# On The Size of Training Set and The Benefit from Ensemble

Zhi-Hua Zhou[1], Dan Wei[1], Gang Li[2], and Honghua Dai[2]

[1] National Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
zhouzh@nju.edu.cn dwei@ai.nju.edu.cn
[2] School of Information Technology
Deakin University, Burwood, Vic3125, Australia
{gangli, hdai}@deakin.edu.au

**Abstract.** In this paper, the impact of the size of the training set on the benefit from ensemble, i.e. the gains obtained by employing ensemble learning paradigms, is empirically studied. Experiments on Bagged/ Boosted J4.8 decision trees with/without pruning show that enlarging the training set tends to improve the benefit from Boosting but does not significantly impact the benefit from Bagging. This phenomenon is then explained from the view of bias-variance reduction. Moreover, it is shown that even for Boosting, the benefit does not always increase consistently along with the increase of the training set size since single learners sometimes may learn relatively more from additional training data that are randomly provided than ensembles do. Furthermore, it is observed that the benefit from ensemble of unpruned decision trees is usually bigger than that from ensemble of pruned decision trees. This phenomenon is then explained from the view of error-ambiguity balance.

## 1 Introduction

Ensemble learning paradigms train a collection of learners to solve a problem. Since the generalization ability of an ensemble is usually better than that of a single learner, one of the most active areas of research in supervised learning has been to study paradigms for constructing good ensembles [5].

This paper does not attempt to propose any new ensemble algorithm. Instead, it tries to explore how the change of the training set size impacts the benefit from ensemble, i.e. the gains obtained by employing ensemble learning paradigms. Having an insight into this may be helpful to better exerting the potential of ensemble learning paradigms. This goal is pursued in this paper with an empirical study on ensembles of pruned or unpruned J4.8 decision trees [9] generated by two popular ensemble algorithms, i.e. Bagging [3] and Boosting (In fact, *Boosting* is a family of ensemble algorithms, but here the term is used to refer the most famous member of this family, i.e. AdaBoost [6]). Experimental results show that enlarging training set does not necessarily enlarges the benefit from ensemble. Moreover, interesting issues on the benefit from ensemble, which is related to the characteristics of Bagging and Boosting and the effect of decision tree pruning, have been disclosed and discussed.

The rest of this paper is organized as follows. Section 2 describes the empirical study. Section 3 analyzes the experimental results. Section 4 summarizes the observations and derivations.

## 2 The Empirical Study

Twelve data sets with 2,000 to 7,200 examples, 10 to 36 attributes, and 2 to 10 classes from the UCI Machine Learning Repository [2] are used in the empirical study. Information on the experimental data sets are tabulated in Table 1.

**Table 1.** Experimental data sets

| Data set | Size | Attribute | | Class |
|---|---|---|---|---|
| | | Categorical | Continuous | |
| *allbp* | 2,800 | 22 | 7 | 3 |
| *ann* | 7,200 | 15 | 6 | 3 |
| *block* | 5,473 | 0 | 10 | 5 |
| *hypothyroid* | 3,772 | 22 | 7 | 2 |
| *kr-vs-kp* | 3,196 | 36 | 0 | 2 |
| *led7* | 2,000 | 7 | 0 | 10 |
| *led24* | 2,000 | 24 | 0 | 10 |
| *sat* | 6,435 | 0 | 36 | 6 |
| *segment* | 2,310 | 0 | 19 | 7 |
| *sick* | 3,772 | 22 | 7 | 2 |
| *sick-euthyroid* | 3,156 | 22 | 7 | 2 |
| *waveform* | 5,000 | 0 | 21 | 3 |

Each original data set is partitioned into ten subsets with similar distributions. At the first time, only one subset is used; at the second time, two subsets are used; and so on. The earlier generated data sets are proper subsets of the later ones. In this way, the increase of the size of the data set is simulated.

On each generated data set, 10-fold cross validation is performed. In each fold, Bagging and Boosting are respectively employed to train an ensemble comprising 20 pruned or unpruned J4.8 decision trees. For comparison, a single J4.8 decision tree is also trained from the training set of the ensembles. The whole process is repeated for ten times, and the average error ratios of the ensembles generated by Bagging and Boosting against the single decision trees are recorded, as shown in Tables 2 to 5, respectively. The predictive error rates of the single decision trees are shown in Tables 6 and 7. In these tables the first row indicates the percentage of data in the original data sets that are used, and the numbers following '±' are the standard deviations.

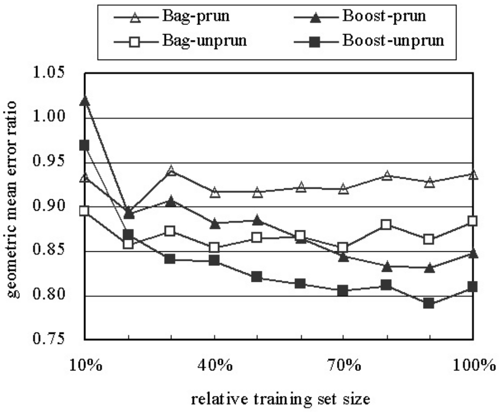

**Fig. 1.** The geometrical mean error ratio of Bagged/Boosted J4.8 decision trees with/without pruning against single J4.8 decision trees with/without pruning

Here the error ratio is defined as the result of dividing the predictive error rate of an ensemble by that of a single decision tree. A smaller error ratio means relatively bigger benefit from ensemble, while a bigger error ratio means relative smaller benefit from ensemble. If an ensemble is worse than a single decision tree, then its error ratio is bigger than 1.0.

In order to exhibit the overall tendency, the geometric mean error ratio, i.e. average ratio across all data sets, are also provided in Tables 2 to 5, which is then explicitly depicted in Fig. 1.

## 3 Discussion

### 3.1 Bagging and Boosting

An interesting phenomenon exposed by Fig. 1 and Tables 2 to 5 is that the benefit from Bagging and Boosting exhibit quite different behaviors on the change of the training set size. In detail, although there are some fluctuations, the benefit from Bagging remains relatively unvaried while that from Boosting tends to be enlarged when the training set size increases. In order to explain this phenomenon, it may be helpful to consider the different characteristics of Bagging and Boosting from the view of bias-variance reduction.

Given a learning target and the size of training set, the expected error of a learning algorithm can be broken into the sum of three non-negative quantities, i.e. the intrinsic noise, the bias, and the variance [7]. The intrinsic noise is a lower bound on the expected error of any learning algorithm on the target. The bias measures how closely the average estimate of the learning algorithm is able to approximate the target. The variance measures how much the estimate of the learning algorithm fluctuates for the different training sets of the same size.

**Table 2.** Error ratios of Bagged J4.8 decision trees against single J4.8 decision trees. All the trees are pruned.

| Data set | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|
| *allbp* | .909±.193 | .914±.111 | .952±.070 | 1.01±.125 | .961±.149 |
| *ann* | 1.15±.203 | .899±.135 | 1.03±.180 | .952±.209 | .931±.086 |
| *block* | .911±.186 | .908±.061 | .906±.119 | .886±.067 | .885±.046 |
| *hypothyroid* | 1.10±.236 | 1.01±.223 | 1.14±.243 | 1.04±.135 | 1.19±.247 |
| *kr-vs-kp* | .938±.104 | .897±.180 | 1.06±.376 | .991±.138 | .892±.119 |
| *led7* | .976±.054 | .976±.049 | .978±.017 | .986±.025 | .988±.019 |
| *led24* | .902±.063 | .920±.042 | .941±.048 | .961±.027 | .958±.035 |
| *sat* | .786±.052 | .751±.031 | .736±.043 | .722±.040 | .729±.021 |
| *segment* | .800±.142 | .898±.128 | .913±.106 | .851±.105 | .816±.102 |
| *sick* | 1.22±.352 | .938±.201 | .999±.121 | .975±.150 | 1.08±.143 |
| *sick-euthyroid* | .826±.257 | .952±.212 | .955±.159 | .945±.139 | .911±.080 |
| *waveform* | .672±.138 | .663±.068 | .664±.084 | .671±.073 | .655±.056 |
| geometric-mean | .933 | .894 | .940 | .916 | .916 |

| Data set | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|
| *allbp* | .958±.078 | 1.02±.085 | .940±.080 | 1.01±.116 | 1.01±.038 |
| *ann* | 1.08±.170 | .985±.152 | 1.04±.140 | 1.09±.161 | 1.15±.162 |
| *block* | .876±.043 | .907±.057 | .873±.055 | .864±.044 | .862±.048 |
| *hypothyroid* | 1.14±.255 | .908±.132 | 1.04±.194 | .955±.076 | .973±.089 |
| *kr-vs-kp* | .810±.169 | .957±.136 | 1.02±.142 | 1.02±.130 | 1.07±.082 |
| *led7* | .996±.014 | 1.01±.014 | 1.00±.012 | 1.01±.011 | 1.01±.013 |
| *led24* | .960±.036 | .967±.023 | .969±.028 | .972±.028 | .979±.021 |
| *sat* | .703±.029 | .719±.027 | .724±.014 | .715±.019 | .704±.026 |
| *segment* | .849±.121 | .859±.085 | .845±.086 | .826±.095 | .857±.092 |
| *sick* | 1.07±.096 | 1.10±.250 | 1.14±.365 | .978±.104 | .954±.081 |
| *sick-euthyroid* | .950±.129 | .956±.119 | .941±.085 | .992±.068 | 1.00±.108 |
| *waveform* | .670±.057 | .650±.037 | .690±.051 | .699±.028 | .679±.030 |
| geometric-mean | .922 | .920 | .935 | .928 | .937 |

Since the intrinsic noise is an inherent property of the given target, usually only the bias and variance are concerned.

Previous research shows that Bagging works mainly through reducing the variance [1][4]. It is evident that such a reduction is realized by utilizing bootstrap sampling to capture the variance among the possible training sets under the given size and then smoothing the variance through combining the trained component learners. Suppose the original data set is $D$, a new data set $D'$ is bootstrap sampled from $D$, and the size of $D'$ is the same as that of $D$, i.e. $|D|$. Then, the size of the shared part between $D$ and $D'$ can be estimated according to Eq. 1, which shows that the average overlap ratio is a constant, roughly 63.2%.

$$\left(1-(1-1/|D|)^{|D|}\right)|D| \approx (1-0.368)\,|D| = 0.632\,|D| \tag{1}$$

**Table 3.** Error ratios of Boosted J4.8 decision trees against single J4.8 decision trees. All the trees are pruned.

| Data set | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|
| *allbp* | 1.00±.190 | .930±.202 | .895±.096 | 1.02±.163 | .883±.157 | .896±.077 | .931±.092 | .817±.059 | .844±.058 | .854±.041 |
| *ann* | 1.26±.349 | 1.01±.238 | 1.07±.138 | .926±.116 | .924±.145 | 1.04±.135 | .981±.089 | .894±.073 | .977±.118 | 1.09±.146 |
| *block* | .858±.101 | .934±.110 | .963±.149 | .951±.073 | 1.01±.044 | .986±.050 | .979±.061 | .986±.074 | .997±.039 | .968±.033 |
| *hypothyroid* | 1.63±.506 | .983±.197 | 1.07±.236 | .924±.394 | 1.12±.394 | .984±.314 | .797±.214 | .886±.255 | .720±.184 | .771±.130 |
| *kr-vs-kp* | .707±.244 | .669±.151 | .823±.261 | .827±.179 | .751±.113 | .549±.220 | .522±.160 | .559±.153 | .623±.243 | .652±.177 |
| *led7* | .998±.008 | 1.00±.000 | 1.00±.000 | 1.00±.000 | 1.00±.000 | 1.00±.000 | 1.00±.000 | 1.00±.000 | 1.00±.000 | 1.00±.000 |
| *led24* | 1.06±.104 | 1.09±.079 | 1.16±.069 | 1.15±.054 | 1.17±.059 | 1.16±.033 | 1.19±.042 | 1.20±.030 | 1.18±.028 | 1.18±.032 |
| *sat* | .717±.041 | .679±.047 | .653±.040 | .658±.038 | .657±.025 | .621±.031 | .656±.037 | .645±.028 | .641±.015 | .636±.018 |
| *segment* | .783±.138 | .702±.226 | .654±.084 | .605±.139 | .559±.141 | .589±.149 | .519±.086 | .497±.079 | .491±.110 | .523±.092 |
| *sick* | 1.47±.749 | 1.05±.251 | 1.02±.130 | .898±.113 | .956±.071 | .967±.149 | .934±.179 | .896±.167 | .838±.165 | .820±.078 |
| *sick-euthyroid* | 1.09±.329 | 1.02±.272 | .971±.221 | .980±.209 | .969±.162 | .952±.224 | .990±.193 | .975±.170 | 1.02±.217 | 1.02±.229 |
| *waveform* | .675±.165 | .644±.075 | .604±.057 | .628±.061 | .626±.056 | .624±.047 | .629±.043 | .655±.042 | .650±.032 | .669±.029 |
| geometric-mean | 1.02 | .893 | .907 | .881 | .885 | .864 | .844 | .834 | .832 | .849 |

This means that the variance among the possible samples with the same size that could be captured by a given number of trials of bootstrap sampling might not significantly change when the training set size changes. Therefore, when the training set size increases, the improvement of the ensemble owes much to the improvement of the component learners caused by the additional training data instead of the capturing of more variance through bootstrap sampling. Since the single learner also improves on the additional training data in the same way as the component learners in the ensemble do, the benefit from Bagging might not be significantly changed when the training set is enlarged.

As for Boosting, previous research shows that it works through reducing both the bias and variance but primarily through reducing the bias [1][4]. It is evident that such a reduction on bias is realized mainly by utilizing adaptive sampling,

**Table 4.** Error ratios of Bagged J4.8 decision trees against single J4.8 decision trees. All the trees are unpruned.

| Data set | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|
| *allbp* | 1.03±.047 | .733±.029 | .965±.005 | .890±.129 | .898±.107 |
| *ann* | .995±.111 | .980±.109 | .985±.146 | 1.00±.178 | .879±.153 |
| *block* | .856±.205 | .902±.058 | .862±.120 | .847±.059 | .830±.049 |
| *hypothyroid* | 1.08±.149 | .867±.283 | 1.02±.236 | .864±.078 | 1.07±.292 |
| *kr-vs-kp* | .865±.151 | .998±.325 | .891±.200 | .899±.141 | .887±.252 |
| *led7* | .979±.041 | .968±.045 | .978±.021 | .999±.026 | .988±.019 |
| *led24* | .823±.082 | .803±.046 | .803±.043 | .827±.030 | .806±.043 |
| *sat* | .750±.061 | .735±.036 | .716±.051 | .693±.039 | .702±.017 |
| *segment* | .795±.149 | .871±.127 | .898±.121 | .812±.085 | .810±.110 |
| *sick* | 1.08±.337 | .927±.188 | .866±.091 | .883±.148 | 1.00±.112 |
| *sick-euthyroid* | .818±.203 | .848±.100 | .827±.079 | .858±.113 | .836±.074 |
| *waveform* | .669±.134 | .664±.067 | .668±.090 | .665±.070 | .656±.062 |
| geometric-mean | .895 | .858 | .873 | .853 | .864 |

| Data set | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|
| *allbp* | .906±.089 | .885±.035 | .876±.077 | .858±.030 | .901±.025 |
| *ann* | .991±.149 | 1.01±.095 | .953±.111 | 1.11±.245 | 1.23±.273 |
| *block* | .836±.092 | .841±.028 | .872±.065 | .844±.030 | .804±.044 |
| *hypothyroid* | 1.11±.453 | .788±.108 | .993±.355 | .870±.160 | .858±.066 |
| *kr-vs-kp* | .864±.158 | .869±.144 | 1.02±.181 | .893±.133 | 1.01±.135 |
| *led7* | .997±.015 | 1.01±.012 | .996±.016 | 1.01±.009 | 1.01±.013 |
| *led24* | .796±.029 | .785±.022 | .800±.018 | .799±.017 | .807±.013 |
| *sat* | .680±.026 | .687±.027 | .699±.014 | .694±.023 | .685±.023 |
| *segment* | .792±.138 | .822±.105 | .808±.071 | .806±.078 | .806±.096 |
| *sick* | .976±.145 | 1.07±.170 | 1.03±.242 | .940±.121 | .909±.064 |
| *sick-euthyroid* | .795±.097 | .824±.091 | .832±.077 | .843±.062 | .924±.062 |
| *waveform* | .663±.051 | .647±.028 | .682±.053 | .693±.025 | .669±.028 |
| geometric-mean | .867 | .853 | .880 | .863 | .884 |

i.e. adaptively changing the data distribution to enable a component learner focus on hard examples for its predecessor. When the training set is enlarged, the adaptive sampling process becomes more effective since more hard examples for a component learner could be effectively identified and then passed on to the successive learner, some of which might not be identified when the training set is a relatively smaller one. Therefore, the reduction on bias may be enhanced along with the increase of the size of training set, which leads to that the benefit from Boosting tends to be enlarged.

It is worth noting that Fig. 1 and Tables 2 to 5 also show that the benefit from ensemble, even for Boosting, does not always increase consistently when the training set size increases. This is not difficult to understand because the chances for an ensemble to get improved from the additional training data that

**Table 5.** Error ratios of Boosted J4.8 decision trees against single J4.8 decision trees. All the trees are unpruned.

| Data set | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|
| *allbp* | .953±.187 | .746±.090 | .984±.128 | .883±.175 | .851±.131 | .875±.138 | .875±.054 | .751±.105 | .755±.069 | .727±.050 |
| *ann* | 1.47±.300 | 1.07±.123 | .939±.146 | 1.06±.366 | .832±.168 | .989±.221 | 1.00±.108 | .948±.159 | 1.02±.150 | 1.17±.103 |
| *block* | .860±.210 | .898±.122 | .911±.075 | .890±.032 | .943±.048 | .928±.048 | .938±.075 | .975±.048 | .930±.011 | .923±.051 |
| *hypothyroid* | 1.44±.453 | .949±.345 | .951±.350 | .753±.316 | 1.00±.262 | .869±.366 | .742±.291 | .811±.118 | .696±.121 | .726±.103 |
| *kr-vs-kp* | .646±.243 | .757±.236 | .762±.281 | .865±.315 | .656±.197 | .673±.282 | .585±.075 | .719±.175 | .619±.066 | .613±.143 |
| *led7* | 1.00±.000 | 1.00±.000 | 1.00±.000 | 1.00±.000 | 1.00±.000 | 1.00±.000 | 1.00±.000 | 1.00±.000 | 1.00±.000 | 1.00±.000 |
| *led24* | .968±.107 | .974±.061 | .956±.052 | .980±.022 | .965±.036 | .953±.026 | .958±.020 | .976±.024 | .959±.020 | .972±.016 |
| *sat* | .691±.071 | .672±.070 | .640±.038 | .631±.029 | .624±.015 | .614±.038 | .615±.027 | .611±.012 | .620±.020 | .624±.017 |
| *segment* | .723±.087 | .646±.103 | .549±.064 | .594±.097 | .539±.065 | .541±.064 | .506±.051 | .501±.047 | .497±.081 | .511±.057 |
| *sick* | 1.23±.886 | 1.11±.259 | .917±.130 | .876±.123 | .960±.152 | .892±.103 | .936±.147 | .931±.209 | .860±.122 | .864±.074 |
| *sick-euthyroid* | .985±.242 | .948±.171 | .860±.162 | .906±.097 | .877±.095 | .823±.077 | .870±.079 | .870±.089 | .895±.064 | .936±.078 |
| *waveform* | .665±.114 | .643±.101 | .621±.066 | .631±.063 | .600±.066 | .601±.038 | .638±.028 | .637±.045 | .636±.050 | .659±.040 |
| geometric-mean | .969 | .868 | .841 | .839 | .821 | .813 | .805 | .811 | .791 | .810 |

are randomly provided might be less than that of a single learner since the ensemble is usually far stronger than the single learner. It is analogous to the fact that improving a poor learner is more easier than improving a strong learner. Therefore, the benefit of ensemble shrinks if the improvement of the ensemble is not so big as that of the single learner on additional training data. However, if the additional training data have been adequately selected so that most of them can benefit the ensemble, then both the ensemble and the single learner could be significantly improved while the benefit from ensemble won't be decreased.

### 3.2 Pruned and Unpruned Trees

Another interesting phenomenon exposed by Fig. 1 and Tables 2 to 5 is that the benefit from ensemble comprising unpruned decision trees is always bigger than

**Table 6.** Predictive error rate (%) of pruned single C4.5 decision trees.

| Data set | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|
| *allbp* | 4.40±1.36 | 4.00±0.58 | 3.57±0.52 | 3.25±0.55 | 3.35±0.38 |
| *ann* | 0.98±0.41 | 0.73±0.25 | 0.52±0.16 | 0.49±0.14 | 0.45±0.08 |
| *block* | 4.39±1.00 | 3.97±0.27 | 3.59±0.53 | 3.60±0.32 | 3.19±0.29 |
| *hypothyroid* | 1.45±0.57 | 1.24±0.39 | 0.97±0.34 | 0.69±0.14 | 0.57±0.16 |
| *kr-vs-kp* | 4.62±1.37 | 2.75±0.58 | 1.75±0.57 | 1.36±0.40 | 1.14±0.26 |
| *led7* | 35.21±2.92 | 31.83±3.07 | 30.08±1.87 | 28.81±1.67 | 28.01±1.48 |
| *led24* | 36.19±6.13 | 33.17±3.35 | 30.67±2.83 | 30.93±1.49 | 29.93±1.41 |
| *sat* | 19.22±1.68 | 17.02±1.04 | 15.87±0.59 | 15.33±0.52 | 14.63±0.54 |
| *segment* | 9.16±3.07 | 6.94±2.00 | 5.85±1.41 | 5.59±1.37 | 4.94±1.21 |
| *sick* | 2.21±0.97 | 2.27±0.97 | 1.86±0.36 | 1.79±0.27 | 1.63±0.33 |
| *sick-euthyroid* | 4.07±1.97 | 3.39±1.68 | 3.20±1.56 | 2.95±1.00 | 2.67±0.80 |
| *waveform* | 11.27±1.87 | 10.43±1.59 | 10.44±1.00 | 9.70±0.93 | 9.96±0.58 |

| Data set | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|
| *allbp* | 3.10±0.32 | 2.88±0.28 | 2.94±0.30 | 2.87±0.29 | 2.76±0.17 |
| *ann* | 0.39±0.07 | 0.41±0.06 | 0.39±0.05 | 0.33±0.03 | 0.30±0.03 |
| *block* | 3.20±0.18 | 3.11±0.21 | 3.15±0.21 | 3.08±0.15 | 3.03±0.11 |
| *hypothyroid* | 0.53±0.16 | 0.54±0.12 | 0.53±0.11 | 0.49±0.06 | 0.45±0.04 |
| *kr-vs-kp* | 1.01±0.20 | 0.93±0.16 | 0.80±0.13 | 0.72±0.13 | 0.57±0.08 |
| *led7* | 27.82±1.41 | 27.39±0.68 | 27.02±0.56 | 26.90±0.50 | 26.73±0.27 |
| *led24* | 29.02±1.28 | 28.56±1.03 | 28.30±0.72 | 28.20±0.34 | 27.78±0.54 |
| *sat* | 14.83±0.59 | 14.11±0.46 | 14.11±0.53 | 13.70±0.49 | 13.54±0.30 |
| *segment* | 4.11±1.06 | 3.82±0.86 | 3.46±0.82 | 3.21±0.63 | 2.93±0.59 |
| *sick* | 1.50±0.30 | 1.42±0.28 | 1.33±0.35 | 1.39±0.25 | 1.38±0.29 |
| *sick-euthyroid* | 2.69±1.08 | 2.51±0.62 | 2.42±0.52 | 2.23±0.49 | 2.21±0.49 |
| *waveform* | 9.88±0.48 | 9.75±0.56 | 9.38±0.45 | 9.13±0.46 | 8.95±0.31 |

that comprising pruned decision trees, despite whether Bagging or Boosting is employed. In order to explain this phenomenon, it may be helpful to consider the effect of decision tree pruning from the view of error-ambiguity balance.

It has been shown that the generalization error of an ensemble can be decomposed into two terms, i.e. $E = \bar{E} - \bar{A}$, where $\bar{E}$ is the average generalization error of the component learners while $\bar{A}$ is the average ambiguity [8]. The smaller the $\bar{E}$ and the bigger the $\bar{A}$, the better the ensemble.

In general, the purpose of decision tree pruning is to avoid overfitting. With the help of pruning, the generalization ability of a decision tree is usually improved. Thus, the $\bar{E}$ of an ensemble comprising pruned decision trees may be smaller than that of an ensemble comprising unpruned decision trees. But on the other hand, pruning usually causes the decrease of the ambiguity among the decision trees. This is because some trees may become more similar after pruning. Thus, the $\bar{A}$ of an ensemble comprising pruned decision trees may be smaller than that of an ensemble comprising unpruned decision trees.

**Table 7.** Predictive error rate (%) of unpruned single C4.5 decision trees.

| Data set | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|
| *allbp* | 4.44±1.27 | 4.62±0.59 | 3.91±0.66 | 3.56±0.72 | 3.65±0.44 | 3.36±0.35 | 3.16±0.37 | 3.17±0.34 | 3.09±0.42 | 3.02±0.14 |
| *ann* | 1.00±0.40 | 0.73±0.26 | 0.52±0.16 | 0.51±0.14 | 0.47±0.10 | 0.40±0.09 | 0.40±0.07 | 0.38±0.05 | 0.33±0.05 | 0.27±0.04 |
| *block* | 4.50±1.04 | 4.00±0.26 | 3.69±0.54 | 3.77±0.27 | 3.34±0.20 | 3.33±0.24 | 3.24±0.20 | 3.26±0.24 | 3.23±0.13 | 3.24±0.09 |
| *hypothyroid* | 1.74±0.88 | 1.40±0.43 | 1.11±0.37 | 0.75±0.16 | 0.61±0.18 | 0.56±0.16 | 0.59±0.15 | 0.55±0.15 | 0.51±0.09 | 0.48±0.05 |
| *kr-vs-kp* | 4.41±1.16 | 2.51±0.56 | 1.65±0.42 | 1.34±0.37 | 1.16±0.19 | 1.00±0.21 | 0.91±0.14 | 0.77±0.12 | 0.71±0.10 | 0.60±0.08 |
| *led7* | 35.36±2.85 | 32.04±3.15 | 30.06±1.72 | 28.61±1.63 | 28.09±1.66 | 27.81±1.48 | 27.34±0.69 | 27.08±0.71 | 27.06±0.49 | 26.92±0.26 |
| *led24* | 40.29±7.11 | 38.42±2.67 | 36.95±2.76 | 36.96±1.32 | 36.58±2.23 | 35.88±1.52 | 35.89±1.04 | 35.19±0.75 | 35.23±0.76 | 34.41±0.63 |
| *sat* | 19.92±1.69 | 17.40±0.99 | 16.34±0.63 | 15.93±0.56 | 15.09±0.52 | 15.27±0.62 | 14.66±0.46 | 14.54±0.53 | 14.06±0.55 | 13.84±0.30 |
| *segment* | 9.92±1.89 | 7.60±1.17 | 6.43±0.76 | 6.02±0.43 | 5.23±0.52 | 4.55±0.62 | 4.11±0.31 | 3.78±0.34 | 3.44±0.31 | 3.17±0.17 |
| *sick* | 2.38±1.10 | 2.14±0.67 | 2.03±0.43 | 1.89±0.31 | 1.59±0.27 | 1.52±0.23 | 1.34±0.19 | 1.26±0.29 | 1.26±0.20 | 1.22±0.09 |
| *sick-euthyroid* | 3.84±1.09 | 3.42±0.97 | 3.22±0.87 | 2.96±0.68 | 2.80±0.20 | 2.90±0.21 | 2.78±0.44 | 2.66±0.15 | 2.53±0.15 | 2.39±0.14 |
| *waveform* | 11.27±1.71 | 10.37±1.58 | 10.36±1.02 | 9.77±0.88 | 9.99±0.61 | 9.99±0.54 | 9.80±0.49 | 9.47±0.44 | 9.21±0.44 | 9.02±0.32 |

In other words, in constituting an ensemble, the advantage of stronger generalization ability of pruned decision trees may be killed to some degree by its disadvantage of smaller ambiguity. Thus, although an ensemble comprising pruned decision trees may be stronger than that comprising unpruned decision trees, the gap between the former and the pruned single decision trees may not be so big as that between the latter and the unpruned single decision trees. Therefore, the benefit from ensemble of unpruned decision trees is usually bigger than that from ensemble of pruned decision trees.

## 4 Conclusion

In summary, the empirical study described in this paper discloses:

- Enlarging the training set tends to enlarge the benefit from Boosting but does not significantly impact the benefit from Bagging. This is because the increase of the training set size may enhance the bias reduction effect of adaptive sampling but may not significantly benefit the variance reduction effect of bootstrap sampling.

- The benefit from ensemble does not always increase along with the increase of the size of training set. This is because single learners sometimes may learn relatively more from randomly provided additional training data than ensembles do.
- The benefit from ensemble of unpruned decision trees is usually bigger than that from ensemble of pruned decision trees. This is because in constituting an ensemble, the relatively big ambiguity among the unpruned decision trees counteracts their relatively weak generalization ability to some degree.

These findings suggest that when dealing with huge volume of data, ensemble learning paradigms employing adaptive sampling are more promising, adequately selected training data are more helpful, and the generalization ability of the component learners could be sacrificed to some extent if this leads to a very significant increase of the ambiguity.

## Acknowledgement

This work was partially supported by the National Outstanding Youth Foundation of China under the Grant No. 60325207, the Excellent Young Teachers Program of MOE of China, the Fok Ying Tung Education Foundation under the Grant No. 91067, and the National 973 Fundamental Research Program of China under the Grant No. 2002CB312002.

## References

1. Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, Boosting, and variants. Machine Learning **36** (1999) 105–139
2. Blake, C., Keogh, E., 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, CA (1998)
3. Breiman, L.: Bagging predictors. Machine Learning **24** (1996) 123–140
4. Breiman, L.: Bias, variance, and arcing classifiers. Technical Report 460, Statistics Department, University of California, Berkeley, CA (1996)
5. Dietterich, T.G.: Machine learning research: four current directions. AI Magazine **18** (1997) 97–136
6. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Proceedings of the 2nd European Conference on Computational Learning Theory, Barcelona, Spain (1995) 23–37
7. German, S., Bienenstock, E., Doursat, R.: Neural networks and the bias/variance dilemma. Neural Computation **4** (1992) 1–58
8. Krogh, A., Vedelsby, J.: Neural network ensembles, cross validation, and active learning. In: Tesauro, G., Touretzky, D.S., Leen, T.K. (eds.): Advances in Neural Information Processing Systems, Vol. 7. MIT Press, Cambridge, MA (1995) 231–238
9. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, CA (2000)