An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, Boosting, and Randomization

Ana I. González Acuña

by Thomas G. Dietterich, *Machine Learning* (2000)

27/01/2012

Outline

Introduction
Methods
Experimental Results
Conclusions

Bagging
Boosting family
Randomization

Introduction

- Ensemble learning methods: a collection of individual classifiers.
- Construction: Base learning algorithm over different training sets.
- Techniques for constructing ensembles:
  - Bagging (bootstrap aggregation)
  - Boosting (Adaboost family)

Introduction
Methods
Experimental Results
Conclusions

Bagging
Boosting family
Randomization

Bagging

- Given a training set S of m examples, a new training set S' is constructed by drawing m examples uniformly (with replacement) from S.

- Bagging generates diverse classifiers only if the base learning algorithm is *unstable*— that is, if small changes to the training set cause large changes in the learned classifier.

Introduction
Methods
Experimental Results
Conclusions

Bagging
Boosting family
Randomization

Boosting family

- *Adaboost* algorithm: maintains a set of weights over the original training set S and adjusts these weights after each classifier is learned by the base learning algorithm:
    - increase the weight of examples that are misclassified.
    - decrease the weight of examples that are correctly classified.

- Methods to construct a new training set S':
    - *boosting by sampling*, examples are drawn with replacement from S with probability proportional to their weights.
    - *boosting by weighting*, the entire training set S (with associated weights) is given to the base learning algorithm, if it can accept a weighted training set directly.

- Adaboost requires less instability, because it can make much larger changes in the training set (large weights on few examples).

Introduction
Methods
Experimental Results
Conclusions

Bagging
Boosting family
Randomization

Randomization

- Proposition of an alternative method for constructing good ensembles that does not rely on instability.

- Idea: randomize the internal decisions of the learning algorithm.

- Modified version of the C4.5 (Release 1) learning algorithm in which the decision about which split to introduce at each internal node of the tree is randomized.

- Implementation: computes the 20 best splits (among those with non-negative information gain ratio) and then chooses uniformly randomly among them.

- For continuous attributes, each possible threshold is considered to be a distinct split, so the 20 best splits may all involve splitting on the same attribute.

Introduction
Methods
Experimental Results
Conclusions

Description
Pruning
Statistical tests

Methods

Methods:

- C4.5 Release 1 (alone),
- C4.5 with bagging,
- C4.5 with Adaboost.M1 (boosting by weighting), and
- Randomized C4.5.

Datasets:

- 33 domains drawn from the UCI Repository

Introduction
Methods
Experimental Results
Conclusions

Description
Pruning
Statistical tests

Methods

- Validation: train/test (3), stratified 10-fold cross-validation.
- Size ensembles:
  - Randomization and bagging: 200 classifiers
  - Boosting: at most 100 classifiers
- Iterations with convergence (reached the same accuracy as an ensemble of size 200) for most domains :
  - Randomization and bagging: 50 iterations
  - Boosting: 40 iterations

Introduction
Methods
Experimental Results
Conclusions

Description
Pruning
Statistical tests

Methods

- Pruning
  - Pruned and unpruned decision trees
  - Pruning confidence level 0.10
  - Test data to determine pruning

- Pruning difference:
  - Boosting: no significant difference in any of the 33 domains
  - C4.5 and randomized C4.5: significant difference in 10 domains
  - Bagged C4.5: significant differences in only 4 domains

- Does the lack of differences is due to low pruning confidence level?

Introduction
Methods
Experimental Results
Conclusions

Description
Pruning
Statistical tests

Methods

- Statistical tests to compare algorithm configurations:
  - in the 30 domains:
    - 10-fold cross-validated t test to construct a 95% confidence interval for the difference in the error rates of the algorithms
    - if the interval includes zero, there is not a difference in performance between the algorithms
  - in the 3 domains:
    - a single test that constructs a confidence interval based on the normal approximation to the binomial distribution

Introduction
Methods
Experimental Results
Conclusions

Error rates
Statistical tests
Classification noise
Diversity error diagrams

Experimental Results

Table 1. The 33 domains employed in this study.

Index	Name	C4.5		Randomized C4.5		Bagged C4.5		Adaboosted C4.5	
		P	Error rate	P	Error rate	P	Error rate	P	Error rate
1	sonar		0.3257 ± 0.0637		0.2018 ± 0.0545	*	0.2752 ± 0.0607	*	0.1651 ± 0.0505
2	letter		0.1225 ± 0.0045		0.0285 ± 0.0023		0.0552 ± 0.0032	*	0.0271 ± 0.0023
3	splice	*	0.0575 ± 0.0081	*	0.0397 ± 0.0068	*	0.0506 ± 0.0076		0.0503 ± 0.0076
4	segment		0.0328 ± 0.0073		0.0203 ± 0.0058		0.0263 ± 0.0065		0.0151 ± 0.0050
5	glass	*	0.3437 ± 0.0636		0.2277 ± 0.0562		0.2723 ± 0.0596	*	0.2277 ± 0.0562
6	soybean		0.1262 ± 0.0371	*	0.0852 ± 0.0312	*	0.1009 ± 0.0337	*	0.0757 ± 0.0296
7	autos		0.2326 ± 0.0578	*	0.1581 ± 0.0499		0.1814 ± 0.0528		0.1814 ± 0.0528
8	satimage	*	0.1515 ± 0.0157		0.0890 ± 0.0125		0.1020 ± 0.0133		0.0850 ± 0.0122
9	annealing	*	0.0132 ± 0.0075		0.0088 ± 0.0061		0.0099 ± 0.0065		0.0055 ± 0.0048
10	krk		0.1887 ± 0.0046		0.1309 ± 0.0039		0.1463 ± 0.0041	*	0.1026 ± 0.0036

Error rate $\pm$ 95% confidence limit.
Error rate estimated by 10-fold cross validation (except 8, 14, 21)
P * –> pruned trees

Introduction
Methods
Experimental Results
Conclusions

Error rates
Statistical tests
Classification noise
Diversity error diagrams
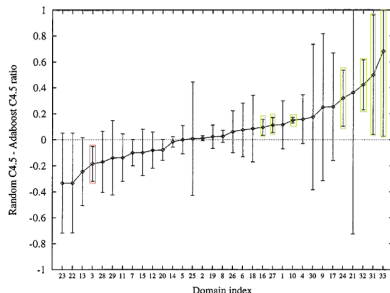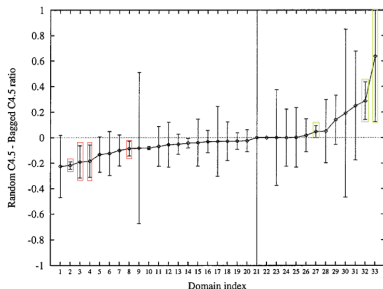
Experimental Results

Results of statistical tests:

- All three ensemble methods do well against C4.5 alone: Randomized C4.5 is better in 14 domains, Bagged C4.5 is better in 11, and Adaboosted C4.5 is better in 17.

- C4.5 is never able to do better than any of the ensemble methods.

Table 2. All pairwise combinations of the four ensemble methods. Each cell contains the number of wins, losses, and ties between the algorithm in that row and the algorithm in that column.

	C4.5	Adaboost C4.5	Bagged C4.5
Random C4.5	14 - 0 - 19	1 - 7 - 25	6 - 3 - 24
Bagged C4.5	11 - 0 - 22	1 - 8 - 24	
Adaboost C4.5	17 - 0 - 16		

Introduction
Methods
Experimental Results
Conclusions

Error rates
Statistical tests
Classification noise
Diversity error diagrams

Experimental Results

"Kohavi" plots:

Each point plots the difference in the performance that is scaled by the error rate of C4.5 alone.

Error bars give a 95% confidence interval according to the cross-validated t test.

Introduction
Methods
Experimental Results
Conclusions

Error rates
Statistical tests
Classification noise
Diversity error diagrams

Classification noise

- How well these ensemble methods perform in situations where there is a large amount of classification noise (i.e., training and test examples with incorrect class labels)?
  Some previous experiments demonstrate the poor performance of Adaboosted C4.5 and Randomized against classification noise, but are applied over small ensembles.

- Larger ensembles can be able to overcome the effects of noise?

Introduction
Methods
Experimental Results
Conclusions

Error rates
Statistical tests
Classification noise
Diversity error diagrams

Classification noise

- Effect of classification noise:
  Add random class noise to 9 domains (present statistically significantly different performance)

- To add classification noise at a given rate r :
  Choose a fraction r of the data points (randomly, without replacement) and change their class labels to be incorrect (the label for each example was chosen uniformly randomly from the incorrect labels).

- The data were split into 10 subsets for the stratified 10-fold cross-validation (the stratification was performed using the new labels).

Introduction
Methods
Experimental Results
Conclusions

Error rates
Statistical tests
Classification noise
Diversity error diagrams

Classification noise

	C4.5	Adaboost C4.5	Bagged C4.5
Noise = 0%			
Random C4.5	5 - 0 - 4	1 - 6 - 2	3 - 3 - 3
Bagged C4.5	4 - 0 - 5	0 - 5 - 4	
Adaboost C4.5	6 - 0 - 3		
Noise = 5%			
Random C4.5	5 - 2 - 2	3 - 2 - 4	1 - 5 - 3
Bagged C4.5	6 - 0 - 3	5 - 1 - 3	
Adaboost C4.5	3 - 3 - 3		
Noise = 10%			
Random C4.5	4 - 1 - 4	5 - 1 - 3	1 - 6 - 2
Bagged C4.5	5 - 0 - 4	6 - 1 - 2	
Adaboost C4.5	2 - 3 - 4		
Noise = 20%			
Random C4.5	5 - 2 - 2	5 - 0 - 4	0 - 2 - 7
Bagged C4.5	7 - 0 - 2	6 - 0 - 3	
Adaboost C4.5	3 - 6 - 0		

Table 3. Shows the win-lose-tie counts for all pairs of learning methods at the four noise levels (0%, 5%, 10%, and 20%) and 9 domains.

Introduction
Methods
Experimental Results
Conclusions

Error rates
Statistical tests
Classification noise
Diversity error diagrams

Classification noise

- Confirmation of previous works:
  Adding noise to these problems, Randomized C4.5 and
  Adaboosted C4.5 lose some of their advantage over C4.5 while
  Bagged C4.5 gains advantage over C4.5.

- Conclusion:
  The best method in applications with large amounts of
  classification noise is Bagged C4.5, with Randomized C4.5
  behaving almost as well.
  In contrast, Adaboost is not a good choice in such
  applications.

Introduction
Methods
Experimental Results
Conclusions

Error rates
Statistical tests
Classification noise
Diversity error diagrams

κ-error diagrams

- Scatter plot in which each point corresponds to a pair of classifiers. Its x coordinate is the diversity value (κ) and its y coordinate is the mean accuracy of the classifiers.

- The κ statistic is defined as follows:

$$\kappa = \frac{\Theta_1 - \Theta_2}{1 - \Theta_2}$$

  - $\kappa = 0$ when the agreement of the two classifiers equals that expected by chance.
  - $\kappa = 1$ when the two classifiers agree on every example.
  - $\kappa < 0$ when there is systematic disagreement between the classifiers.

Introduction
Methods
Experimental Results
Conclusions

Error rates
Statistical tests
Classification noise
Diversity error diagrams

κ-error diagrams

Θ_1 is an estimate of the probability that the two classifiers agree.
Θ_2 is an estimate of the probability that the two classifiers agree by chance.

$$\Theta_1 = \frac{\sum_{i=1}^{L} C_{ii}}{m} \qquad \Theta_2 = \sum_{i=1}^{L} \left(\sum_{j=1}^{L} \frac{C_{ij}}{m} \cdot \sum_{j=1}^{L} \frac{C_{ji}}{m} \right)$$

where m is the total number of test examples, L classes, and C be an $L \times L$ square array such that C_{ij} contains the number of test examples assigned to class i by the first classifier and into class j by the second classifier.

Introduction
Methods
Experimental Results
Conclusions

Error rates
Statistical tests
Classification noise
Diversity error diagrams

κ-error diagrams

κ-error diagrams for the sick data set using Bagged C4.5 (a), Randomized C4.5 (b), and Adaboosted C4.5 (c). Accuracy and diversity increase as the points come near the origin.

Introduction
Methods
Experimental Results
Conclusions

Error rates
Statistical tests
Classification noise
Diversity error diagrams
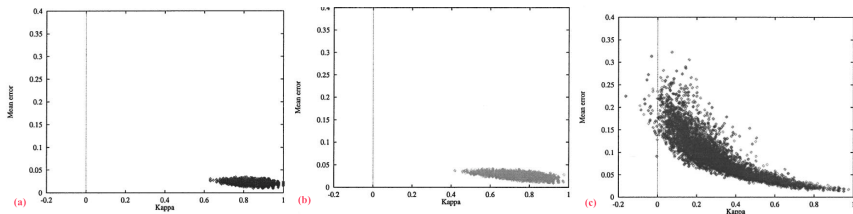
κ-error diagrams

κ-error diagrams for the sick data set with 20% random classification noise using Bagged C4.5 (a), Randomized C4.5 (b), and Adaboosted C4.5 (c).

Introduction
Methods
Experimental Results
Conclusions

Error rates
Statistical tests
Classification noise
Diversity error diagrams
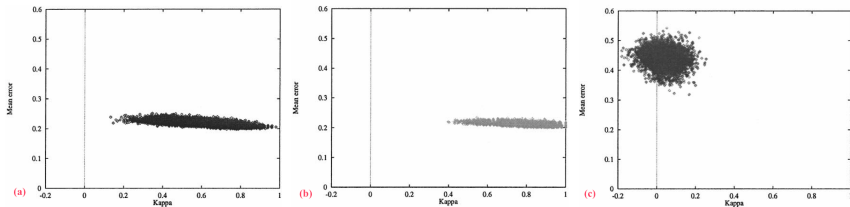
Adaboost behaviour

- Hypothesis: Adaboost is placing more weight on the noisy examples
- Test: Mean weight per training example for the 560 corrupted training examples and the remaining 2,240 uncorrupted training examples in the sick data set.
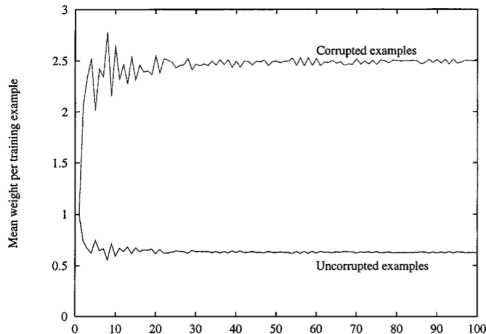
Conclusions

- Proposition of a new method for constructing ensemble classifiers using C4.5
- Without classification noise:
  - Boosting gives the best results in most cases
  - Randomizing and Bagging give quite similar results
- With added classification noise:
  - Bagging is the best method.
  - Randomized C4.5 is not as good as Bagging.

📄 Dietterich, T.G. (2000) *An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization*, Machine Learning, 40, pp 139-157.

📄 Rokach, L. (2010) *Pattern Recognition using Ensemble methods,* Series in Machine Perception and Artificial Intelligence - Vol 75. World Scientific Publishing.

📄 Duda, R.O., Hart, P.E. and Stork, D.G. (2001), *Pattern Classification* (ch8), 2nd edition, John Wiley & Sons