The values of ϕ_e for the data sets presented in Table 2 were estimated on the *test* set of examples. Therefore, they cannot be used by the learning algorithm but provide us with an understanding of why error is reduced more in some data sets.

There is an intimate link between making uncorrelated errors and making errors in a statistically independent manner. Let C denote the number of classes and let ϕ^* denote the

Table 1: Relationship between correlatedness and statistical independence.

$\phi_e > \phi^*$	Positively correlated	Dependent errors
$\phi_e = \phi^*$	Uncorrelated	Independent errors
$\phi_e < \phi^*$	Negatively correlated	Dependent errors

following special value of ϕ_e:

$$\phi^*(\mathcal{F}) = \sum_{i=1}^{T} \sum_{j \neq i}^{T} \sum_{k=1}^{C} p(f(x) = k) \times \left[\sum_{l \neq k} p(\hat{f}_i(x) = l | f(x) = k) \times p(\hat{f}_j(x) = l | f(x) = k) \right] \quad (2)$$

This is the value of ϕ_e that would be obtained if all the members of the ensemble made errors in a pairwise, statistically independent manner for each class in the data. This value is used to define the meaning of "negatively correlated" as shown in Table 1.

Some authors (Hansen & Salamon, 1990; Perrone, 1993) have demonstrated that making errors in an uncorrelated (independent) manner leads to a lower error rate for the ensemble and produces some desirable results relating to error reduction. Kong & Dietterich (1995) attribute the success of their error-correcting output code method to its ability to learn functions that make uncorrelated errors. However, our analysis above suggests that because ϕ^* is not the lowest possible value obtainable, one should aim to learn ensembles whose models make errors in an "negatively correlated" manner. In Section 5.3 we present further arguments for the hypothesis that ensembles whose members make errors in an negatively correlated manner will have lower ensemble error rates than ensembles whose members make errors in an uncorrelated (independent) manner.

3 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. Pruning is not used in this section because we do not want the error reductions to be confounded with the pruning method. Section 6 however shows that even if pruning is used, there is still a correlation between the degree of error reduction and the degree to which models make uncorrelated errors. Unknown attribute values are handled by the method of token averaging (Quinlan, 1986).

Stochastic search is used to generate multiple trees. We consider all decision tree splits whose resultant entropy (Quinlan, 1986) is within some factor β of the entropy of the split with the lowest entropy. For our experiments, we set this factor to 1.25. The probability of choosing a split from this set is proportional to 1/Entropy.[3] We have not experimented

[3] To prevent zero values for Entropy, we used the Laplace approximation for the probabilities involved in the Entropy expression. Briefly, the Laplace approximation for the probability of some discrete random

with other values of β - future work should check if the negative correlation between degree of error reduction and correlatedness of errors holds for other values.

4 Evidence combination

The only other decision one needs to make in making a stochastic version of an algorithm is how to combine evidence and classifications of the learned models in order to make an overall classification by the ensemble. 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. For the other combination functions, each tree must provide a measure of confidence in addition to its classification.

- 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 *at al.*, 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 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

variable which has been observed to occur in f of T trials is $\frac{f+1}{T+k}$ where k denotes the number of possible values for the variable.

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).

5 Experimental results

For our experiments we chose domains from the UCI repository of machine learning databases (Murphy & Aha, 1992) ensuring that at least one domain from each of the major groups (molecular biology, medical diagnosis ...) was chosen. These include molecular-biology domains (2), medical diagnosis domains (7), relational domains (6 variants of the King-Rook-King (KRK) domain, Muggleton *et al.*, 1989), a chess domain with a "small disjuncts problem" (KRKP; Holte *et al.*, 1989), and attribute-value domains (4 LED variants and the tic-tac-toe problem).

For most of the domains tested here, we used thirty independent trials, each time training on two-thirds of the data and testing on the remaining one-third. The exceptions to this are the DNA promoters domain for which leave-one-out testing has traditionally been used and we follow this tradition to allow comparability. Whenever possible we tried to test learned models on noise-free examples (including noisy variants of the KRK and LED domains) but for the natural domains we tested on possibly noisy examples. The "large" variant of the Soybean data set (Murphy & Aha, 1992) was used and the 5-class Heart data set variant was used.

5.1 Link between error reduction and error correlation

Table 2 presents results using 29 data sets from 21 domains. (We distinguish the terms "data set" and "domain" in that a data set also involves specifiying parameters such as training set size, level of class noise etc.) For 72% of the data sets in Table 2 there is a significant reduction in error rate when classifications are made using an ensemble of eleven trees (combined using Uniform Voting). Significant reductions in error are labeled by "–"; significant increases by "+." No significant change in error occurs on the remaining data sets. Hence, the ensemble approach never significantly increases error rate.

Table 2: Comparison of errors made by single decision tree and an ensemble consisting of eleven stochastically-learned decision trees combined with the Uniform Voting function. Suffixes: i: number of irrelevant attributes; e: number of training examples; a: level of attribute noise; c: level of class noise.

Domain	Base	Number	1 Dec. Tree	11 Dec. Trees	
	Error	Training	Error	Uniform Voting	
	Rate	Examples	Rate	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	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	45.9%	200	49.9%	45.3%	–
Primary-tumor	75.3%	225	64.0%	59.8%	–

However, the main point of this paper is not to demonstrate that error is reduced due to the multiple models approach. Rather, we seek to explain the amount of error reduction as a function of the tendency to make correlated errors, ϕ_e. 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. If the error ratios and ϕ_e values for the 19 data sets for which there was a statistically significant degree of error reduction are plotted in a scatter-plot and a least mean-squares linear fit is done, it can be determined that the tendency to make correlated errors explains 60% of the variance in the error ratio variable ($r^2_{E_r,\phi_e} = 0.60$). This is empirical evidence for the hypothesis that there is a negative correlation between the degree to which error is reduced and the degree to which individual model errors in the ensemble are made in an correlated manner. However, it is better to conduct several trials to estimate the statistic $r^2_{E_r,\phi_e}$ so we conducted 10 trials. Within each of these "meta-trials," 30 trials per data set were run. For the i-th meta-trial we decided to use $i \times 10\%$ of the training data. So, for example, on the Tic-Tac-Toe domain, the original training set size was 670 training examples so this was augmented by 30 trials at 67 examples, 30 trials at 134 training examples etc.

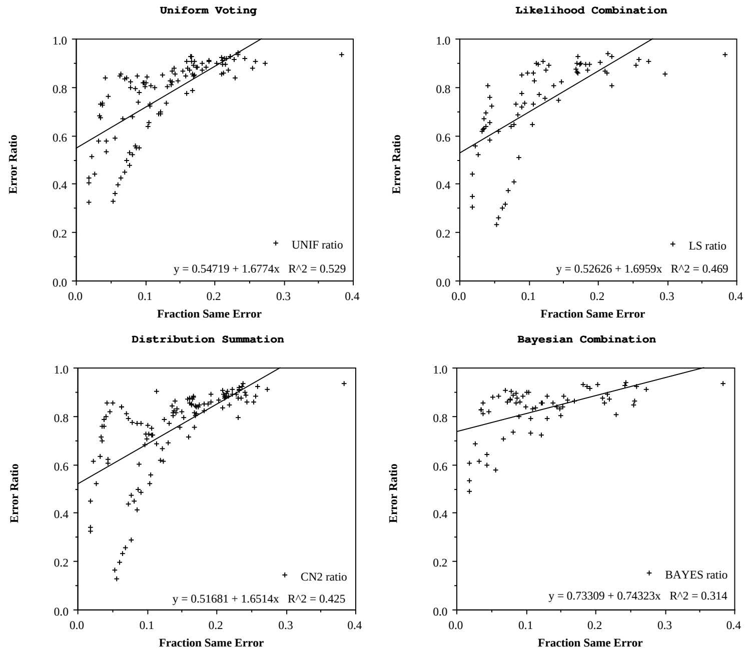

Figure 1: The figures above illustrate that greatest error reduction is obtained for ensembles which make less correlated errors (have lower values of ϕ_e). One point represents one data set - a combination of a domain and a specific training set size.

Figure 1 shows that under the Uniform Voting combination function (top left of Figure 1), 49% of the variance in *amount* of error reduction is explained by the tendency to make the same correlated errors. This value is more reliable than the 60% value for r^2 mentioned earlier because it uses approximately ten times as many data sets. The figure 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 a r of 0.70 ($r^2 = 0.49$) under the null hypothesis, H_0, would be for 162 degrees of freedom. In this case, the null hypothesis would be that the population correlation, ρ, between E_r and ϕ_e, given that there is a significant degree of error reduction is 0. For each of the four combination functions, the probability of attaining the observed r values under H_0, is less than 0.0005 (120 degrees of freedom were used). Therefore, we can confidently say that the perceived linear correlation between ϕ_e and E_r is very unlikely to arise by chance. The 95% confidence intervals around r^2 are $[38\%, 60\%]$ for Uniform Voting, $[20\%, 44\%]$ for Bayesian combination, $[28\%, 49\%]$ for Distribution Summation and $[32\%, 56\%]$ for Likelihood Combination.

That we can empirically discover the negative correlation between amount of error reduction and tendency to make correlated errors is quite encouraging given that the data sets vary widely in optimal Bayes error and along other dimensions. Secondly, ϕ_e is a pairwise measure, whereas what the error rate under Uniform Voting counts is the proportion of the test examples on which at least six models made an error (assuming an ensemble size of eleven). Another limitation of ϕ_e is that it assumes all models have equal voting weight. This is only true under the Uniform Voting combination function and that is why the r^2 under that function is higher than under other functions.

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 which may vary from domain to domain. 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.

In order to gain insight into why ϕ_e explains so much of the variance in error ratio consider the simpler problem of modeling variation in error *rate* within a given domain. Assume that the simplest evidence combination method (Uniform Voting) is used and that the data set contains two classes and that the ensemble contains just two models. In this situation, an ensemble error occurs if both the models make an error or if the models disagree and the tie is broken so as to cause an error. Assume that ties occur on a negligible proportion of the cases. Under these assumptions, ϕ_e is an exact measure of ensemble error (E_e). As ϕ_e is a pairwise measure, how well it models within-dataset ensemble error depends on the size of the ensemble.

To summarize: our results provide an explanation of why the multiple models approach leads to great error reduction in some domains but hardly any in other domains. The results show that there is a negative correlation between the amount of error reduction and the amount of correlatedness of errors - the less correlated the individual model errors, the better the ensemble is at reducing error.

5.2 Gain Ties

The amount of error correlation provides a post-hoc way of understanding the degree of error reduction. Now we want to predict *during learning* the expected amount of error reduction. We seek to understand why the stochastic learning algorithm produces models that make less correlated errors in some data sets.

The motivation for postulating this hypothesis is the observation that each time the stochastic generation method is run, it uses the same training data. However, it is able to generate different descriptions because it randomly picks from the decision tree nodes whose gain is within some factor β $(\beta \in (1, \infty))$ of the entropy of the best node. If there are many such nodes then the possibility for syntactic variation from description to description is greater. It is our hypothesis that greater syntactic variety leads to descriptions that make less correlated errors. Hence, if we can measure (during learning) the amount of potential syntactic diversity, we can estimate the degree to which the resulting models will make

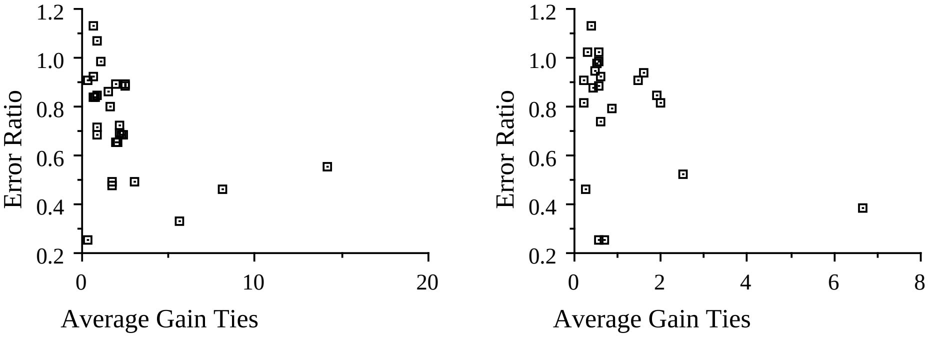

Figure 2: Error ratio as a function of average gain ties for decision trees (left) and rule sets (right) (Ali & Pazzani, submitted). The ensembles of decision trees contained eleven, stochastically learned decision trees with respect to the entropy gain function. The ensembles of eleven rule sets were learned using stochastic hill-climbing and combined using likelihood combination. Similar plots are obtained for other evidence combination methods.

correlated errors.

As a first approximation measure of the amount of syntactic variety in a data set as experienced by a learning algorithm, consider the number of literals that tie for the highest information gain. If n literals tie for gain, that event is recorded as representing $n-1$ ties in gain. The total number of ties experienced during learning a model is then divided by the number of literals in the model to produce the quantity g - the "average number of gain ties" for that data set. Values of g used in Figure 2 represent averages over thirty trials. A large number of such ties are a problem for a hill-climbing deterministic learner but represent an opportunity for the multiple model learner. Figure 2 plots error ratio as a function of average gain ties. The figure shows that some of the largest reductions in error are obtained for data sets for which such ties are frequent. For instance, there were, on average, 14.2 gain ties on the Wine data set and 8.1 for the Splice data set! This underscores the difficulty experienced by a greedy, hill-climbing algorithm on such data sets. Even with further look-ahead, the single-model approach still can only produce a single model as its output. Therefore, unless the sample size is very large, the single-model approach will still be at a disadvantage. The figure also shows that a high average value for ties in gain appears to be a sufficient but not *necessary* condition for significant reduction of error. For example, the multiple models approach is able to achieve low error ratios on the Tic-Tac-Toe and the noise-free LED variants (bottom left of Figure 2) even though there are not many ties in gain for those data sets.

The gain ties measure is a rough approximation of the potential for syntactic diversity and the hypothesized resulting diversity of errors. For instance, the measure can be fooled by multiple definitions of attributes. We tried some variants that took the variability in the extensions of the candidate decision nodes into account but found no measure that provided

Table 3: An arbitrary arrangement of individual model errors on 100 test examples. A "X" indicates an individual model error. An ensemble error occurs for test examples (columns) in which there are more than 2 individual model errors.

	Eg1	Eg2		Eg99	Eg100
Model1	X	X			X
Model2		X			X
Model3	X		...	X	
Model4				X	X
Model5		X			

a significantly better estimation of error reduction. Other measures that counted "near gain-ties" or gain ties weighted by the number of examples at that node also did not yield a better estimate of error reduction.

5.3 Negatively Correlated Errors

In this section we consider whether making errors in a negatively-correlated manner leads to lower values of error reduction than if the errors are made in an uncorrelated manner. Consider what an optimimal arrangement of errors (that minimizes ensemble error rates) would look like. We cannot vary each model's error rate but we can permute the examples on which it makes errors.

Consider an arbitrary pattern of errors as shown in Table 3 for 5 learned models and 100 test examples. Assume that Uniform Voting is used so an ensemble error occurs if there are more than 2 errors for any example. Therefore, in order to minimize the ensemble error rate, the models should make errors in a pattern that minimizes the number of columns in which more than 2 errors are made.

Now, because we are able to rearrange the errors but we are not able to modify the error rates it follows that we can permute each row. The ensemble error minimization procedure operates by ordering the models - most error-prone first. Then the errors of the second model are permuted so that as many of them as possible occur on examples that were correctly classified by the first model. That is, the models should make errors on *disjoint* subsets (to as great a degree as possible) rather than on independently drawn subsets. This process continues so that for each model we arrange for the mistakes to be made on examples on which the fewest mistakes have been made by previous models. However, once the number of errors on an example exceeds $\lfloor \frac{T}{2} \rfloor$ (T is the number of models) then an ensemble error is conceded on that example. Then in order to keep minimizing ensemble error rate, it is best to arrange for subsequent models to make their errors on such "conceded" examples. From this analysis it becomes clear that in order to minimize the number of ensemble errors, it is better for the constituent models to make errors in a *dependent* but negatively correlated way rather than in an independent (uncorrelated) way.

ϕ_e is a perfect measure of ensemble error rate *within* a domain given that only two models are in the ensemble, the domain only contains two classes and that ties occur on a negligible proportion of the examples. However, for ensembles of larger sizes, consider how the arrangement of errors such as that in Table 3 impacts the ability of ϕ_e to measure ensemble error rate. The number of ensemble errors simply counts the number of columns in the table on which more than $\lfloor \frac{T}{2} \rfloor$ errors occurred. Therefore, any rearrangement in the pattern of errors on the columns in which $\lfloor \frac{T}{2} \rfloor$ or fewer errors occurred has no bearing on the number of ensemble errors as long as the rearrangement does not cause more than $\lfloor \frac{T}{2} \rfloor$ error to occur in any given column. But these rearrangements do have an impact on ϕ_e which is simply a pairwise (2nd order) measure. This explains why ϕ_e does not do a perfect job of modeling error rate within data sets or of modeling error ratio between data sets.

6 Results with Pruning

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. Using 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.

7 Related work

Our work is related to the recent work of Breiman (1994, submitted) 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 (e.g. Aha, 1990) 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.

Kwok & Carter (1990) have also done related work showing that decision tree ensembles whose trees' root nodes were varied led to better results than ensembles whose trees had variations further down the tree. Their conclusion (using two domains) was that greater syntactic diversity led to lower ensemble error rates. Our gain ties measure is an attempt to quantitatively measure the potential for syntactic diversity. However, it would be best to measure diversity in the functional space. Work in functional diversity has been done by Perrone & Cooper (1993) although they do not incorporate the goal of functional diversity into their learning algorithm or offer an explanation of why it is possible to learn functionally more diverse ensembles on some domains.

Our work on correlation is also related to that of Kong & Dietterich (1995) in which they attribute the power of the Error-Correcting Output Codes (ECOC) approach to the fact that it involves learning several approximations of the target function f and then voting among those approximations. Kong and Dietterich hypothesize that the ECOC approach works because the approximations make uncorrelated errors. However, they use "uncorrelated" to mean "non-identical" and their work is not concerned with a quantitative measure of error correlation or with explaining the amount of error reduction.

Finally, our work is 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. However, adaptive boosting relies on the assumption that the data is not overfitted. If the first model achieves 100% accuracy over the training set, the adaptive boosting algorithm terminates having just learned a single, overfitted model.

8 Conclusions

The paper provides an understanding of why the multiple models approach leads to striking reductions in error on some domains whilst on other domains there is no reduction in error. Our finding is that the *amount* of error reduction is negatively correlated with the degree to which the models in the ensemble make errors in a correlated manner. We use quantitative definitions for error reduction and the degree to which models make errors in a correlated manner to empirically show that there is a linear relationship between these two variables. The results are based on experiments using 260 data sets from 20 domains crossed with four evidence combination methods. Although this paper only presents results for decision trees, our earlier work (Ali & Pazzani, submitted) shows that the linear relationship between error reduction and amount of error correlation also holds for models consisting of rule sets.

But why does stochastic learning produce models in one domain whose errors are uncorrelated whilst in another domain it produces models with highly correlated errors? This is answered by the second result of the paper: in domains in which many ties in gain are experienced, the errors of the resulting models are relatively uncorrelated and so the reduction in error is relatively large. Although this simple measure - gain-ties - has limitations, it is as useful in predicting error reduction as some of its more complex variants.

The third result of the paper is that our analysis predicts that ensembles whose models make errors in a dependent but "negatively correlated" manner should have lower ensemble error rates than ensembles whose models make errors in an independent (uncorrelated) manner. This supersedes previous beliefs that one of the goals of multiple models learning is to learn models that make errors in an independent (uncorrelated) manner.

Acknowledgements- We would like to acknowledge the help of Wray Buntine and David Wolpert in reading early versions of this paper and the support of NSF grant # IRI-9310413.

9 References

Aha, D. (1990.) *A Study of Instance-Based Algorithms for Supervised Learning Tasks.* Doctoral dissertation. Department of Information and Computer Science, University of California, Irvine, California, USA.

Ali, K., & Pazzani, M. (1995a.) Learning Multiple Relational Rule-based Models. In Fisher, D., & Lenz, H. (Eds.), *Learning from Data: Artificial Intelligence and Statistics, Vol. 5.* Fort Lauderdale, FL: Springer-Verlag.

Ali, K.M., & Pazzani, M. (submitted.) Error Reduction through Learning Multiple Descriptions *Machine Learning Journal, , .*

Breiman, L. (1994.) *Heuristics of instability in model selection..* (Technical Report ?). University of California at Berkeley, Statistics Department.

Breiman, L. (submitted.) Bagging Predictors *?, ?,* ?.

Buntine, W.L. (1990.) *A Theory of Learning Classification Rules.* Doctoral dissertation. School of Computing Science, University of Technology, Sydney, Australia.

Drucker, H., Cortes, C., Jackel, L., LeCun, Y. & Vapnik V. (1994.) Boosting and Other Machine Learning Algorithms In *Machine Learning: Proceedings of the Eleventh International Conference* New Brunswick, NJ: Morgan Kaufmann.

Freund, Y., & Schapire, R.E. (1995.) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. In Vitanyi, P. (Ed.), *Lecture Notes in Artificial Intelligence, Vol. 904.* Berlin, Germany: Springer-Verlag.

Gams, M., & Petkovsek, M. (1988.) Learning From Examples in the Presence of Noise In *8th International Workshop; Expert Systems and their applications, Vol. 2* Avignon, France: .

Hansen, L.K, & Salamon, P. (1990.) Neural Network Ensembles *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12, 10,* 993-1001.

Holte, R., Acker, L., & Porter, B. (1989.) Concept Learning and the Problem of Small Disjuncts. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence* Detroit, MI: Morgan Kaufmann.

Howell, D. (1987.) *Statistical Methods for Psychology.* Boston, MA: Duxbury Press.

Kong, E.B., & Dietterich, T.G. (.) Machine Learning Bias, Statistical Bias, and Statistical Variance Decision Tree Algorithms In : .

Kononenko, I., & Kovacic, M. (1992.) Learning as Optimization: Stochastic Generation of Multiple Knowledge In *Machine Learning: Proceedings of the Ninth International Workshop* Aberdeen, Scotland: Morgan Kaufmann.

Kovacic, M (1994.) MILP - a stochastic approach to Inductive Logic Programming In *Proceedings of the Fourth International Workshop on Inductive Logic Programming* Bad Honnef/Bonn, Germany: GMD Press.

Kwok, S., & Carter, C. (1990.) Multiple decision trees *Uncertainty in Artificial Intelligence, 4,* 327-335.

Madigan, D., & York, J. (1993.) *Bayesian Graphical Models for Discrete Data.* (Technical Report UW-93-259). University of Washington, Statistics Department.

Muggleton, S., Bain, M., Hayes-Michie, J., & Michie, D. (1989.) An experimental comparison of human and machine-learning formalisms. In *Proceedings of the Sixth International Workshop on Machine Learning* Ithaca, NY: Morgan Kaufmann.

Oliver, J.J., & Hand, D.J. (1995.) On Pruning and Averaging Decision Trees In *Machine Learning: Proceedings of the Twelfth International Conference on Machine Learning* Tahoe City, CA: Morgan Kaufmann.

Perrone, M.P., & Cooper, L.N. (1993.) When Networks Disagree: Ensemble Methods for Hybrid Neural Networks. In Mammone, R.J. (Ed.), *Neural Networks for Speech and Image Processing.* : Chapman Hall.

Quinlan, R. (1986.) Induction of Decision Trees. *Machine Learning, 1, 1,* 81-106.

Smyth, P., Goodman, R.M., & Higgins, C. (1990.) A Hybrid Rule-Based/Bayesian Classifier In *Proceedings of the 1990 European Conference on Artificial Intelligence* London, UK: Pitman.