

A Comparison of Ensemble Creation Techniques

The Fifth International Conference on Multiple Classifier
Systems
Cagliari, Italy

**Robert Banfield, Larry Hall, Kevin Bowyer,
Divya Bhadoria, Philip Kegelmeyer, Steven Eschrich**

<http://morden.csee.usf.edu/avatar/>

June 9, 2004

Bagging Vs. Everything Else

The accuracy of seven ensemble methods is compared; bagging, boosting, and several later algorithms.

When statistical significance is carefully computed ...

none of the methods, including Adaboost, are generally statistically significantly more accurate than bagging.

(No, we weren't expecting this either.)

The Ensemble of Ensemble Algorithms

Bagging: Bag size set equal to data set size.

Boosting: Adaboost.M1W, default parameters.

Random Subspaces (RS): $N/2$ attributes in each tree.

Random Trees B (RTB): Use random test out of 20 best tests at each node.

Random Forests: n attributes considered at each node.
Bag size set equal to data set size.

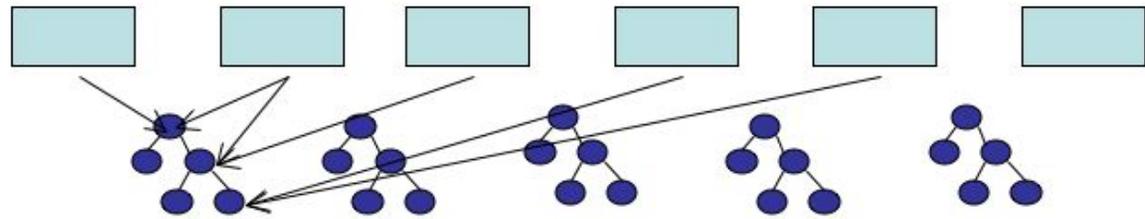
Random Forests-lg (RF-lg): $n = \log_2(N) + 1$

Random Forests-1 (RF-1): $n = 1$

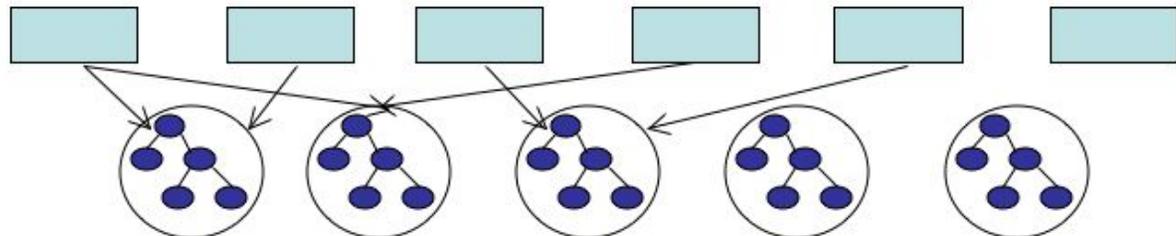
Random Forests-2 (RF-2): $n = 2$

Ensemble Method Distinctions

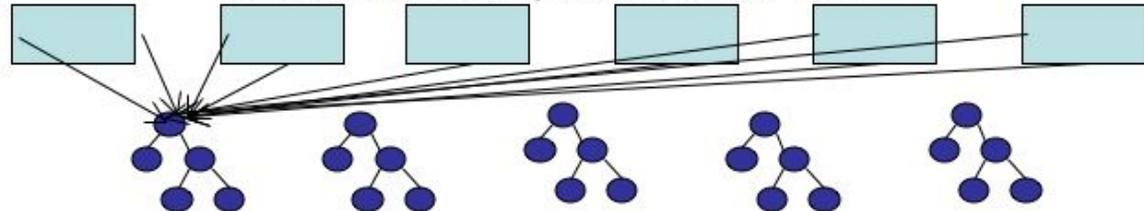
Random Forests



Random Subsets



Random Trees B; Randomized C4.5



Thirty Four Data Sets

anneal	glass	hypo	lymph	sick
audiology	heart-c	ion	page	sonar
autos	heart-h	iris	pendigits	soybean
breast-w	heart-s	krkp	phoneme	vehicle
breast-y	heart-v	labor	pima	voting
credit-a	hepatitis	led-24	primary	waveform
credit-g	horse-colic	letter	satimage	

Experimental Details

Pruning: All trees unpruned.

Size of Ensembles: *Boosting:* 50 trees. *Others:* 1000 trees.

Ensemble Performance: 10-fold cross ensemble validation.

As an example, pendigits has around 10,000 samples.

- Divide the 10,000 into 10 bins of 1000.
- Set aside one bin, merge the rest into 9000 samples.
- Grow an ensemble from the 9000 samples.
- Determine its accuracy on set aside 1000 samples.
- Repeat for all 10 bins.

Result: 10 measurements of accuracy for an ensemble method.

Simple But Misleading Comparison

Score Average Accuracy: 1 point for beating bagging,
score 0.5 points for a tie:

RF-lg	RF-2	RS	Boosting	RTB	RF-1	Bagging
24.5	23.0	21.5	19	19	18.5	17

Borda Count: B points for 1st place, $B - 1$ for
2nd, etc, where B is the number of methods (here, $B = 7$):

RF-lg	RF-2	RTB	RS	Boosting	Bagging	RF-1
167	166	152	150	128	118	117

Tentative Conclusions:

- RF-lg and RF-2 consistently lead the pack.
- Bagging is consistently outperformed by everything.

Statistical Details

Point: Determine if algorithm X has a *statistically significant* improvement over bagging.

ANOVA: First, use analysis of variance to determine the $p \leq 34$ datasets for which average accuracies are significantly different across the ensemble methods,

t-test: Then use a paired t-test on just those p datasets, at the 99% confidence level, to determine statistical significance of differences in accuracy.

Why 99%? 34 data sets, 6 comparisons each, so 204 statistical tests. 95% confidence would be 10 errors! Even 99% means 2 errors.

Accurate But Surprising Comparison

For 30 of 34 datasets, nothing was significantly better than bagging. (Though some were significantly worse!)

Statistically significant comparisons against bagging:

	Wins	Losses	Ties
Random Forests-lg	4	1	29
Random Forest-2	4	2	28
Random Trees B	3	1	30
Random Forests-1	3	2	29
Random Subspaces	3	4	27
Boosting	2	1	31

Significant Accuracy Comparisons

Only 7 of 34 datasets had any significant differences.

Data set	Boost	RS	RTB	Bag	RF-lg	RF-1	RF-2
krkp	99.56	95.75	98.72	99.66	99.47	97.94	99.13
led-24	71.43	69.44	72.41	73.57	74.93	74.27	74.77
letter	96.74	97.03	96.44	94.90	96.84	95.66	96.81
pendigits	99.21	99.30	99.25	98.59	99.25	99.02	99.14
phoneme	91.46	83.70	90.37	91.42	91.26	91.02	91.35
sick	98.91	96.29	98.86	98.94	98.49	97.96	98.17
waveform	84.21	85.27	85.55	84.01	85.01	85.59	85.41

Blue: better than Bagging, Red: worse than Bagging

Data and Algorithm Observations

- In 2 of the 7 significant data sets, letter and pendigits, *everything* was statistically superior to bagging. Why?
 - The two largest data sets?
 - Of datasets with all continuous attributes, they had the most classes?
- Random forests is faster than bagging.
- Random subspaces requires half as much memory as bagging.

Bagging Vs. Everything Else

The accuracy of seven ensemble methods is compared; bagging, boosting, and several later algorithms.

When statistical significance is carefully computed ...

none of the methods, including Adaboost, are generally statistically significantly more accurate than bagging.

But Random Forests is no worse, and is faster.

End of Talk

Supplemental Slides Follow

Data Set	A	CA	E	C	Data Set	A	CA	E	C
anneal	38	6	898	6	krkp	36	0	3196	2
audiology	69	0	226	24	labor	16	8	57	2
autos	25	15	205	7	led-24	24	0	5000	10
breast-w	9	9	699	2	letter	16	16	20000	26
breast-y	9	0	286	2	lymph	18	3	148	4
credit-a	15	6	690	2	page	10	10	5473	5
credit-g	20	7	1000	2	pendigits	16	16	10992	10
glass	9	9	214	7	phoneme	5	5	5404	2
heart-c	13	5	303	2	pima	8	8	768	2
heart-h	13	5	294	2	primary	17	0	339	22
heart-s	13	5	123	2	satimage	36	36	6435	7
heart-v	13	5	200	2	sick	29	7	3772	2
hepatitis	19	6	155	2	sonar	60	60	208	2
horse-colic	22	8	368	2	soybean	35	0	683	19
hypo	25	7	3163	2	vehicle	18	18	846	4
ion	34	34	351	2	voting	15	0	435	2
iris	4	4	150	3	waveform	21	21	5000	3

A: # attributes. CA: # continuous attributes. E: # examples. C: # classes.