JOURNAL OF APPLIED MATHEMATICS AND DECISION SCIENCES, 8(3), 141–154

Variance Reduction Trends on 'Boosted' Classifiers

VIRGINIA WHEWAY† vlw04@uow.edu.au
School of Mathematics & Applied Statistics, the University of Wollongong

Abstract. Ensemble classification techniques such as *bagging*, (Breiman, 1996a), *boosting* (Freund & Schapire, 1997) and *arcing* algorithms (Breiman, 1997) have received much attention in recent literature. Such techniques have been shown to lead to reduced classification error on unseen cases. Even when the ensemble is trained well beyond zero training set error, the ensemble continues to exhibit improved classification error on unseen cases. Despite many studies and conjectures, the reasons behind this improved performance and understanding of the underlying probabilistic structures remain open and challenging problems. More recently, diagnostics such as *edge* and *margin* (Breiman, 1997; Freund & Schapire, 1997; Schapire et al., 1998) have been used to explain the improvements made when ensemble classifiers are built. This paper presents some interesting results from an empirical study performed on a set of representative datasets using the decision tree learner C4.5 (Quinlan, 1993). An exponential-like decay in the variance of the edge is observed as the number of boosting trials is increased. i.e. boosting appears to 'homogenise' the edge. Some initial theory is presented which indicates that a lack of correlation between the errors of individual classifiers is a key factor in this variance reduction.

Keywords: Classification, ensemble classifiers, boosting, variance reduction.

1. Introduction

This paper is concerned with the classification problem, whereby a model builder (classifier) is presented with a training set comprising of a series of n labelled training examples of the form $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, with $y_i \in (1, \ldots, k)$. The classifier's task is to use these training examples to produce an hypothesis, $h(\mathbf{x})$, which is an estimate of the unknown relationship $y = f(\mathbf{x})$. This 'hypothesis' then allows future prediction of y_i given new input values of $\mathbf{x}$. A classifier built by combining individual $h(\mathbf{x})$'s to form a single classifier is known as an ensemble. Whilst there are many ensemble building methods in existence, this discussion focusses on the method of boosting which is based on a weighted subsampling of the training examples.

† Requests for reprints should be sent to Virginia Wheway, School of Mathematics and Applied Statistics, the University of Wollongong, Australia.

Introduced by Freund and Schapire in 1997, boosting is recognised as being one of the most significant recent advances in classification (Freund & Schapire, 1997). Since its introduction, boosting has been the subject of many theoretical and empirical studies (Breiman, 1996b; Quinlan, 1996; Schapire et al., 1998). Empirical studies have shown that ensembles grown from repeatedly applying a learning algorithm over different randomly chosen subsamples of the data of size n improves generalisation error for unstable learners (i.e. methods where a small change in the input data leads to large changes in the learned classifier).

2. Current Explanations of the Boosting Mechanism

Ensemble classifiers and the reasons for their improved classification accuracy have provided a fertile ground for research and significant gains may still be made if these reasons are addressed. In theory, as the combined classifier complexity increases, the gap between training and test set error should increase. However, this is not reflected in empirical studies. There is strong empirical support for the view that overfitting is less of a problem (or perhaps a different problem) when boosting and other resampling methods are used to improve a learner. Some authors have addressed this issue via bias and variance decompostions in an attempt to understand the stability of a learner (Breiman, 1997; Breiman, 1996b; Friedman, 1997).

Boosting is an iterative procedure which trains a classifier over the n weighted observations. Boosting begins with all training examples being weighted equally. (i.e. $\frac{1}{n}$) At the $m+1$-th iteration, examples which were classified incorrectly at the m-th iteration have their weight increased multiplicatively so that the total weight on incorrect observations is equal to 0.5 for the $m+1$-th iteration. Hence, the learning algorithm will be given more opportunity to explore areas of the training set which are more difficult to classify. Hypotheses from these parts of the space make fewer mistakes on these areas and play an important role in prediction when all hypotheses are combined via weighted voting. Weighted voting takes places by having each hypothesis assigned a voting weight which is a function of the error made on that particular hypothesis. Hypotheses which make fewer errors are given a higher voting weight when the ensemble is formed. Accuracy of the final hypothesis depends on the accuracy of *all* the hypotheses returned at each iteration and the method exploits hypotheses that predict well in more difficult parts of the instance space. An advantage of boosting is that it does not require any background knowledge of the performance of the underlying classification algorithm. Refer to Table 1 for details of the boosting algorithm.

Input: n training instances x_i with labels y_i.
Maximum trials, M. Classifier, H.
Initialization:
All training instances begin with selection weight $w_i^0 = 1/n$.
Repeat for M trials:
form classifier, h_m, using weighted training data and H;
put ϵ_m = weighted error for h_m on the training data
($\epsilon_m = \sum_{i=1}^{n} w_i^m \times I(H_m(\mathbf{x}_i \neq y_i)$ if $\epsilon_m > 1/2$,
discard h_m and stop boosting;
If $\epsilon_m = 0$ then
h_m gets infinite weight.
classifier h_m voting weight = β_m, where
$\beta_m = \log \frac{\epsilon_m}{1-\epsilon_m}$
re-weight training instances:
if $h_m(x_i) \neq y_i$ then
$w_i^{m+1} = w_i^m/(2\epsilon_m)$
else
$w_i^{m+1} = w_i^m/2(1-\epsilon_m)$
Unseen instances are classified by voting the ensemble of classifiers h_m with weights $\log(\frac{1}{\beta_m})$.

The following comments and conclusions on boosting and ensemble classification have been made to date.

- Boosting is capable of variance and bias reduction (Breiman, 1996b).
- Breiman (1996b) claims the main effect of the adaptive resampling when building ensembles is to reduce variance, where the reduction comes from adaptive resampling and not the specific form of the ensemble forming algorithm.
- A weighted algorithm in which the classifiers are built via weighted observations performs better than weighted resampling at each iteration, apparently due to removing the randomisation (Friedman, Hastie & Tibshirani, 1998).
- Confidence rated predictions outperform boosting algorithms where a 0/1 loss is applied to incorrect classification (Freund & Schapire, 1996; Schapire & Singer, 1998).

- Successful ensemble classification is due to the non-overlap of errors (Dietterich, 1997) i.e. observations which are classified correctly by one hypothesis are classified incorrectly by others and vice versa.
- Margin and edge analysis are recent explanations for Breiman, 1997 and Schapire et al., 1998. More detail on these measures and related studies is provided in the next section.

3. Edge and Margin Analysis

Recent explanations as to the success of boosting algorithms have their foundations in margin and edge analysis. These two measures are defined for the ith training observation at trial m as follows: Assume we have a base learner which produces hypothesis $h_m(\mathbf{x})$ at the m-th iteration, and an error indicator function, $I_m(\mathbf{x}_i) = I(h_m(\mathbf{x}_i) \neq y_i)$. Let c_m represent the vote for the m-th hypothesis with $\sum_m c_m = 1$. Then,

- $edge_i(m, \mathbf{c})$ = total weight assigned to all incorrect classes. The edge is defined formally in (Breiman, 1997) as

$$edge_i(m, \mathbf{c}) = \sum_{j=1}^{m} c_j I_j(\mathbf{x}_i) \tag{1}$$

- $margin_i(m, \mathbf{c})$ = total weight assigned to the correct class minus the maximal weight assigned to any incorrect class.

For the two class case $margin_i(m, \mathbf{c}) = 1 - 2\ edge_i(m, \mathbf{c})$ and in general, $margin_i(m, \mathbf{c}) \geq 1 - 2\ edge_i(m, \mathbf{c})$ (Schapire et al. (1998)).

Whilst more difficult to compute, the value of the margin is relatively simple to interpret. Margin values will always fall in the range $[-1, 1]$, with high positive margins indicating confidence of correct classification. An example is classified incorrectly if it has a negative margin. The edge on the other hand cannot be used as an indicator variable for correct classification (except in the two-class case). Whilst the margin is a useful measure due to its interpretability, mathematically it is perhaps not as robust and tractable as the edge.

Schapire et al. (1998) claim that boosting is successful because it works to increase low margins for difficult observations, hence increasing the confidence of correct classification. Similarly, Breiman (1997) claims that a lower average edge (or higher average margin) should lead to higher classification accuracy on unseen cases.

This study examines the variance and average of the edge values versus the number of boosting trials performed. Methodology for the study is discussed in detail in the next section.

4. Empirical Results and Initial Theory

In all experiments, the decision tree learner C4.5 (Quinlan, 1993) with default values and pruning was used as the base classifier, with a boosted ensemble being built from $M = 50$ iterations. Datasets used are a selection from the UCI[1] Machine Learning Repository. The datasets were chosen to provide a representative mixture of dataset size and boosting performance previously reported (Quinlan, 1996), Schapire et al., 1998). 10-fold crossvalidation was applied whereby the original training data were shuffled randomly and split into 10 equal-sized partitions. Each of the ten partitions was used in turn as a test set for the ensemble generated using the remaining 90% of the original data as a training set.

At each iteration, the values for edge were calculated for each observation in the 10 training sets (cross-validation folds). The average and variance of $edge_i(m, \mathbf{c})$ were calculated as follows:

$$\hat{E}[edge_i(m, \mathbf{c})] = \tfrac{1}{n} \sum_{i=1}^{n} edge_i(m, \mathbf{c})$$
$$\hat{Var}[edge_i(m, \mathbf{c})] = \tfrac{1}{n} \sum_{i=1}^{n} (edge_i(m, \mathbf{c}) - \hat{E}[edge_i(m, \mathbf{c})])^2$$

The results of these trials for the colic, glass and letter datasets appear in graphical form below, - (these results are indicative of the average and variance trends for all datasets tested).

oNte an apparent exponential decrease in variance perhaps indicating an asymptote of zero or some small value, ϵ. These results prompt the question *"does boosting homogenise the edge?"*. The most dramatic variance decay is seen in boosting trials $m \leq 5$ i.e. most of the 'hard' work appears to be done in the first few trials. This observation is consistent with several authors noting in earlier published empirical studies that little additional benefit is gained after 10 boosting trials when a relatively strong learner is used as the base learner. If the above variance reduction trends are truly exponential, replotting on a log scale will show a linear trend. This is not the case, however, and we may conclude that the decrease is not purely exponential.

It appears that observations with low inital edges are 'sacrificed' for observations with high initial edges. i.e observations which were initially classified correctly are classified incorrectly in later rounds in order to classify 'harder' observations correctly. This notion is consistent with margin

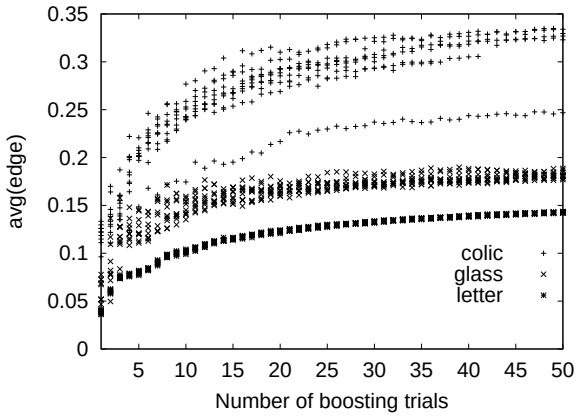

Figure 1. Average edge versus number of boostingtrials.

distribution results presented by Schapire et al. (1998) and with results yet to be published by Breiman (personal communication with Leo Breiman).

It has been suggested in recent studies that reduction in test set error may correlate with a reduction in the edge values. With the exception of the letter dataset, plotting the average edge versus test set error for all crossvalidated folds showed no relationship with reduction in test set error. However, the glass and colic datasets are quite small, resulting in crossvalidated test sets containing 10-20 observations only. It was also noted that boosting with unweighted votes where the vote for each classifier was equal to $\frac{1}{m}$ resulted in a similar 'exponential' decrease, perhaps alluding to a voting effect rather than the specific form of the algorithm.

Interestingly, an increase in the average edge is apparent as the number of boosting trials increases. Refer to Figure 4 below for results on the colic, glass and letter datasets. Again these trends are indicative of trends for all other datsets tested.

5. Developing Expressions for $\mathbf{Var[edge_i(m,c)]}$ and $\mathbf{E[edge_i(m,c)]}$

An alternative expression for the variance of the edge is derived below. Firstly, the definition of edge follows that given in equation (1) (Breiman, 1997). Letting $c_j = a_j / \sum a_j$, and defining the *unweighted* error of the j-th

hypothesis as e_j, this expression may be rewritten as :

$$edge_i(m, \mathbf{c}) = \frac{1}{\sum a_j} \sum_{j=1}^{m} a_j I_j(\mathbf{x}_i).$$

Now for **Adaboost**,

- $a_j = \log(\frac{\epsilon_j}{1-\epsilon_j}) = \log(\frac{1}{\beta_j})$
- $Var[I_j(\mathbf{x}_i)] = e_j(1 - e_j)$

Therefore,

$$Var[edge_i(m, \mathbf{c})] = \frac{\sum_{j=1}^{m} (\log \beta_j)^2 e_j(1 - e_j)}{(\sum_{j=1}^{m} \log \beta_j)^2}$$

$$+ \frac{2 \sum \sum_{j<k}^{m} \log \beta_j \log \beta_k Cov[I_j(\mathbf{x}_i), I_k(\mathbf{x}_i)]}{(\sum_{j=1}^{m} \log \beta_j)^2} \tag{2}$$

The expression derived for $Var[edge_i(m, \mathbf{c})]$ above is dependent only on ϵ_j, e_j and $Cov[I_j(\mathbf{x}_i), I_k(\mathbf{x}_i)]$. Intuitively, this result makes sense, since, at each iteration, the learner attempts to correctly predict observations that were predicted incorrectly at the previous iteration. For this to happen, the indicator variables for unweighted error should be negatively correlated or uncorrelated for pairwise iterations. If errors were positively correlated, voting could degrade performance since individual hypotheses may consistently vote incorrectly on some observations and never be given the chance to explore different areas of the training set. Hence, from the expression derived, negative or zero covariance terms will result in non-increasing values for $Var[edge_i(m, \mathbf{c})]$. Now, $Cov[I_j(\mathbf{x}_i), I_k(\mathbf{x}_i)] = E[I_j(\mathbf{x}_i), I_k(\mathbf{x}_i)] - e_j e_k$ and a loose lower bound for $Cov[I_j(\mathbf{x}_i), I_k(\mathbf{x}_i)]$ is given by $min(e_j, e_k) - e_j e_k$.

A simplified expression for the average edge is given by:

$$E[edge_i(m, \mathbf{c})] = \frac{\sum_{j=1}^{m} \log(\frac{1}{\beta_j}) e_j}{\sum_{j=1}^{m} \log(\frac{1}{\beta_j})} \tag{3}$$

After algebraic manipulation it can be shown that the condition for average edge to increase between the mth and $m+1$th trials is $e_{m+1} \geq E[edge_i(m, \mathbf{c})]$. That is, the average edge will increase between the mth and $m+1$th trials if the unweighted error of the $m+1$th trial is greater than or equal to the average edge on the mth trial. Generally the average edge increases but stays below a threshold of the maximum unweighted error of the hypotheses. The maximum unweighted error is an upper bound on the average edge. This may imply the following:

$e_{m+1} \uparrow \Rightarrow E[edge_i(m, \mathbf{c})] \uparrow \Rightarrow Var[edge_i(m, \mathbf{c})] \downarrow \Rightarrow TestError \downarrow$, perhaps indicating that further improvements in test set error may be obtained by actively minimising the variance of the edge (or margin) values.

6. Is $Var[edge_i(m, \mathbf{c})]$ a Monotonic Non-Increasing Function?

A theoretical expression for $Var[edge_i(m, \mathbf{c})]$ has been given in Section 4.1 in (2). To prove that this is monotonic non-increasing function, it must be shown that:

$$Var[edge_i(m, \mathbf{c})] \geq Var[edge_i(m+1, \mathbf{c})]$$

Algebraically, it is quite straightforward to show that:

$$Var[edge_i(m+1, \mathbf{c})] = C_m Var[edge_i(m, \mathbf{c})] + A_m$$

Now, $C_m < 1 \ \forall m$ and hence A_m must be non-negative in the limit as the LHS variance expression would fall below zero. But $A_m = A_{m,1} + A_{m,2}$ with $A_{m,1}$ being negative in the limit and $A_{m,2}$ being strictly positive. If A_m is non-negative in the limit, $\frac{A_{m,1}}{A_{m,2}}$ must be less than -1 in the limit. Refer to Figure 7 below where it appears that this is the case, implying that A_m has a non-negative limit.

Alternatively, define δ_m as follows:

$$\delta_m = Var[edge_i(m+1, \mathbf{c})] - Var[edge_i(m, \mathbf{c})]$$

If $Var[edge_i(m, \mathbf{c})]$ is monotonic non-increasing, $\delta_m \leq 0$. The values of δ_m are plotted below.

Again, via algebraic manipulation it can be shown that:

$$\delta_m = \delta_{m,1} + \delta_{m,2} + A_m$$

Now, $\delta_{m,1} < 0$ and $\delta_{m,2} < \|\delta_{m,1}\|$. It can be shown via mathematical induction that $\delta_{m,2} < 0$. Hence in both cases above we seek the distribution of A_m to prove the non-increasing property of $Var[edge_i(m, \mathbf{c})]$. Since A_m is essentially a correlation term with a slightly positive value in the limit, this confirms the notion that the success of boosting is due to the lack of correlation between errors made by individual classifiers (or a non-overlap of errors between individual classifiers).

7. Empirical Trials of New Terms

Using the glass, colic and letter datasets, the values of A_m, $A_{m,1}$, $A_{m,2}$, C_m, δ_m, $\delta_{m,1}$ and $\delta_{m,2}$ were evaluated. As with previous empirical results, 10-fold crossvalidation was applied to the same shuffled datasets with each plotted point representing the result from one fold. M=50 boosting trials were employed.

It can be seen in Figure that A_m appears to be zero or slightly positive in the limit with a tighter scatter about the zero line as m increases. Figure shows $A_{m,1}$ being negative in the limit but not exhibiting the same scatter decrease as A_m. From Figure it is clear that $A_{m,2}$ is strictly positive.

It can be seen in Figure that C_m is always negative with an apparent limit of 1. It can be seen in Figures and that $\delta_{m,1}$, $\delta_{m,2}$ are always negative, both with an apparent limit of 0.

It can be noted in all the above figures, the variation in each term for the colic dataset is larger than the variation observed for the other 2 datasets plotted (i.e. glass, letter)

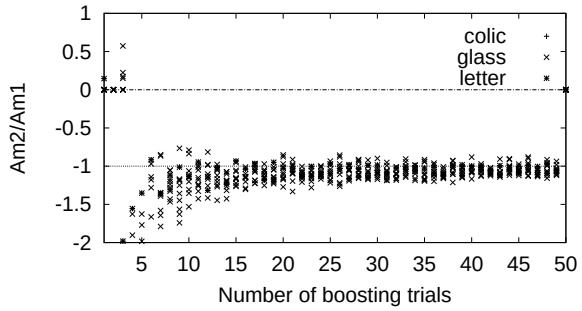

Figure 2. $\frac{A_{m,2}}{A_{m,1}}$.

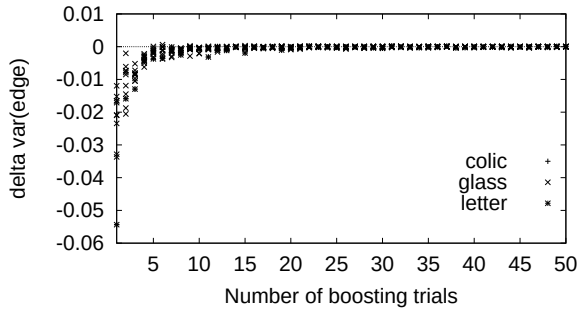

Figure 3. δ_m versus number of boosting trials.

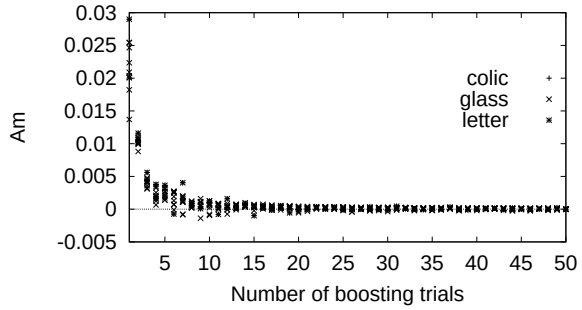

Figure 4. A_m versus number of boosting trials.

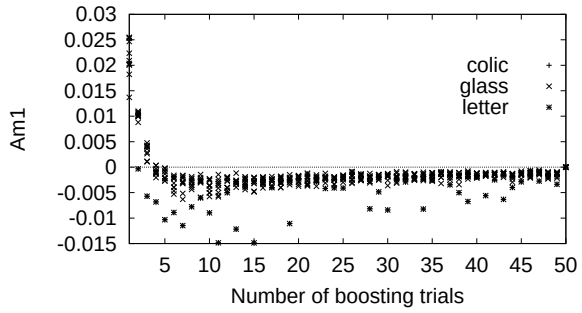

Figure 5. $A_{m,1}$ versus number of boosting trials.

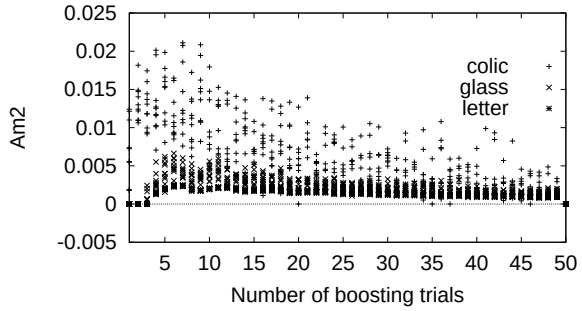

Figure 6. $A_{m,2}$ versus number of boosting trials.

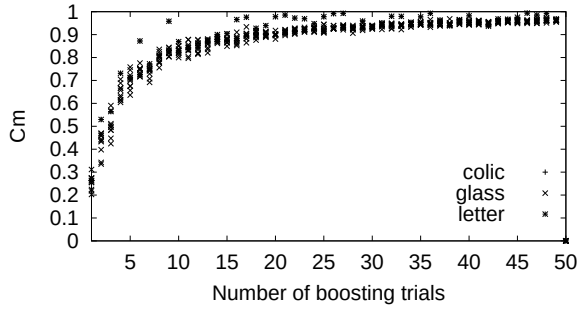

Figure 7. C_m versus number of boosting trials.

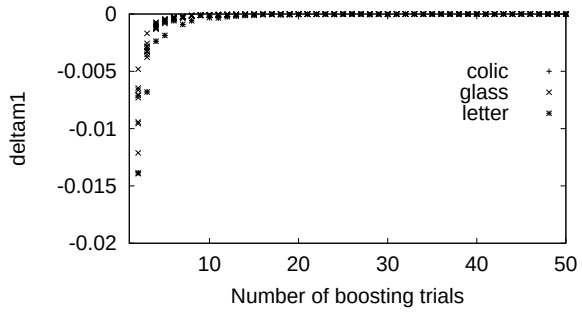

Figure 8. $\delta_{m,1}$ versus number of boosting trials.

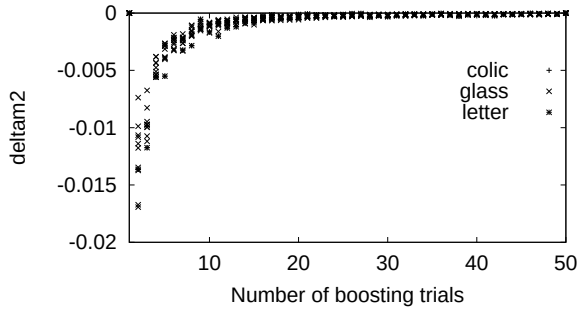

Figure 9. $\delta_{m,2}$ versus number of boosting trials.

8. Interesting Empirical Observations on $E[I_j(\mathbf{x}_i), I_k(\mathbf{x}_i)]$ and $\log \beta_j$

For the glass,bands, colic and letter datasets, the values of $\log \beta_j$ were calculated for all j and $E[I_j(\mathbf{x}_i), I_k(\mathbf{x}_i)]$ for $j, k \leq 5, j \neq k$.

- for all but the colic dataset, $E[I_j(\mathbf{x}_i), I_k(\mathbf{x}_i)]$=0 for all $j, k \leq 3$.
- the colic dataset had $E[I_j(\mathbf{x}_i), I_k(\mathbf{x}_i)]$¿0 for all $j, k \leq 5$. i.e. there are observations which are predicted incorrectly in one round and also predicted incorrectly in subsequent rounds. Could this be a factor in the degradation of performance of a learner on colic data when boosting is applied?
- it appears that $E[I(\mathbf{x}_j), I(\mathbf{x}_j + 1)] = 0$.
- It has been noted earlier that $E[I_j(\mathbf{x}_i), I_k(\mathbf{x}_i)] \leq min(e_i, e_j)$. This upper bound is now seen to be very loose when the exact values of $E[I_j(\mathbf{x}_i), I_k(\mathbf{x}_i)]$ are calculated empirically. i.e. $E[I_j(\mathbf{x}_i), I_k(\mathbf{x}_i)] << min(e_i, e_j)$ in this study.
- $\log \beta_m$ shows no trending as m increases for glass and colic but shows variance reduction and possible cycles for letter. Additionally, the values of β_j are highly variable for the colic datatset. This is an interesting result as boosting degrades performance on the colic dataset.
- $\sum_{j=1}^{m} \log \beta_j$ is linear in m, suggesting that $\log \beta_j$ is constant.
- Since $\log \beta_j$ appears to be constant, $(\sum \log \beta_j)^{-2}$ is strongly a x^{-2} type curve and this normalising factor has a strong decaying effect.

9. General Forms of Voting Systems

Mathematical analysis of variance reduction may be simplified by considering general forms of voting systems. This may also allow us to partition the variance into components pertaining to the voting mechanism and those pertaining to the method of formation of a sequence of classifiers. In boosting, consecutive classifiers are formed via an adaptive procedure but for bagging they are formed via a sequence of bootstrap replicates. Examples of possible schemes to consider are :

- all m classifiers make identical predictions at each iteration and hence have the same individual error rate with $corr[I_j(\mathbf{x}_i), I_k(\mathbf{x}_i)]$= 1. Voting weight of the mth classifier = $\frac{1}{m}$; in this case $Var[edge_i(m, \mathbf{c})]$= $e(1-e)$, which is constant and independent of m. Therefore, if this type of voting scheme was employed, no reduction in the variance of the edge would occur.
- the m classifiers do not make identical predictions at each iteration but have the same individual error rates with $corr[I_j(\mathbf{x}_i), I_k(\mathbf{x}_i)] = \rho$. ($\frac{-1}{m-1} \leq \rho <$ 1). Voting weight of the mth classifier = $\frac{1}{m}$; in this case, $Var[edge_i(m, \mathbf{c})]$ = $\frac{e(1-e)}{m}\{1 + \rho(m-1)\}$.

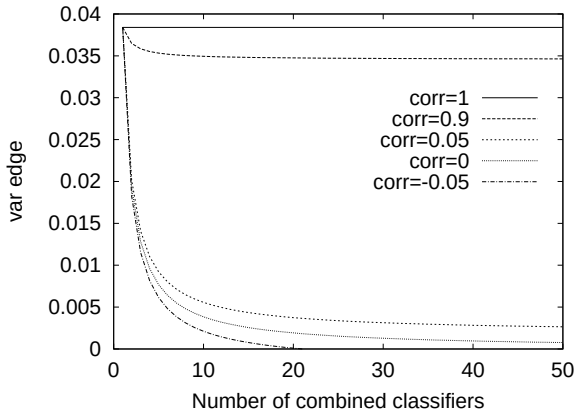

Figure 10. Plot of calculated $Var[edge_i(m, \mathbf{c})]$ for varying ρ versus number of combined classifiers ($e = 0.04$).

The first term in the variance expression above may be considered to be the voting component and the second term involving ρ the component pertaining to the method of classifier formation. Figure 9 below shows the value of this variance with e fixed at 0.04 and ρ varying; $\frac{-1}{1-m} \leq \rho \leq 1$. The reason ρ has a lower limit of $\frac{-1}{m-1}$ is given by Kendall & Stuart (1963). It may be possible to apply this lower limit on ρ in future work when trying to prove asymptotic limits on $Var[edge_i(m, \mathbf{c})]$.

To check the degree of correlation between individual hypotheses for the letter dataset, the variance values obtained empirically are overlaid onto the variance trend graph above. The value of e is again 0.04, which is a close match to the values of e_j obtained empirically for the letter data. Refer to Figure 5 below where we may conclude that $corr[I_j(\mathbf{x}_i), I_k(\mathbf{x}_i)]$ for the letter data is in the range $0 \leq corr[I_j(\mathbf{x}_i), I_k(\mathbf{x}_i)] \leq 0.10$. Clearly the assumptions of equal individual error rates and equal correlation between classifiers is too loose but we can still gain an appreciation for the type of variance reduction occuring and the degree of correlation between hypotheses.

10. Conclusion

This study has presented some interesting results on the variance of the edge when a boosted ensemble is formed. Variance reduction trends are consistent across all datasets tested. The initial theory and associated empirical results presented confirm that a key factor in this reduction is lack of correlation between errors of individual classifiers. Some initial theoretical work presented in Section

5 on generalised forms of voting systems also suggests low correlation of errors between individual hypotheses leads to reduction in the variance of the edge (and margin) values. Analysing the edge and margin distributions as a whole rather than sample statistics such as average and variance may lead to further insights into the boosting mechanism with possibilities for improved ensemble classifiers.

Notes

1. http://www.ics.uci.edu/˜mlearn/MLRepository.html

References

1. Breiman, L. (1996a). Bagging predictors. *Machine Learning* 26(2), 123-140.
2. Breiman, L. (1996b). *Bias, Variance and Arcing Classifiers* (Technical Report 460). Statistics Department, University of California, Berkeley.
3. Breiman, L. (1997). *Arcing the edge* (Technical Report 486). Statistics Department, University of California, Berkeley.
4. Dietterich, T.G. (1997). Machine learning research: Four current directions. *AI Magazine* 18(4), 99-137.
5. Freund, Y., & Schapire, R.E. (1996). Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning*, 148-156. Morgan Kaufmann.
6. Freund, Y., & Schapire, R.E. (1997). A decision-theoretic generalisation to on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55(1), 119-139.
7. Friedman, J.H.(1997). On bias, variance, 0/1-loss and the curse of dimensionality. *Data Mining and Knowledge Discovery* 1(1), 55-77.
8. Friedman, J.H., Hastie, T. & Tibshirani, R.(1998). *Additive logisitic regression: a statistical perspective on boosting.* (Technical Report 199). Department of Statistics, Stanford University.
9. Kendall, M.G. & Stuart, A. (1963). *The Advanced Theory of Statistics.* Oxford University Press, Oxford, UK.
10. Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann.
11. Quinlan, J.R.(1996). Bagging, boosting and C4.5. *Proceedings of the Thirteenth National Conference on Articifical Intelligence.* (pp. 725-730). Menlo Park California, American Association for Artificial Intelligence.
12. Schapire, R.E., Freund, Y., Bartlett, P. & Lee, W.S. (1998). Boosting the margin: a new explanation for the effectiveness of voting methods. *Annals of Statistics* 26(5), 1651-1686.
13. Schapire, R.E. & Singer, Y. (1998). Improved boosting algorithms using confidence rated predictions. *Proceedings of the Eleventh Computational Learning Theory*, 80-91.