Bagging, Boosting, and C4.5

J.R. Quinlan

University of Sydney

Sydney, Australia 2006

quinlan@cs.su.oz.au

Report by:

Bassem Fadlia

Bagging, Boosting, and C4.5

C4.5

A training set $\xrightarrow{\text{Decision tree algorithm (C4.5)}}$ Classifier C(x)

Bagging, Boosting, and C4.5

C4.5

A training set $\xrightarrow{\text{Decision tree algorithm (C4.5)}}$ Classifier C(x)

Bagging and Boosting

Altered training set $\xrightarrow{\text{Decision tree algorithm (C4.5)}}$ Classifier $C^1(x)$

Altered training set $\xrightarrow{\text{Same decision tree algorithm (C4.5)}}$ Classifier $C^2(x)$

Altered training set $\xrightarrow{\text{Same decision tree algorithm (C4.5)}}$ Classifier $C^t(x)$

Aggregation of the t classifiers: $\longrightarrow$ Classifier $C^*(x)$

	Windy	Outlook	Temperature	Play
100	True	Sunny	90	Play
200	False	Sunny	80	Play
300	True	Overcast	65	Don't play
400	False	Rain	95	Don't Play
500	False	Sunny	70	Play
600	False	Rain	70	Don't Play
700	True	Overcast	75	Play
800	False	Sunny	95	Play

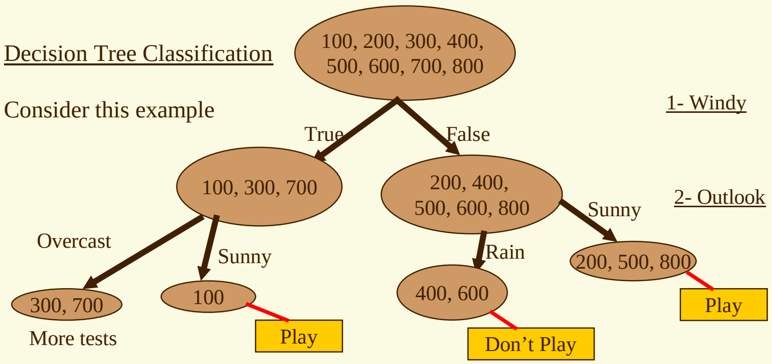
Decision Tree Classification

Consider this example

	Windy	Outlook	Temperature	Play
100	True	Sunny	90	Play
200	False	Sunny	80	Play
300	True	Overcast	65	Don't play
400	False	Rain	95	Don't Play
500	False	Sunny	70	Play
600	False	Rain	70	Don't Play
700	True	Overcast	75	Play
800	False	Sunny	95	Play

Decision Tree Classification

Consider this example



Attribute Selection

- Choose the most informative attribute first
- Entropy is one measure of how informative the attribute is

Entropy I(P) = - (p1 * log(p1) + p2 * log(p2)+ ..+ pn * log(pn))

$$\text{Info}(X,T) = \sum_{I=1..n} (T_i / T)) * \text{Info}(T_i)$$

Attribute Selection

- Choose the most informative attribute first
- Entropy is one measure of how informative the attribute is

Entropy I(P) = - (p1 * log(p1) + p2 * log(p2)+ ..+ pn * log(pn))

$$\text{Info}(X,T) = \sum_{I=1..n} (T_i / T)) * \text{Info}(T_i)$$

Id	100	200	300	400	500	600	700	800
Windy	True	False	True	False	False	False	True	False
Play	Play	Play	Don't Play	Don't Play	Play	Don't Play	Play	Play

Info (Windy, T) = ?

Attribute Selection

- Choose the most informative attribute first
- Entropy is one measure of how informative the attribute is

Entropy I(P) = - (p1 * log(p1) + p2 * log(p2)+ ..+ pn * log(pn))

$$\text{Info}(X,T) = \sum_{I=1..n} (T_i / T)) * \text{Info}(T_i)$$

Id	100	200	300	400	500	600	700	800
Windy	True	False	True	False	False	False	True	False
Play	Play	Play	Don't Play	Don't Play	Play	Don't Play	Play	Play

Info (Windy, T) = 3/8 * I(2/3, 1/3) + 5/8 * I(3/5, 2/5)
= 0.918

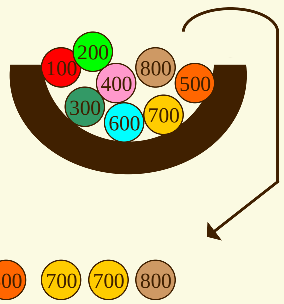
Bagging

•Unordered with replacement sampling.

•An instance can appear more than once, while another doesn't appear in the sample.

•Each sample is independent of previous samples and previous classifier results.

•Each classifier has an equal standing in the final vote



Boosting

• Weights are given to the instances

• Adjust the weights each time to give more attention to misclassified instances.

•Usually performs better than Bagging, but more risky.

•Different ways to incorporate the weights in the algorithm.

Freund and Schapire's method for introducing weights

Id	100	200	300	400	500	600	700	800
Windy	True	False	True	False	False	False	True	False
Play	Play	Play	Don't Play	Don't Play	Play	Don't Play	Play	Play
Weight	1	3	2	1	1	1	2	1

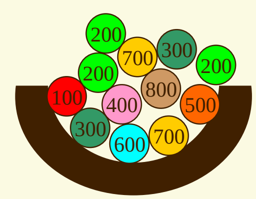
Fruend and Schapire's method for introducing weights

Id	100	200	300	400	500	600	700	800
Windy	True	False	True	False	False	False	True	False
Play	Play	Play	Don't Play	Don't Play	Play	Don't Play	Play	Play
Weight	1	3	2	1	1	1	2	1

• They incorporate the weight in a manner analogous to bagging.

•Still, an instance can appear more than once, while another doesn't appear in the sample.

•Doesn't benefit from the major advantage of boosting over bagging. And gives misleading results that bagging is competitive to boosting.



Quinlan's method of introducing weights

Id	100	200	300	400	500	600	700	800
Windy	True	False	True	False	False	False	True	False
Play	Play	Play	Don't Play	Don't Play	Play	Don't Play	Play	Play
Weight	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

Info (Windy, T) = 3/8 * I(2/3, 1/3) + 5/8 * I(3/5, 2/5)

Quinlan's method of introducing weights

Id	100	200	300	400	500	600	700	800
Windy	True	False	True	False	False	False	True	False
Play	Play	Play	Don't Play	Don't Play	Play	Don't Play	Play	Play
Weight	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

Info (Windy, T) = 3/8 * I(2/3, 1/3) + 5/8 * I(3/5, 2/5)

Info (Windy, T) = $(1/8_{100}+1/8_{300}+1/8_{700}) * I\left(\frac{1/8+1/8}{1/8+1/8+1/8}, \frac{1/8}{1/8+1/8+1/8}\right)$

$+\ (1/8+1/8+1/8+1/8+1/8) * I\left(\frac{1/8+1/8+1/8}{1/8+1/8+1/8+1/8+1/8}, \frac{1/8+1/8}{1/8+1/8+1/8+1/8+1/8}\right)$

Quinlan's method of introducing weights

Id	100	200	300	400	500	600	700	800
Windy	True	False	True	False	False	False	True	False
Play	Play	Play	Don't Play	Don't Play	Play	Don't Play	Play	Play
Weight	W1	W2	W3	W4	W5	W6	W7	W8

Info (Windy, T) = 3/8 * I(2/3, 1/3) + 5/8 * I(3/5, 2/5)

Info (Windy, T) = $(w1+w3+w7) * I\left(\frac{w1+w7}{w1+w3+w7}, \frac{w3}{w1+w3+w7}\right)$

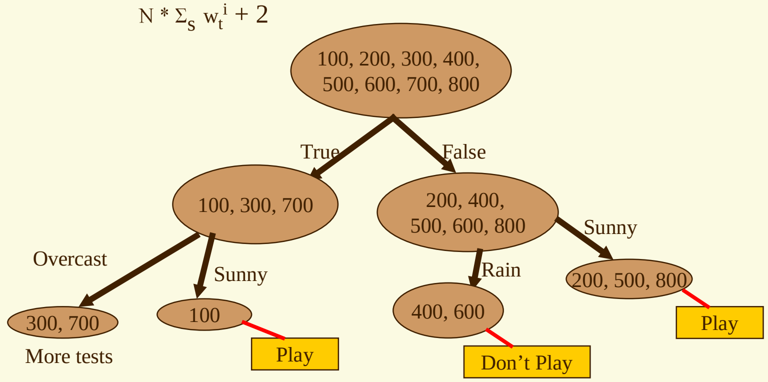
$+\ (w2+w4+w5+w6+w8) * I\left(\frac{w2+w5+w8}{w2+w4+w5+w6+w8}, \frac{w4+w6}{w2+w4+w5+w6+w8}\right)$

Boosting, adjusting the weights

- Initially, $w_x^1 = 1/N$
- Multiply the weights of correctly classified instances by $\beta^t = \varepsilon^t / 1- \varepsilon^t$
- Divide by normalization constant
- The worth of each classifier's vote depends on its accuracy
  - $\log 1/\beta^t$
  - $H^t(x)$

Boosting, adjusting the weights

$$H^t(x) = \frac{N * \Sigma_{sk}\ w_t^i + 1}{N * \Sigma_s\ w_t^i + 2}$$



Bagging, Boosting, and C4.5

- Requirements for Boosting and Bagging
- Experiments
- Conclusion

Requirement 1: Instability

- Small changes to the training set should lead to different classifiers.
- Quinlan reports "The vital element is the instability of the prediction method. If perturbing the learning set can cause significant changes in the predictor constructed, accuracy is improved"
- Breiman 1994 "Bagging goes a ways toward making a silk purse out of a sow's ear, especially if the sow's ear is twitchy"

Requirement 2: Classifier should not be poor

- A poor learner is one that does not perform better than random guessing.
- Quinlan requires that the predictor's error on the given distribution should be kept below 50% (Binary Classifier, K =2)
- Aggregating weak learners produces a strong learner. Aggregating poor learners produces even more poor learners

Experiments

Settings

• C4.5, Bagged C4.5 and Boosted C4.5 have been evaluated on a collection of 27 datasets from the UCI learning repository

• Parameter T was set at 10 for these experiments.

Experiments

Settings

• C4.5, Bagged C4.5 and Boosted C4.5 have been evaluated on a collection of 27 datasets from the UCI learning repository

• Parameter T was set at 10 for these experiments.

Results

• Bagging reduces C4.5's error by 10% and is superior to C4.5 on 24 of the 27 datasets.

• Boosting reduces C4.5's error by 15% but it is superior on only 21 datasets.

Experiments

Settings

• C4.5, Bagged C4.5 and Boosted C4.5 have been evaluated on a collection of 27 datasets from the UCI learning repository

• Parameter T was set at 10 for these experiments.

Results

• Bagging reduces C4.5's error by 10% and is superior to C4.5 on 24 of the 27 datasets.

• Boosting reduces C4.5's error by 15% but it is superior on only 21 datasets.

• Comparing bagging and boosting, boosting leads to higher reduction of error and is superior on 20 out of the 27 datasets, but it is more risky.

Experiments

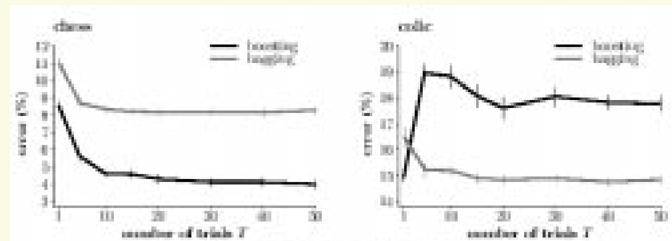


Figure 1: Comparison of bagging and boosting on two datasets

Experiments

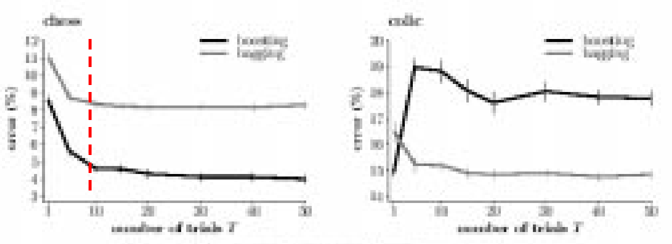


Figure 1: Comparison of bagging and boosting on two datasets

• If you choose very big T, you just cost yourself more computation with no improvement in the classification error.

• Breiman describes this as "love's labor lost"

Conclusion

• Boosting and Bagging both require T times the computation time of C4.5

• A 10-fold increase in computation buys an average reduction of between 10% and 19% of the classification error.

• Boosting seems to be more effective than bagging when applied to C4.5, although the performance of the bagged C4.5 is less variable.

• Fruend and Schapire attribute Boosting failure to overfitting. Quinlan thinks this hypothesis is insufficient and calls for more investigation.

END