Pre-publication draft of a paper which appeared in the Proceedings of the Fourth Australian Knowledge Acquisition Workshop (AKAW '99), pages 117-128

On the effect of data set size on bias and variance in classification learning

Damien Brain
Geoffrey I Webb
School of Computing and Mathematics
Deakin University
Geelong Vic 3217

Abstract

With the advent of data mining, machine learning has come of age and is now a critical technology in many businesses. However, machine learning evolved in a different research context to that in which it now finds itself employed. A particularly important problem in the data mining world is working effectively with large data sets. However, most machine learning research has been conducted in the context of learning from very small data sets. To date most approaches to scaling up machine learning to large data sets have attempted to modify existing algorithms to deal with large data sets in a more computationally efficient and effective manner. But is this necessarily the best method? This paper explores the possibility of designing algorithms specifically for large data sets. Specifically, the paper looks at how increasing data set size affects bias and variance error decompositions for classification algorithms. Preliminary results of experiments to determine these effects are presented, showing that, as hypothesised variance can be expected to decrease as training set size increases. No clear effect of training set size on bias was observed. These results have profound implications for data mining from large data sets, indicating that developing effective learning algorithms for large data sets is not simply a matter of finding computationally efficient variants of existing learning algorithms.

Introduction

The amount of data being stored by organisations is increasing at a rapid rate, and this trend is likely to continue for the foreseeable future. Therefore, as time passes, we can expect machine learning algorithms to be required to be used on increasingly large data sets - much larger than the size of data sets with which they were originally developed. Hence, machine learning algorithms will be required to perform well on very large data sets. This paper addresses the impact of this trend on classification learning algorithms.

Classification learning algorithms aim to learn a model that maps a multivalued input X into a single valued categorical output Y. Thus, classification algorithms can be used to predict the output Y of an unseen X.

Although much work has been done on evaluating the performance of classification algorithms, these evaluations have generally used relatively small data sets. Therefore, there is little evidence to support the notion that "standard" versions of common classification algorithms perform well on very large data sets. In fact, there is a large body of literature on attempts to "scale up" algorithms to handle large data sets [1, 2, 3]. This body of work primarily addresses the issue of how to reduce the high computational costs of traditional learning algorithms so as to make tractable their application to large data sets.

However, this begs the question of whether machine learning algorithms developed for small data sets are inherently suitable for large data sets. Is it really just a question of making existing algorithms more efficient, or are the demands of effective learning from large data sets fundamentally different from those of effective learning from small data sets?

This paper argues for the later position. We argue that whereas the major problem confronting classification learning from small data sets is the management of error resulting from learning variance, as data set sizes increase the impact of variance can be expected to decrease. Hence, fundamentally different types of learning algorithm are appropriate for effective learning from small and large data sets.

The paper is organised as follows. First we describe the concepts of learning bias and learning variance. Next we outline the reasoning that led us to hypothesise that variance can be expected to decrease as data set sizes increase. Then we present some preliminary experimental results that lend support to this hypothesis. Finally we present our conclusions and outline our proposed directions for future research.

Learning Bias and Learning Variance

A number of recent studies have shown that the decomposition of a learner's error into bias and variance terms can provide considerable insight into the prediction performance of the learner. This decomposition originates from analyses of regression, the learning of models with numeric outputs [4]. Squared *bias* measures the contribution to error of the central tendency of the learner when trained on different data. *Variance* is a measure of the contribution to error of deviations from the central tendency. Bias and variance are evaluated with respect to a distribution of training sets, such as a distribution containing all possible training sets of a specified size for a specified domain.

Analysing learning systems in this way highlights the following issues. If a learning system learns different classifiers from different training sets, then the degree to which the predictions of those classifiers differ provides a lower limit on the average error of those classifiers when applied to subsequent test data. If the predictions from different classifiers differ then not all can be correct!

However, inhibiting such variations between the classifiers will not necessarily eliminate prediction error. The degree to which the correct answer for an object can differ from that for other objects with identical descriptions ("irreducible error") and the accuracy of the learning bias also affect prediction error. Errors will also be caused by predictions from different classifiers that are identical but incorrect!

Unfortunately, the definitions of bias and variance that have been developed for numeric regression do not directly apply to classification learning. In numeric regression a prediction is not just simply right or wrong, there are varying degrees of error. In consequence, a number of alternative formulations of bias and variance for classification learning have emerged [5, 6, 7, 8, 9]. Each of these definitions is able to offer valuable insight into different aspects of a learner's performance. For this research we use Kohavi and Wolpert's definition [6], as it is the most widely employed of those available. Following Kohavi and Wolpert, irreducible error is aggregated into *bias*^{2}. In conseqence, *bias*^{2}and *variance* sum to *error*.

Different learning algorithms may have quite different bias/variance profiles.

An example of a high bias classification algorithm is the Naïve-Bayes classifier [10, 11]. Naïve-Bayes classifiers fit a simple parametric model to the data. They are limited in the extent to which they can be adjusted to different data sets. Such adjustments are restricted to changes in a relatively small number of conditional probability estimates that underlie a Naïve-Bayes classifier. In consequence, the predictions of a Naïve-Bayes classifier will be little affected by small changes in the training data.

Low bias algorithms (such as boosting decision trees[12]), are more flexible. They can not only describe a wider range of concepts, but are usually more adaptive in dealing with the training set. Boosting repeatedly applies a base learning algorithm to a training set, resampling or reweighting the data each time so as to force the learning system toward correctly classifying all the training cases. This process has been shown to result in a general major decrease in learning bias, accompanied by a smaller decrease in learning variance when applied with a standard decision tree learner as the base learning algorithm [13].

Our hypothesis

Variance measures the degree to which the predictions of the classifiers developed by a learning algorithm differ from training sample to training sample. When sample sizes are small, the relative impact of sampling on the general composition of a sample can be expected to be large. For example, if 50% of a population exhibit some characteristic, in a sample of size 10 there is a 0.38 probability that only 40% or less of the sample will exhibit the characteristic and a 0.17 probability that only 30% or less of the sample will exhibit the characteristic. In other words, a small sample is likely to be quite unrepresentative of the population as a whole. In contrast, in a random sample of size 1,000,000, if 50% of the population exhibits the characteristic then the probability that 40% or less of the sample will exhibit the characteristic is less than 10^{-22}. The probability that 30% or less of the sample will exhibit the characteristic is less than 10^{-26}. If 1% of a population exhibits a characteristic, there is a probability of 0.37 that it will not be represented at all in a random sample of size 100. The probability that it will not be represented at all in a sample of size 1,000,000 is less than 10^{-17}. Clearly, the degree of significant differences in the composition of alternative small samples will be greater than that of significant differences in the composition of large samples.

In consequence, it is to be expected that classifiers learned from alternative small samples will differ more significantly than classifiers learned from alternative large examples. It follows that their predictions are likely to differ more and hence that their variance will be higher.

On the basis of this reasoning, we predict that there will be a tendency for standard learning algorithms to exhibit decreasing levels of variance as training set sizes increase.

Experiments

Experiments were performed to investigate whether such a trend towards lower variance as training set size increases exists in practice. Three different classification algorithms were used, each with different bias and variance profiles. The high bias / low variance Naïve-Bayes classifier [10, 11], the machine learning exemplar C4.5 [14] and the bias and (to a lesser extent) variance reducing MultiBoost [9] were chosen. The latter combines the well-known AdaBoost [12] and Bagging [15] algorithms, coupling most of the superior bias reduction of the former with most of the superior variance reduction of the latter. Training set sizes started from 125 cases and increased up to

32,000, doubling at each step. The four data sets used (adult, shuttle, connect-4, and cover type) were the only real-world data sets available from the UCI Machine Learning Repository [16] that were suitable for classification learning and contained at least 32,000 cases.

For each data set, training set size, and classification algorithm, a bias-variance analysis was performed using 10 times 3-fold cross-validation. The Kohavi-Wolpert bias-variance decomposition was measured, along with the average number of nodes for each induced classifier (Note. Naïve-Bayes algorithms do not have a variable number of nodes, hence this measure is not presented for the Naïve-Bayes algorithm). While our hypothesis relates only to variance, we believed that it was also interesting to examine bias, to see whether it was subject to any clear trends with respect to increasing data set size.

Multiboost

Graphs of bias/variance decompositions of error for experiments using MultiBoost are presented in Figures 1 – 4. In these and all subsequent bias/variance decomposition graphs, bias is represented as the upper part of each vertical shaded bar, and variance the remainder. The total height of each bar represents the total average error for that sample size. The figures show a clear trend toward lower total error as sample size increases. The figures also show a clear trend toward lower variance with increasing sample size. The only exception to this is a slight increase in variance when moving from a sample size of 250 to 500 on the Connect-4 data set. Bias also has a clear downward trend in the first three figures, however this is not apparent for the Adult data set.

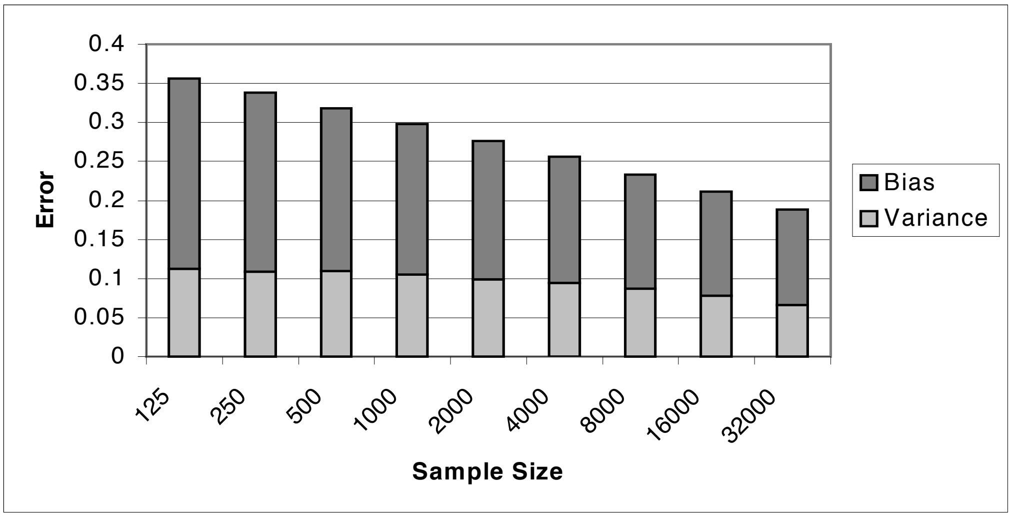

Figure 1. Bias and variance of Multiboost on the Connect-4 data set.

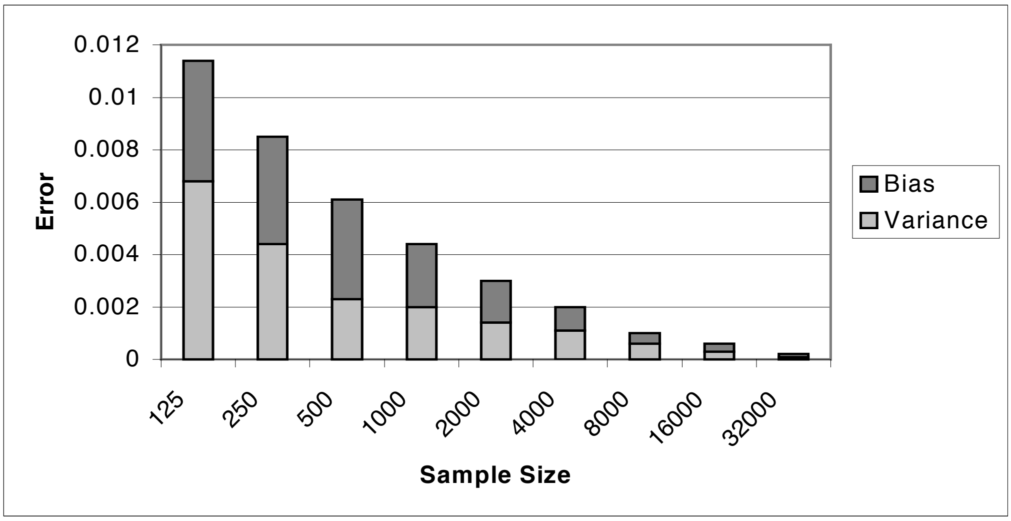

Figure 2. Bias and variance of Multiboost on the Shuttle data set.

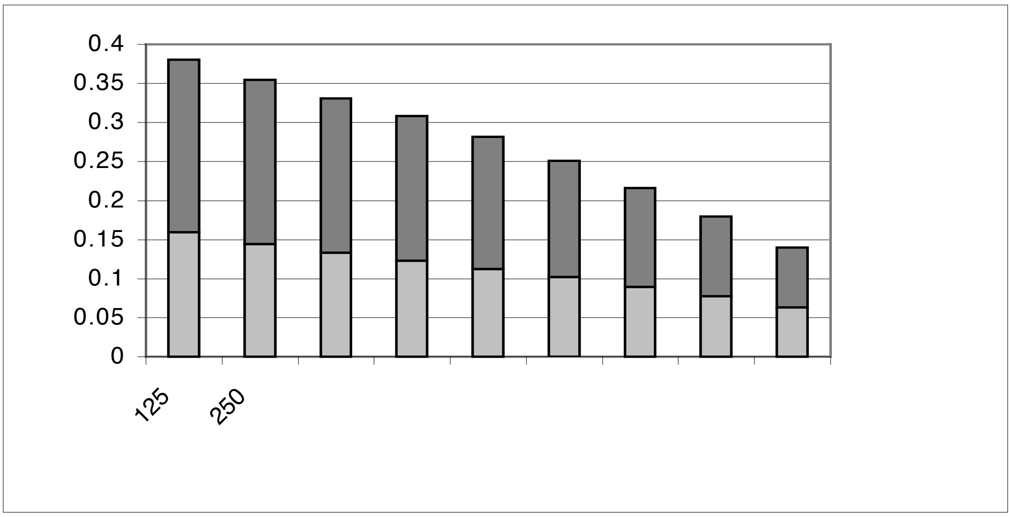

Figure 3. Bias and variance of Multiboost on the Cover Type data set.

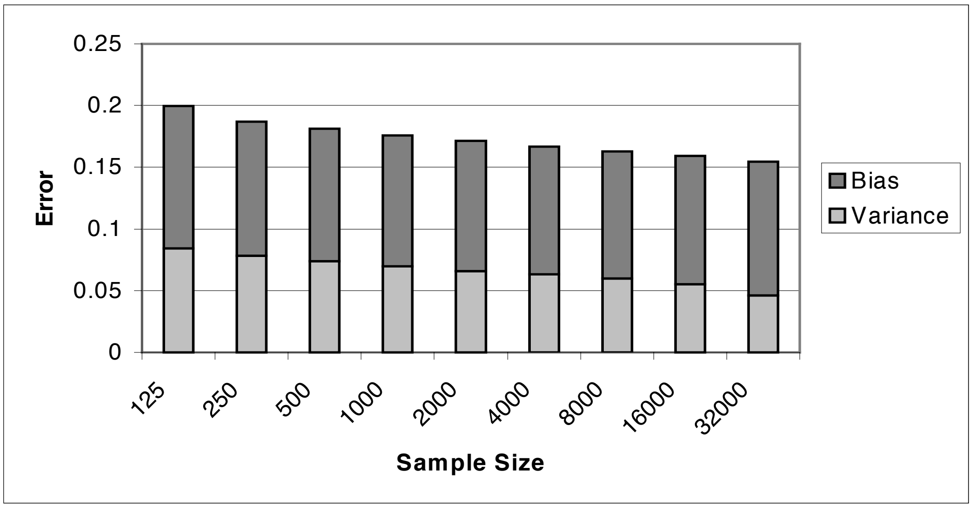

Figure 4. Bias and variance of Multiboost on the Adult data set.

C4.5

The results of C4.5 experiments are contained in Figures 5 – 8. It is clear in Figures 6 and 7 that the trend of lower bias, variance, and total error as sample size increases continues. However, in Figure 5, there is an increase in variance from sample sizes of 125 to 250, and 500 to 1,000. Bias decreases in every case, though. Results for the Adult data set (Figure 8) show many increases in bias – 250 to 500, 1,000 to 2,000, 4,000 to 8,000, and 16,000 to 32,000. There is also a very slight increase in variance when moving from 2,000 to 4,000 cases.

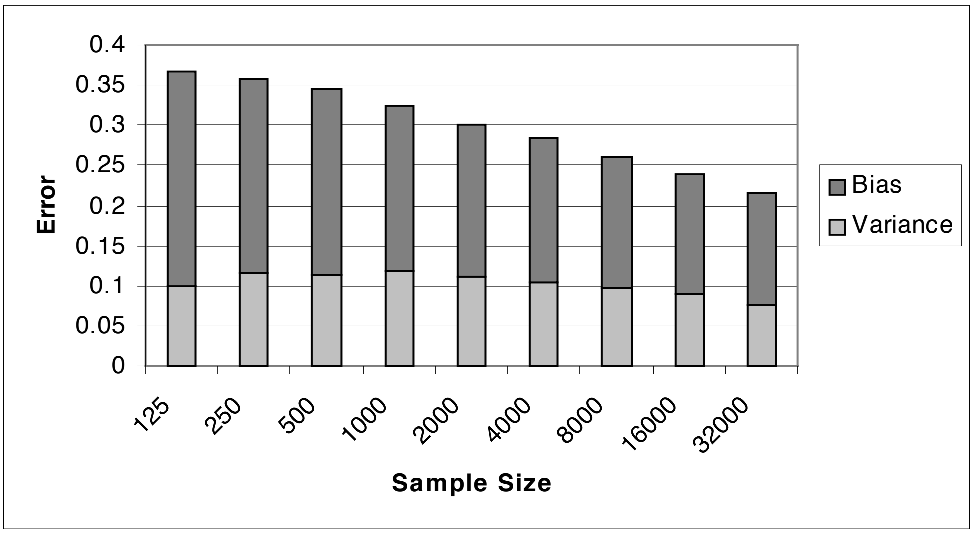

Figure 5. Bias and variance of C4.5 on the Connect-4 data set.

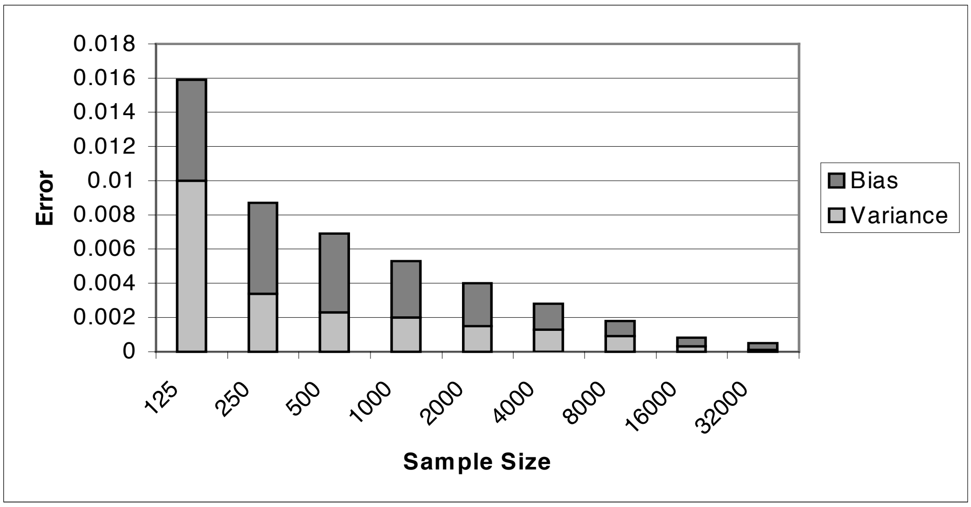

Figure 6. Bias and variance of C4.5 on the Shuttle data set.

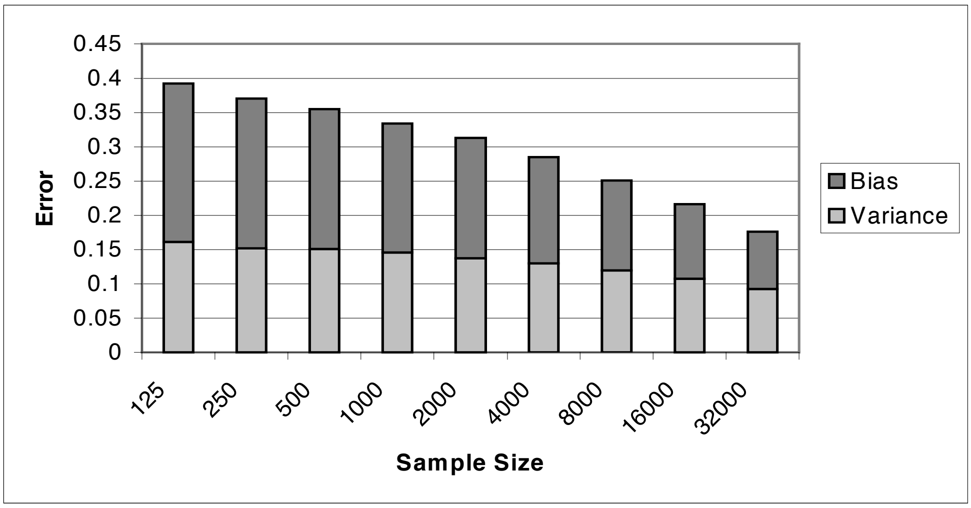

Figure 7. Bias and variance of C4.5 on the Cover Type data set.

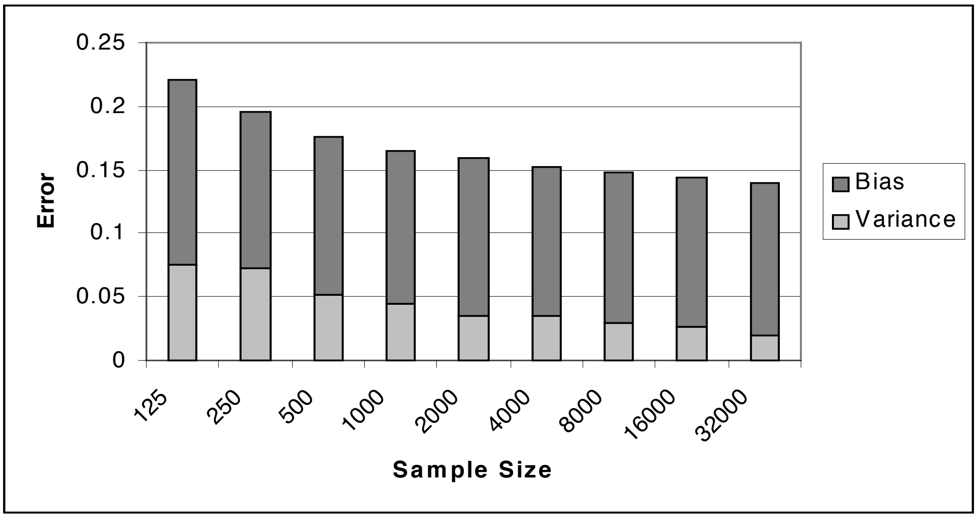

Figure 8. Bias and variance of C4.5 on the Adult data set.

Naïve-Bayes

Figures 9 – 12 show the results of experiments using a Naïve-Bayes classifier. In every case variance decreases as sample size increases. However, this is not true for bias in many situations. In fact, for the Adult data set (Figure 12), bias *increases* as sample size increases. This is also true for Figures 9 and 11, except for increasing sample size from 125 to 250 cases, where bias slightly decreases. The trend for bias on the Shuttle data set is more like those for other classifiers. Apart from a small increase from 8,000 to 16,000 cases, bias decreases as sample size increases. It is interesting to note that the only situation in all of the experiments in which increasing sample size corresponds to an increase in total error is in Figure 12, moving from 8,000 to 16,000 cases. This is undoubtedly due to the increased bias of the Naïve-Bayes classifier.

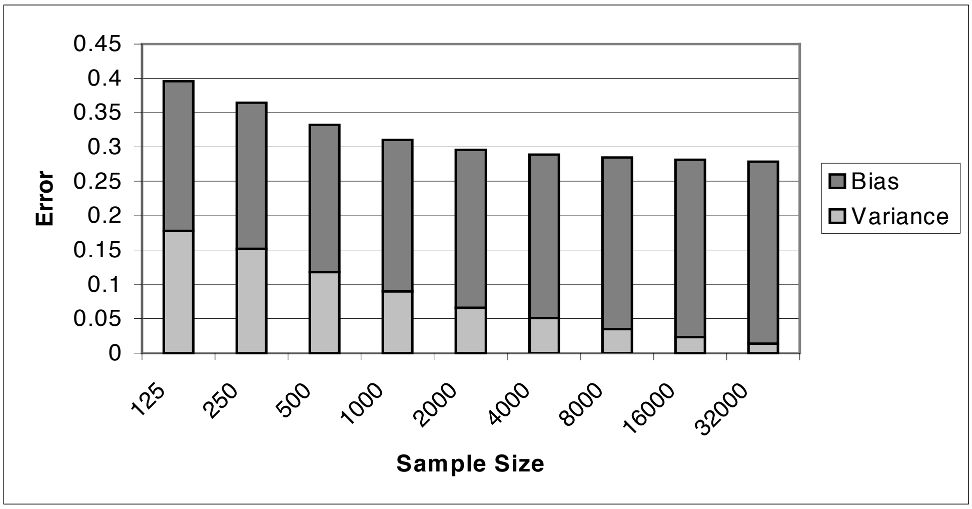

Figure 9. Bias and variance of NB on the Connect-4 data set.

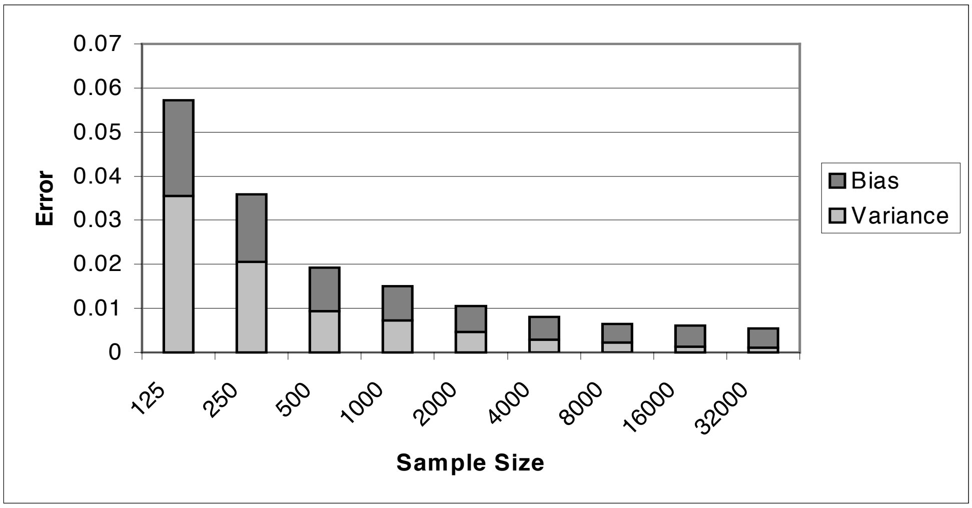

Figure 10. Bias and variance of NB on the Shuttle data set.

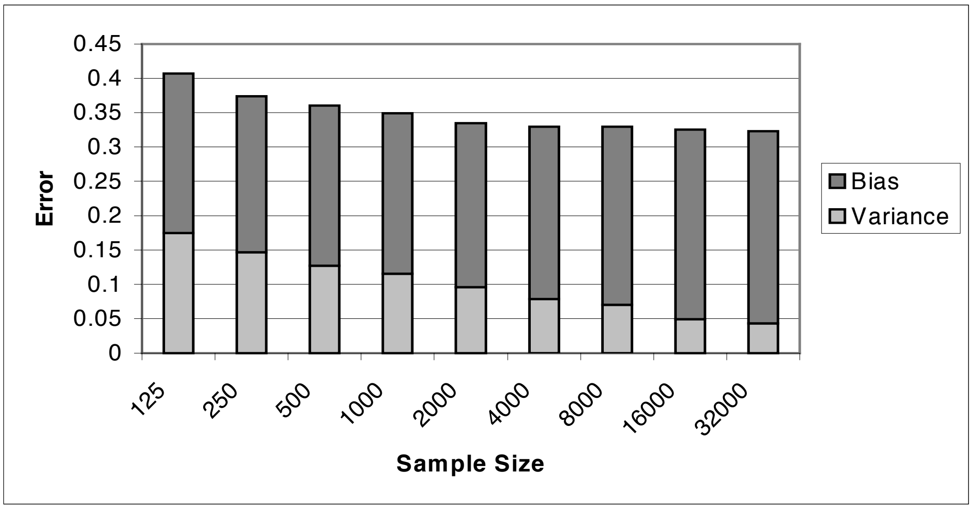

Figure 11. Bias and variance of NB on the Cover Type data set.

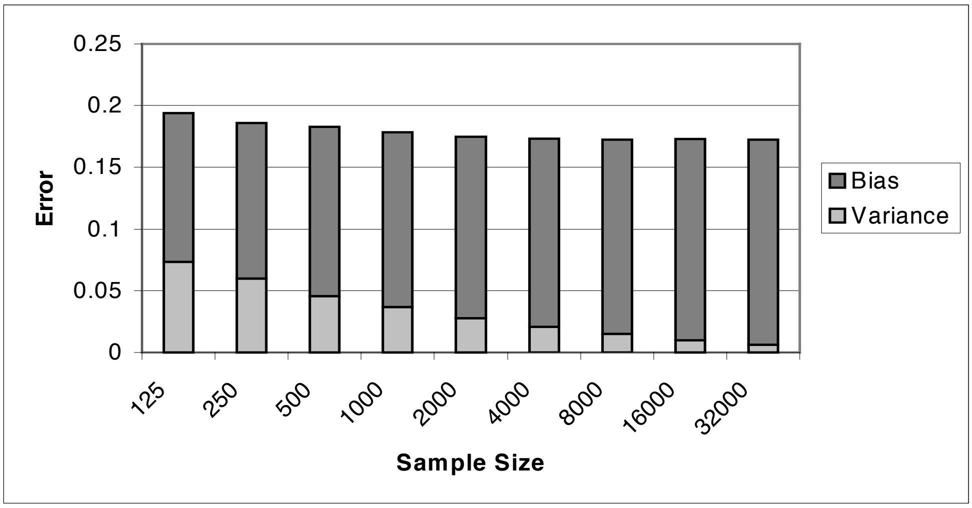

Figure 12. Bias and variance of NB on the Adult data set.

Statistical Significance

Table 1 shows the statistical significance of the above results. Significance was measured by applying a binomial probability test to a count of the number of increases and decreases in bias or variance when moving from one sample size to the next (i.e. sample sizes of 125 were not compared to 500 or more). Because prior predictions were made for variance, one-tailed tests are applied for the variance outcomes. As no predictions were made with respect to bias, two-tailed tests are for applied for the bias outcomes[1]. Results were considered significant if the outcome of the binomial test is less than 0.05.

As can be seen in Table 1, variance is shown to have a statistically significant reduction due to increased sample size in all but one instance (Connect-4 data using C4.5). This result is supportive of our hypothesis that variance will tend to decrease as training set sizes increase. These results suggest a particularly powerful effect given the very low power of the analyses performed, having only eight training set size steps available for each analysis.

Bias, on the other hand, has a significant decrease in four instances, and a significant *increase* in one instance. This suggests that bias is not influenced by training set size in the straightforward manner that variance appears to be.

	MultiBoost		C4.5		Naïve-Bayes	
	Bias	Variance	Bias	Variance	Bias	Variance
Adult	6:2 (0.2891)	**8:0 (0.0039)**	4:4 (0.6367)	**7:1 (0.0352)**	*0:8 (0.0078)*	**8:0 (0.0039)**
Connect-4	**8:0 (0.0078)**	**7:1 (0.0352)**	**8:0 (0.0078)**	6:2 (0.1445)	1:7 (0.0703)	**8:0 (0.0039)**
Cover Type	**8:0 (0.0078)**	**8:0 (0.0039)**	**8:0 (0.0078)**	**8:0 (0.0039)**	1:7 (0.0703)	**8:0 (0.0039)**
Shuttle	**8:0 (0.0078)**	**8:0 (0.0039)**	**8:0 (0.0078)**	**8:0 (0.0039)**	7:1 (0.0703)	**8:0 (0.0039)**

Table 1. Statistical significance of reductions in bias and variance due to an increase in sample size.
The ratio of decreases in bias or variance to increases is shown, followed by the outcome of a binomial probability test. Statistically significant results are shown in bold if positive, italics if negative.

Learned Theory Complexity

Another interesting aspect of the effect of larger training set sizes is the complexity of learned theories. It should be expected that as training sets increase there will be more information available to the classification algorithm, and this will allow a more complex theory to be produced. Figures 13 and 14 present the average numbers of nodes induced by MultiBoost and C4.5 respectively. The results given are the number of nodes induced for a training set size divided by the maximum number of nodes induced on all training sizes for a particular data set. As can be expected, the number of nodes increases dramatically as the training set moves from smaller to larger sizes.

[1] Note that using one-tailed tests for bias would convert the three non-significant results for Naïve Bayes to significant but have no impact on the significance or otherwise of the results for MultiBoost or C4.5. Using two-tailed tests for variance would convert the two 7:1 results to being insignificant, but have no other impact on the assessments of significance.

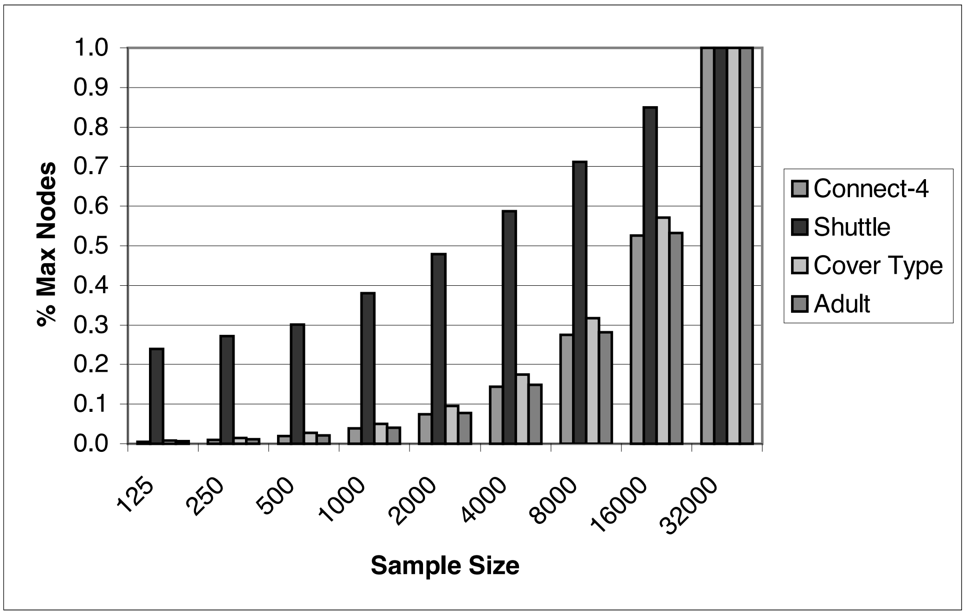

Figure 13. Comparison of the average number of nodes used by MultiBoost. The maximum number of nodes were: Connect-4: 64343.8; Shuttle: 340.5; Cover Type: 38261.7; Adult: 43375.4

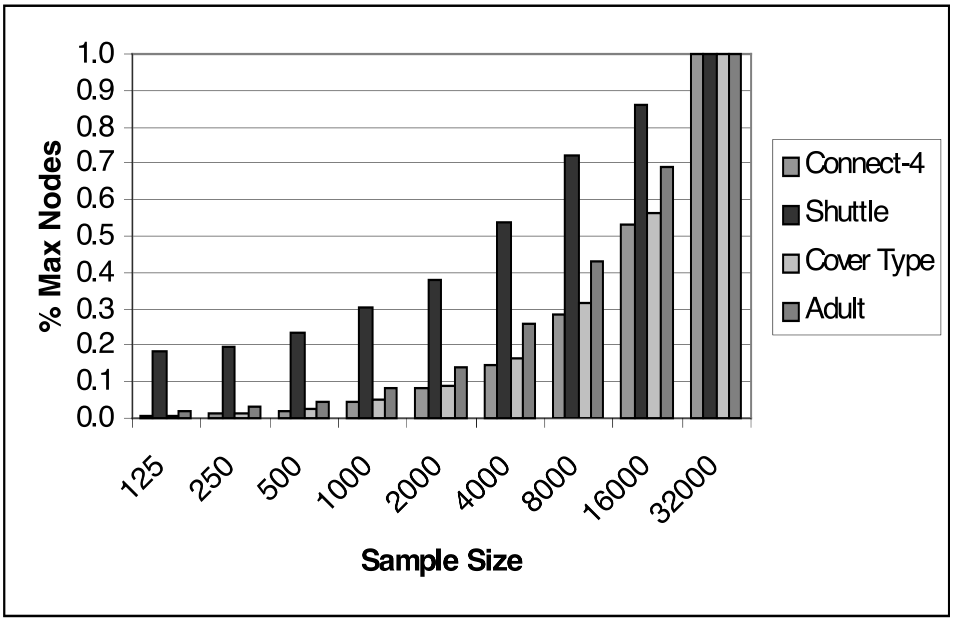

Figure 14. Comparison of the average number of nodes used by C4.5 The maximum number of nodes were: Connect-4: 3477.1; Shuttle: 45.3; Cover Type: 4066.4; Adult: 607.6

Conclusions and Future Work

These preliminary results show statistically significant evidence to support the hypothesis that variance can be expected to decrease as training set size increases. This may not be a surprising result. However, the fact that the results do not show a similar decrease in bias is not entirely expected. The algorithms used represented both high and low bias/variance algorithms, thus the results do not seem specific to a certain type of algorithm.

This is further supported by the huge increases in complexity of learned theories as training size increases. If the presented results are extrapolated to millions of training examples then the complexity of learned models can be expected to be orders of magnitude higher than that for the relatively small training sizes from which models are normally developed. However, it may be that this increase in complexity is exactly what causes the noted decreases in variance.

It is important to note the limitations of the current study. Only four data sets were employed and the largest data set sizes considered were very modest in data mining terms. Further experiments must be performed on more and larger data sets. If such experiments confirm the above results, then there are important implications. If the hypothesis is confirmed, some of the most well known and widely used classification algorithms may be shown to be less suitable for large data sets than for small. Our preliminary results suggest that while variance management is a critical property for good generalisation performance with small data sets, bias management is far more critical for good generalisation performance with large data sets.

References

1. Provost, F.J. and Aronis, J.M. (1996) "Scaling Up Inductive Learning with Massive Parallelism," *Machine Learning*, volume 23, number 1, pages 33-46. Kluwer Academic Publishers.
2. Provost, F.J. and Kolluri, V (1997) "Scaling Up Inductive Algorithms: An Overview," *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, Newport Beach, CA, pages 239-242. AAAI Press.
3. Catlett, J. (1992) "Peepholing: Choosing Attributes Efficiently for Megainduction," *Proceedings of the Ninth International Conference on Machine Learning*, Aberdeen, Scotland, pages 49-54, Morgan Kaufmann.
4. Geman, S. and Bienenstock, E. (1992) "Neural Networks and the Bias/Variance Dilemma," *Neural Computation*, volume 4, pages 1-48.
5. Breiman, L. (1996) "Bias, Variance, and Arcing Classifiers," *Technical Report 486*, Statistics Department, University of California, Berkeley, CA.
6. Kohavi, R. and Wolpert, D.H. (1996) "Bias Plus Variance Decomposition for Zero-One Loss Functions," *Proceedings of the 13th International Conference on Machine Learning*, Bari, Italy, pages 275-283. Morgan Kaufmann.
7. Kong, E.B. and Dietterich, T.G. (1995) "Error-correcting Output Coding Corrects Bias and Variance," *Proceedings of the 12th International Conference on Machine Learning*, Tahoe City, CA, pages 313-321. Morgan Kaufmann.
8. Friedman, J.H. (1997) "On Bias, Variance, 0/1-Loss, and the Curse-of-Dimensionality," *Data Mining and Knowledge Discovery*, volume 1, number 1, pages 55-77. Kluwer Academic Publishers.

9. Webb, G.I. (in press) "MultiBoosting: A Technique for Combining Boosting and Wagging," *Machine Learning*.

10. Kononenko, I. (1990) "Comparison of Inductive and Naïve Bayesian Learning Approaches to Automatic Knowledge Acquisition," In B. Wielinga et al. (eds.), *Current Trends in Knowledge Acquisition*. IOS Press.

11. Langley, P. and Iba, W.F. and Thompson, K. (1992) "An Analysis of Bayesian Classifiers," *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 223-228. AAAI Press.

12. Freund, Y. and Schapire, R.E. (1996) "Experiments with a New Boosting Algorithm," *Proceedings of the 13th International Conference on Machine Learning*, Bari, Italy, pages 148-156. Morgan Kaufmann.

13. Schapire, R.E and Freund, Y. and Bartlett, P. and Lee, W.S. (1998) "Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods," *The Annals of Statistics, 26, 1651-1686*.

14. Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

15. Breiman, L. (1996) "Bagging Predictors," *Machine Learning*, *24*, pages123-140.

16. Blake, C.L. and Merz, C.J. (1998) UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.