The Need for Low Bias Algorithms in Classification Learning from Large Data Sets

Damien Brain and Geoffrey I. Webb

School of Computing and Mathematics, Deakin University Geelong
Victoria, 3217, Australia
{dbrain,webb}@deakin.edu.au

Abstract. This paper reviews the appropriateness for application to large data sets of standard machine learning algorithms, which were mainly developed in the context of small data sets. Sampling and parallelisation have proved useful means for reducing computation time when learning from large data sets. However, such methods assume that algorithms that were designed for use with what are now considered small data sets are also fundamentally suitable for large data sets. It is plausible that optimal learning from large data sets requires a different type of algorithm to optimal learning from small data sets. This paper investigates one respect in which data set size may affect the requirements of a learning algorithm – the bias plus variance decomposition of classification error. Experiments show that learning from large data sets may be more effective when using an algorithm that places greater emphasis on bias management, rather than variance management.

1 Introduction

Most approaches to dealing with large data sets within a classification learning paradigm attempt to increase computational efficiency. Given the same amount of time, a more efficient algorithm can explore more of the hypothesis space than a less efficient algorithm. If the hypothesis space contains an optimal solution, a more efficient algorithm has a greater chance of finding that solution (assuming the hypothesis space cannot be exhaustively searched within a reasonable time). However, a more efficient algorithm results in more or faster search, not better search. If the learning biases of the algorithm are inappropriate, an increase in computational efficiency may not equate to an improvement in prediction performance.

A critical assumption underlies many current attempts to tackle large data sets by creating algorithms that are more efficient [1, 2, 3, 4, 5, 6]: that the learning biases of existing algorithms are suitable for use with large data sets. Increasing the efficiency of an algorithm assumes that the existing algorithm only requires more time, rather than a different method, to find an acceptable solution.

T. Elomaa et al. (Eds.): PKDD, LNAI 2431, pp. 62-73, 2002.

Since many popular algorithms (e.g. C4.5 [7], CART [8], neural networks [9], k-nearest neighbor [10]) were developed using what are now considered small data sets (hundreds to thousands of instances), it is possible that they are tailored to more effective learning from small data sets than large. It is possible that if data set sizes common today were common when these algorithms were developed, the evolution of such algorithms may have proceeded down a different path.

So far, few approaches to dealing with large data sets have attempted to create totally new algorithms designed specifically for today's data set sizes. In fact, many "new" algorithms are actually more efficient means of searching the hypothesis space while finding the same solution. Examples include RainForest [11], and ADTree [12]. It would seem logical that new algorithms, specifically designed for use with large data sets, are at least worth exploring.

This is not to say that efficiency is unimportant. There is little value in an algorithm that can produce miraculously good models, but takes so long to do so that the models are no longer useful. There must be a balance between efficiency and accuracy. This again lends support to the utility of algorithms fundamentally designed for processing large data sets.

The paper is set out as follows. Section 2 looks at issues relating to learning from large data sets. Section 3 details experiments performed and presents and discusses associated results. Conclusions and possible areas for further work are outlined in Section 4.

2 Learning from Large Data Sets

We have hypothesized that effective learning from large data sets may require different strategies to effective learning from small data sets. How then, might such strategies differ? This section examines multiple facets of this issue.

2.1 Efficiency

Certainly, efficiency is of fundamental importance. Small data set algorithms can afford to be of order $O(n^3)$ or higher, as with small data set sizes a model can still be formed in a reasonable time. When dealing with large data sets, however, algorithms of order $O(n^2)$ can be too computationally complex to be realistically useable. Therefore, algorithms for use with large data sets must have a low order of complexity, preferably no higher than $O(n)$.

It is worth noting that many algorithms can be made more efficient through parallelisation [13]. Splitting processing between multiple processors can be expected to reduce execution time, but only when splitting is possible. For example, the popular Boosting algorithms such as AdaBoost [14] and Arc-x4 [15] would seem prime candidates for parallelisation, as multiple models are produced. Unfortunately, this is not so, as the input of a model depends on the output of previous models (although parallelising the model building algorithms may still be possible). Bagging [16] and MultiBoost [17] are, on the other hand, suitable for parallelisation.

However, techniques such as parallelisation (or, for that matter, sampling) only reduce execution time. They do not make an algorithm fundamentally more suitable for large data sets.

2.2 Bias and Variance

What other fundamental properties of machine learning algorithms are required for learning from large data sets? This research focuses on the bias plus variance decomposition of error as a possible method of designing algorithms for use with large data sets.

The bias of a classification learning algorithm is a measure of the error that can be attributed to the central tendency of the models formed by the learner from different samples. The variance is a measure of the error that can be attributed to deviations from the central tendency of the models formed from different samples.

Unfortunately, while there is a straight-forward and generally accepted measure of these terms in the context of regression (prediction of numeric values), it is less straight-forward to derive an appropriate measure in a classification learning context. Alternative definitions include those of Kong & Dietterich [18], Kohavi & Wolpert [19], James & Hastie [20], Friedman [21], and Webb [17]. Of these numerous definitions we adopt Kohavi & Wolpert's definition [19] as it appears to the most commonly employed in experimental machine learning research.

2.3 Bias and Variance and Data Set Size

We assume in the following that training data is an iid sample. As the data set size increases, the expected variance between different samples can be expected to decrease. As the differences between alternative samples decreases, the differences between the alternative models formed from those samples can also be expected to decrease. As differences between the models decrease, differences between predictions can also be expected to decrease. In consequence, when learning from large data sets we should expect variance to be lower than when learning from small data sets.

2.4 Management of Bias and Variance

It is credible that the effectiveness of many of a learning algorithm's strategies can be primarily attributed either to a reduction of bias or a reduction of variance. For example, there is evidence that decision tree pruning is primarily effective due to an ability to reduce variance [22]. If learning from small data sets requires effective variance management, it is credible that early learning algorithms, focusing on the needs of small data sets, lent more weight to strategies that are effective at variance management than those that are effective at bias management.

It is important to note that better management of bias does not necessarily equate to lower error due to bias (the same holds for variance). It can be trivially shown that an algorithm with better bias management can have worse predictive performance than an algorithm with less bias management. Therefore, the level of bias and variance management should be viewed as a good guide to performance, not a guarantee.

Both bias and variance management are important. However, if, as has been discussed above, variance can be expected to decrease as training set size increases regardless of the level of variance management, then it would seem logical that more focus can be placed on bias management without significant expected loss of accuracy due to an increase in variance error. The following experiments investigate whether this is true.

2.5 Hypothesis

There are two parts to the hypothesis. The first is that as training set size increases variance will decrease. The second is that as training set size increases, variance will become a less significant part of error. This is based on the stronger expectation for variance to decrease as training size increases than bias. Therefore, it seems plausible that the proportion of decrease in variance will be greater than that for bias.

3 Experiments

Experiments were performed to provide evidence towards the hypothesis. As discussed previously, different algorithms have different bias plus variance profiles. Thus, algorithms with a range of bias plus variance profiles were selected for testing. The first was Naïve Bayes, selected due to its extremely high variance management and extremely low bias management. The second algorithm was the decision tree exemplar C4.5. The many options of C4.5 allow small but important changes to the induction algorithm, altering the bias plus variance profile. It was therefore possible to investigate multiple profiles using the same basic algorithm. This helps in ensuring that any differences in trends found in different profiles are due to the differences in the profiles, not differences in the basic algorithm. The variants investigated were C4.5 with its default options (including pruning), C4.5 without pruning, and C4.5 without pruning and with the minimum number of instances per leaf set to 1. The MultiBoost [17] "meta-algorithm" was also used (with standard C4.5 as its base algorithm) as it has been shown to reduce both bias and variance. Table 1 details the algorithms used, and their associated expected bias plus variance profiles.

Pruning of decision trees has been shown to reduce variance [22]. Therefore, growing trees without pruning should reduce variance management. Reducing the number of instances required at a decision leaf in C4.5 should also result in lower variance management.

3.1 Methodology

Experiments were performed as follows. A data set was divided into three parts. One part was used as the hold-out test set. The training set was randomly sampled without replacement from the remaining two parts. A model was created and tested on the hold-out set. This sub-process was then repeated using each of the other two parts as the hold-out test set. This guarantees that each instance is classified once. The whole process was repeated ten times. Each instance is therefore classified precisely ten

Table 1. Selected algorithms and their bias plus variance profiles

ALGORITHM	BIAS PLUS VARIANCE PROFILE
NAÏVE BAYES	Very high variance management, very little bias management
C4.5	Medium variance management, medium bias management
MULTIBOOST C4.5	More bias and variance management than C4.5
C4.5 WITHOUT PRUNING	Less variance management than C4.5
C4.5 WITHOUT PRUNING, MINIMUM OF 1 INSTANCE AT LEAF	Very little variance management

times as a test instance, and used up to twenty times as a training instance. Training set sample sizes were powers of two - ranging from 32 instances to the highest power of 2 that was less than two-thirds of the number of instances in the entire data set.

Seven freely available data sets from the UCI Machine Learning Repository [23] were used (outlined in Table 2).

Data sets were required to be: a) useable for classification, b) large enough for use with the methodology, so as to provide a sufficiently large maximum training set size, and c) publicly available.

Table 2. Description of data sets

DATA SET	NUMBER OF INSTANCES	CONT ATTRS	DISC ATTRS	CLASSES
ADULT	48,842	6	8	2
CENSUS INCOME	199,523	7	33	2
CONNECT-4	67,557	0	42	3
COVER TYPE	581,012	10	44	7
IPUMS	88,443	60	0	13
SHUTTLE	58,000	9	0	7
WAVEFORM	1,600,000	21	0	3

3.2 Results

Graphs show the relation of bias or variance to data set size for all algorithms used. Note that the error for Naïve-Bayes on the waveform data set increases dramatically at training set size 32,768. This occurred for all measures, and was investigated with no clear reason for such behavior found.

3.2.1 Variance

See Figure 1(a-g). In general, all algorithms follow the trend to decrease in variance as training set size increases for all data sets. The one exception is Naïve Bayes on the Census-Income data, where there are substantial increases in variance.

3.2.2 Bias

See Figure 2(a-g). For all data sets all algorithms except Naïve-Bayes tend to decrease in bias as training set size increases. Naïve-Bayes, an algorithm with very little bias management, increases in bias for all data sets except waveform. Although no hypothesis was offered regarding the trend of bias alone, this suggests that bias management is extremely important.

3.2.3 Ratio of Bias to Variance

See Figure 3(a-g). Note that results are presented as the proportion of bias of overall error, rather than a direct relation of bias to variance for simplification of scales. The results show that varying training set size can have different effects on bias and variance profiles. To evaluate the effect of increasing training set size on the relative importance of bias and variance, we look at the difference in the ratio of bias to variance between the smallest and the largest training set size for each data set. If the ratio increases then bias is increasing in the amount it dominates the final error term. If the ratio decreases then the degree to which variance dominates the final error term is increasing.

The second part of the hypothesis is that variance will become a larger portion of the error with increasing training set size. The comparison found that of the 35 comparisons, 28 were in favor of the hypothesis, with only 7 against. This is significant at the 0.05 level, using a one-tailed sign test (p=0.0003).

3.3 Summary

The results show a general trend for variance to decrease with increased training set size. The trend is certainly not as strong as that for bias. However, this trend exists with all algorithms used. Even unpruned C4.5 with minimum leaf instance of one, an algorithm with extremely little variance management, shows the trend. This suggests that variance management may not be of extreme importance in an algorithm when dealing with large data sets. This is not to suggest that variance management is unnecessary, since more variance management can still be expected to result in less variance error. However, these results do suggest that, as expected, variance will naturally decrease with larger training set sizes. The results also support the second part of the hypothesis; that bias can be expected to become a larger portion of error.

3.4 Does Lower Variance Management Imply Higher Bias Management?

It might be thought that management of bias and variance are interlinked so that approaches to reduce bias will increase variance and vice versa. This hypothesis was evaluated with respect to our experiments by examining the effects on bias and variance of the variants of C4.5. The bias and variance of each of the three variants (unpruned, unpruned with minimum leaf size of 1, and MultiBoosting) were compared to the bias and variance of C4.5 with default settings. The number of times the signs of the differences differed (190) was compared with the number of times the

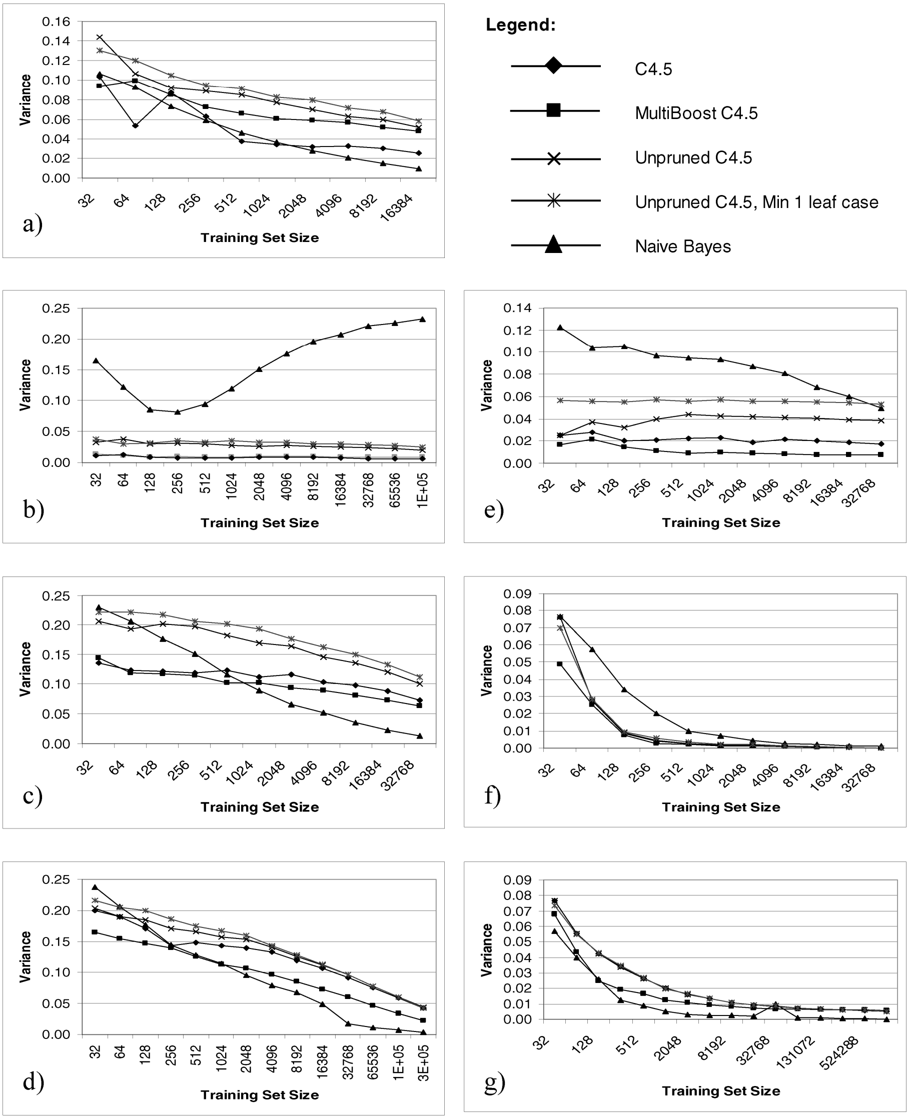

Fig. 1. Variance of algorithms on data sets for different training set sizes. Data sets are a) Adult, b) Census Income, c) Connect-4, d) Cover Type, e) IPUMS, f) Shuttle, g) Waveform

signs were the same (59), with occasions where there was no difference (9) ignored. A one-tailed binomial sign test ($p < 0.0001$) indicates that this outcome indicates a significantly greater chance that a decrease in variance will correspond with an increase in bias and vice versa than that they will both increase or decrease in unison as a result of a modification to a learning algorithm.

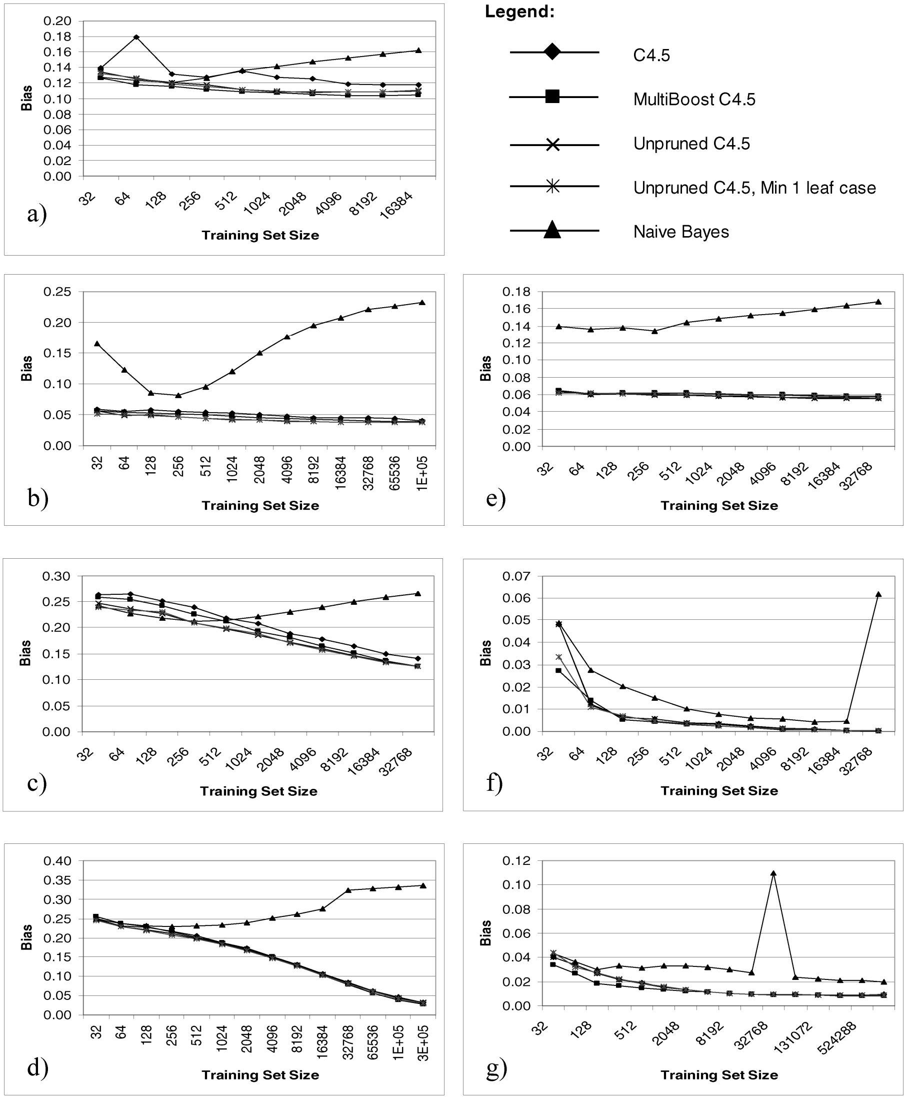

Fig. 2. Bias of algorithms on data sets for different training set sizes. Data sets are a) Adult, b) Census Income, c) Connect-4, d) Cover Type, e) IPUMS, f) Shuttle, g) Waveform

This might be taken as justification for not aiming to manage bias in preference to managing variance at large data set sizes, as managing bias will come at the expense of managing variance. However, while the directions of the effects on bias and variance tend to be the opposite, the magnitudes also differ. The mean absolute difference between the variance of C4.5 with default settings and one of its variants was 0.0221. The mean difference between the bias of C4.5 with default settings and

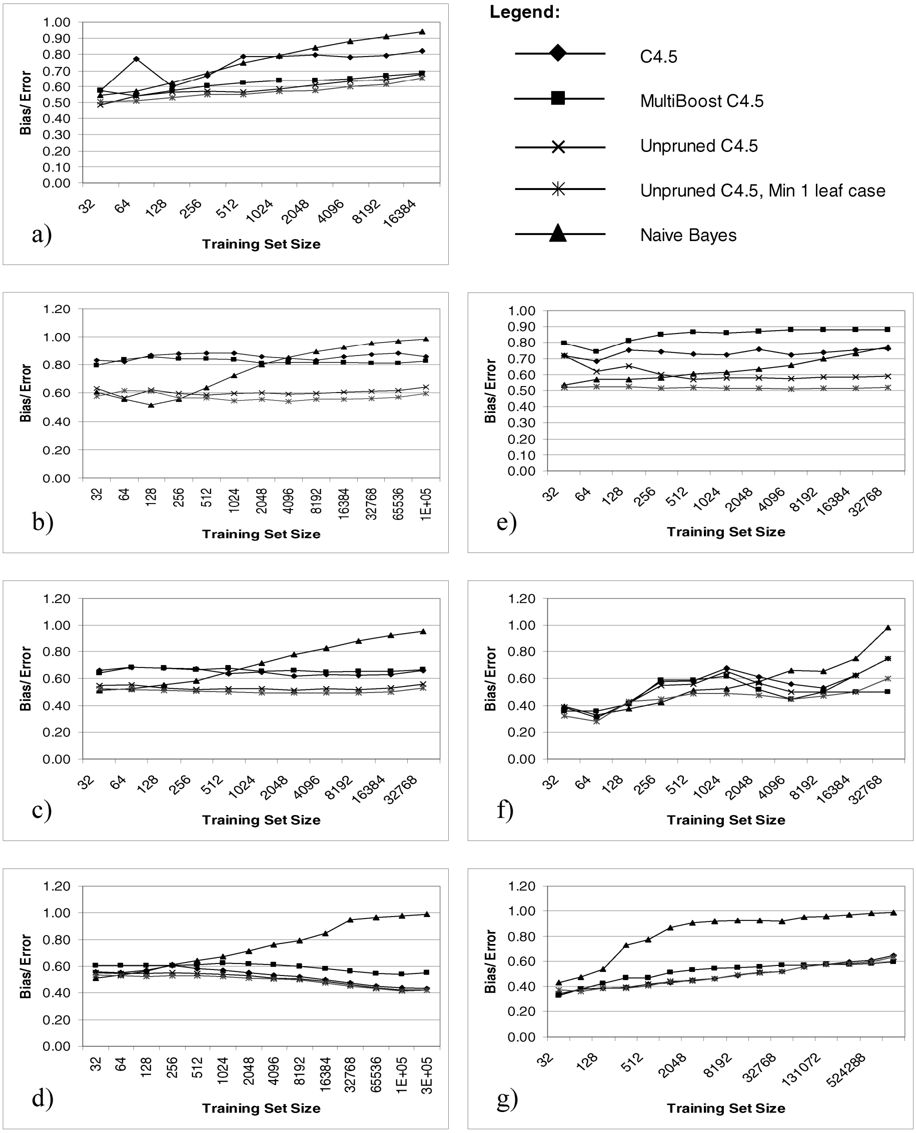

Fig. 3. Ratio of bias to variance of algorithms on data sets for different training set sizes. Data sets are a) Adult, b) Census Income, c) Connect-4, d) Cover Type, e) IPUMS, f) Shuttle, g) Waveform

one of its variants was 0.0101. A one-tailed matched-pair t-test indicates that the effect on variance of the variants of C4.5 is greater than the effect on bias. This adds credibility to the hypothesis that current machine learning algorithms reflect their small data set origins by incorporating primarily variance management measures.

4 Conclusions and Further Work

This paper details experiments performed to investigate whether a) the statistical expectations of bias and variance do indeed apply to classification learning, and b) if bias becomes a larger portion of error as training set size increases. The results support both parts of the hypothesis. Variance, in general, decreases as training set size increases. This appears to be irrespective of the bias plus variance profile of the algorithm. Bias also generally decreases, with more regularity than variance. The one notable exception to this is Naïve-Bayes, an algorithm that employs little bias management. This somewhat surprising result alone suggests that bias management is indeed an important factor in learning from large data sets. An analysis of the impact on bias and variance of changes to a learning algorithm suggest that measures that decrease variance can be expected to increase bias and vice versa. However, the magnitudes of these changes differ markedly, variance being affected more than bias. This suggests that the measures incorporated in standard learning algorithms do indeed relate more to variance management than bias management. The results also show that as training set size increases bias can be expected to become a larger portion of error.

Unfortunately, creating algorithms that focus on bias management seems to be a difficult task. We can, however, identify some methods that may be expected to lower bias. One possibility is to create algorithms with a "larger than usual" hypothesis space. For example, it could be expected that an algorithm that can create non-axis-orthogonal partitions should have less bias than an algorithm that can only perform axis-orthogonal partitions. The drawback of this is an increase in search. Another option might be to introduce a large random factor into the creation of a model. This could be expected to convert at least some of the bias error into variance error. However, the way in which randomization should be included does not appear obvious.

These experiments are by no means exhaustive. Thus, there is scope for continued investigation using a wider range of algorithms, and more and larger data sets.

The data sets used in this research are not considered particularly large by today's standards. Unfortunately, hardware constraints limited the size of the data sets used. However, even with the data sets employed in this study, trends are apparent. It is reasonable to expect that with massive data sets these trends should continue, and possibly become stronger.

We have shown that the importance of bias management grows as data set size grows. We have further presented evidence that current algorithms are oriented more toward management of variance than management of bias. We believe that the strong implication of this work is that classification learning error from large data sets may be further reduced by the development of learning algorithms that place greater emphasis on reduction of bias.

References

1. Provost, F., Aronis, J.: Scaling Up Inductive Learning with Massive Parallelism. Machine Learning, Vol. 23. (1996) 33-46
2. Provost, F., Jensen, D., Oates, T.: Efficient Progressive Sampling. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. ACM Press, New York (1999) 22-32
3. Shafer, J., Agrawal, R., Mehta, M.: SPRINT: A Scalable Parallel Classifier for Data Mining. Proceedings of the Twenty-Second VLDB Conference. Morgan Kaufmann, San Francisco (1996) 544-555
4. Catlett, J.: Peepholing: Choosing Attributes Efficiently for Megainduction. Proceedings of the Ninth International Conference on Machine Learning. Morgan Kaufmann, San Mateo (1992) 49-54
5. Cohen, W.: Fast Effective Rule Induction. Proceedings of the Twelfth International Conference on Machine Learning. Morgan Kaufmann, San Francisco (1995) 115-123
6. Aronis, J., Provost, F.: Increasing the Efficiency of Data Mining Algorithms with Breadth-First Marker Propagation. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining. AAAI Press, Menlo Park (1997) 119-122
7. Quinlan, J. R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993)
8. Breiman, L., Freidman, J. H., Olshen, R. A., Stone, C. J.: Classification and Regression Trees. Wadsworth International, Belmont (1984)
9. Hecht-Nielsen, R.: Neurocomputing. Addison-Wesley, Menlo Park (1990)
10. Cover, T. M., Hart, P. E.: Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory, Vol. 13. (1967) 21-27
11. Gehrke, J., Ramakrishnan, R., Ganti, V.: RainForest – A Framework for Fast Decision Tree Induction. Proceedings of the Twenty-fourth International Conference on Very Large Databases. Morgan Kaufmann, San Mateo (1998)
12. Moore, A., Lee, M. S.: Cached Sufficient Statistics for Efficient Machine Learning with Large Datasets. Journal of Artificial Intelligence Research, Vol. 8. (1998) 67-91
13. Chattratichat, J., Darlington, J., Ghanem, M., Guo, Y., Huning, H., Kohler, M., Sutiwaraphun, J., To, H. W., Yang, D.: Large Scale Data Mining: Challenges and Responses. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining. AAAI Press, Menlo Park (1997)
14. Freund, Y., Schapire, R. E.: A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting. Journal of Computer and System Sciences, Vol. 55. (1997) 95-121
15. Breiman, L.: Arcing Classifiers. Technical Report 460. Department of Statistics, University of California, Berkeley (1996)
16. Breiman, L.: Bagging Predictors. Machine Learning, Vol. 24. (1996) 123-140.
17. Webb, G. (2000). MultiBoosting: A Technique for Combining Boosting and Wagging. Machine Learning, Vol. 40, (2000) 159-196

18. Kong, E. B., Dietterich, T. G.: Error-Correcting Output Coding Corrects Bias and Variance. Proceedings of the Twelfth International Conference on Machine Learning. Morgan Kaufmann, San Mateo (1995)
19. Kohavi, R., Wolpert, D. H.: Bias Plus Variance Decomposition for Zero-One Loss Functions. Proceedings of the Thirteenth International Conference on Machine Learning. Morgan Kaufmann, San Francisco (1996)
20. James, G, Hastie, T.: Generalizations of the bias/variance decomposition for prediction error. Technical Report. Department of Statistics, Stanford University (1997)
21. Friedman, J. H.: On Bias, Variance, 0/1-Loss, and the Curse-of-Dimensionality. Data Mining and Knowledge Discovery, Vol. 1. (1997) 55-77
22. Bauer, E., Kohavi, R.: An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning, Vol. 36. (1999) 105-142
23. Blake, C. L., Merz, C. J. UCI Repository of Machine Learning Databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Department of Information and Computer Science, University of California, Irvine