Bagging

Abhishek Dubey
Vinod Mamtani
Robert Field

Outline

- What is Bagging
- Brieman's Algorithm
- Datasets used by Brieman
- Why bagging works
- How many bootstrap replicates are enough
- How big should the Bootstrap learning set be
- Pros and Cons of Bagging
- Our tests and results

What is Bagging

Acronym for **B**ootstrap **Agg**regat**ing**

Bootstrap
- Form learning sets by sampling with replacement
- New sets form bootstrap distribution that approximates the underlying learning set distribution

Form a predictor for each bootstrap learning set

Aggregating
- Combine (average or vote) predictors trained by bootstrap learning sets

Expected error is reduced; more reduction for unstable predictors

Brieman's Algorithm

1. Dataset is randomly divided into a learning set L and test set T. T is 10% of data.
2. A classification tree is constructed from L using 10 fold cross-validation.
3. A bootstrap sample L_b is selected from L to grow a new tree. This tree is pruned using L. This is repeated 50 times.
4. Class with the most votes is selected as the class label. If there is a tie, select the class with the lowest label. Take an average for numeric response.
5. Compute misclassification rates for steps 2 and 4
6. Repeat steps 1-4 100 times

Datasets used in the paper

Classification
Waveform, Heart, Breast cancer, Ionosphere, Diabetes, Glass, Soybean, Letters, Satellite, Shuttle, DNA

Reduction in misclassification rates ranges from 6% to 77%

Regression
Boston Housing, Ozone, Friedman # 1, 2 and 3

Reduction in mean squared error ranges from 21% to 46%

Why Bagging Works

The total expected error of a classifier is the sum of bias and variance. Aggregating multiple classifiers reduces the variance thereby decreasing the expected error.

For regression, the mean-squared error of the aggregator is lower than the mean-squared error averaged over the training set.

How many bootstrap replicates are enough

Fewer replicates for regression than classification

More replicates are required with increasing number of classes

Computation time may be a factor
- Neural networks progress much more slowly than CART

How big should the Bootstrap Learning Set be

Same as the learning set

Sampling with replacement leaves out approximates 0.37 of the instances

Bootstrap learning sets that are twice as the origianl learning set did not improve accuracy

Pros and Cons of Bagging

Pros

- Works both on numerical and classification data
- Better results for unstable models
- Highly parallel operation
- Known to improve results for any algorithm

Cons

- Error rates may go up with stable models

Conclusion

Bagging works well for methods with high variance such as neural networks and tree based methods

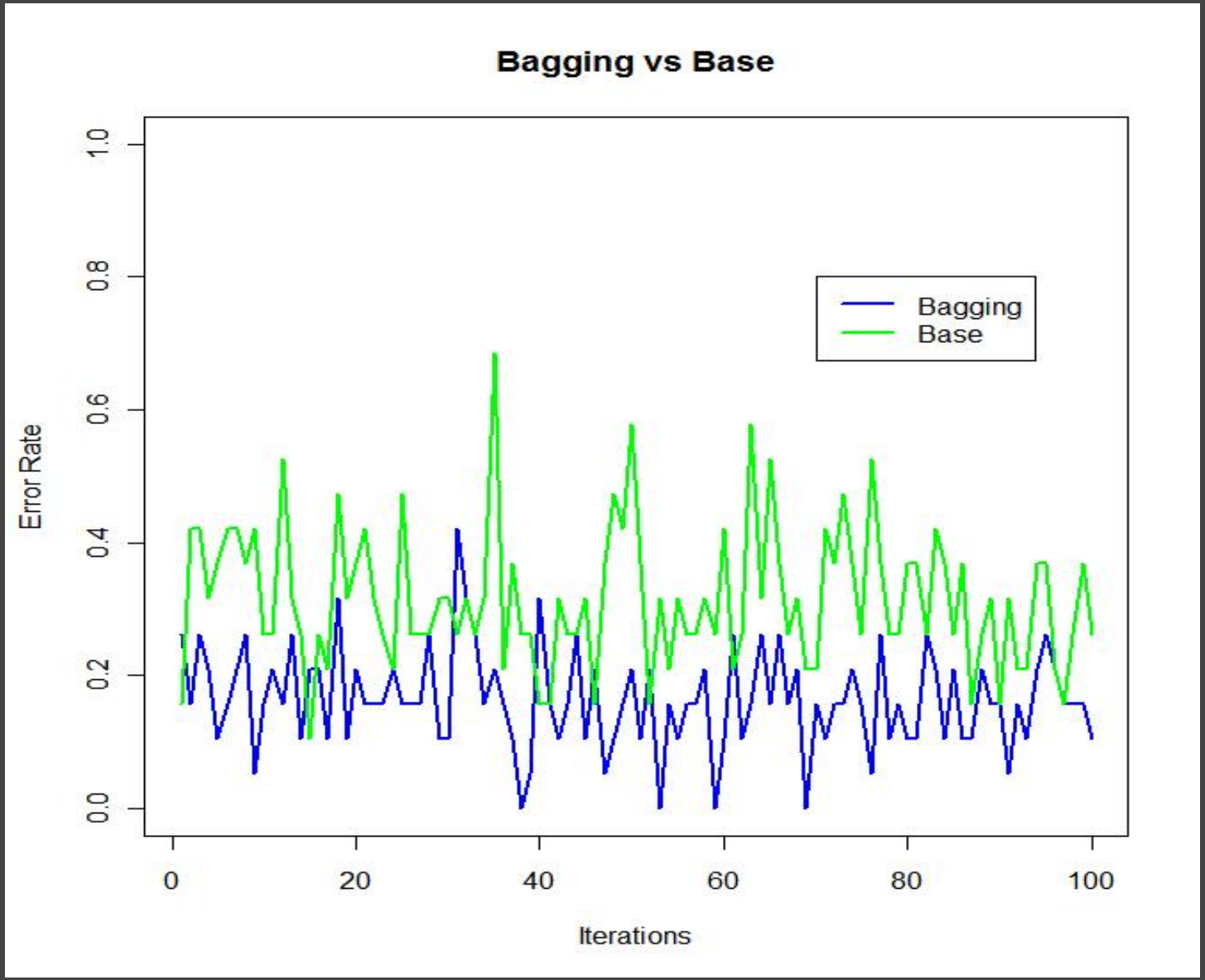
Bagging does not improve the performance of methods with low variance such as linear regression

Unlinke numerical predictors poor classification predictors can be transformed into worse ones.

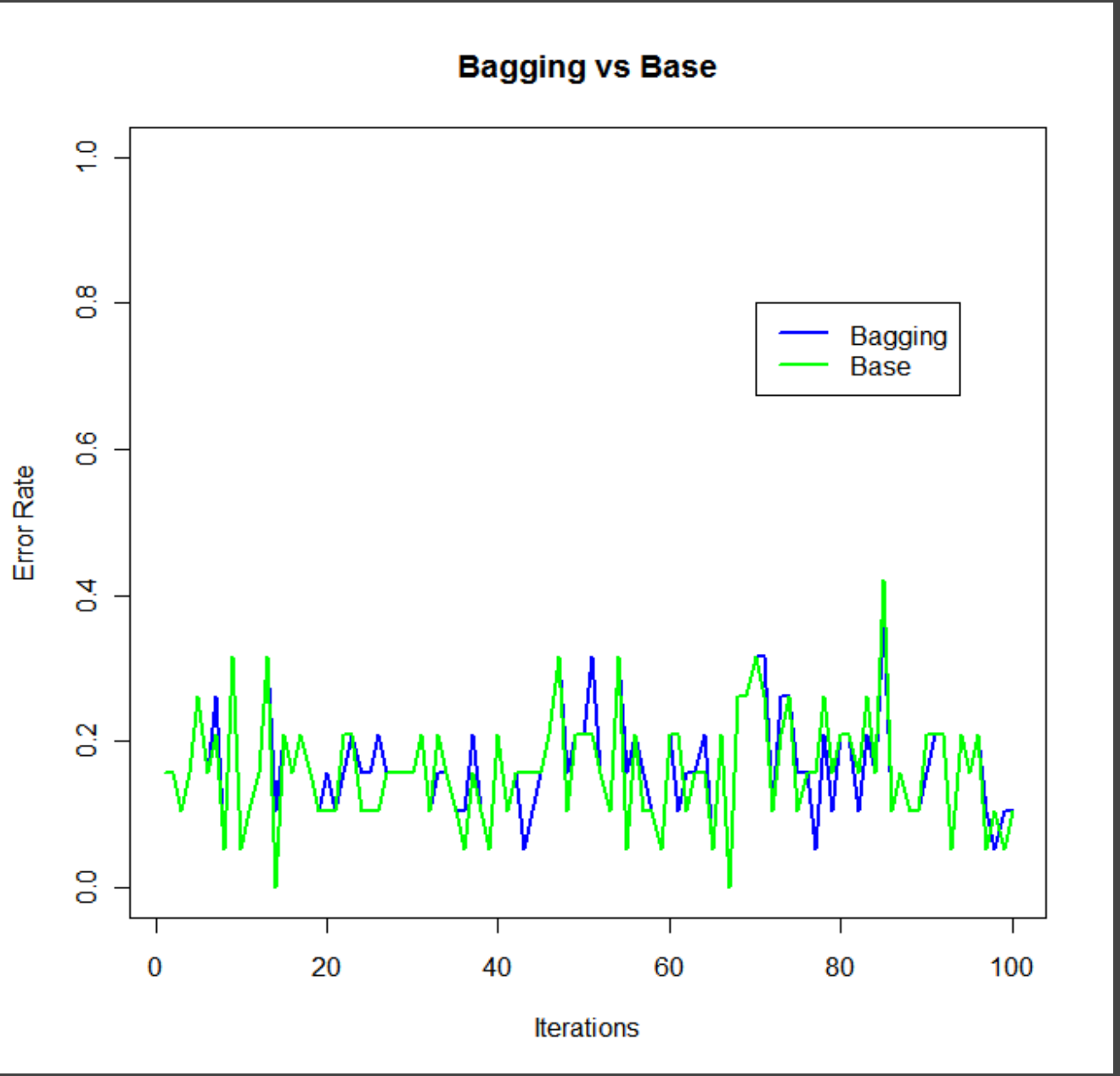
Our Implementation of Bagging Algo

- We implemented Brieman's Bagging Algorithm
- We tested it on Parkinson's Dataset
- We ran our implementation over 100 and 1000 iterations

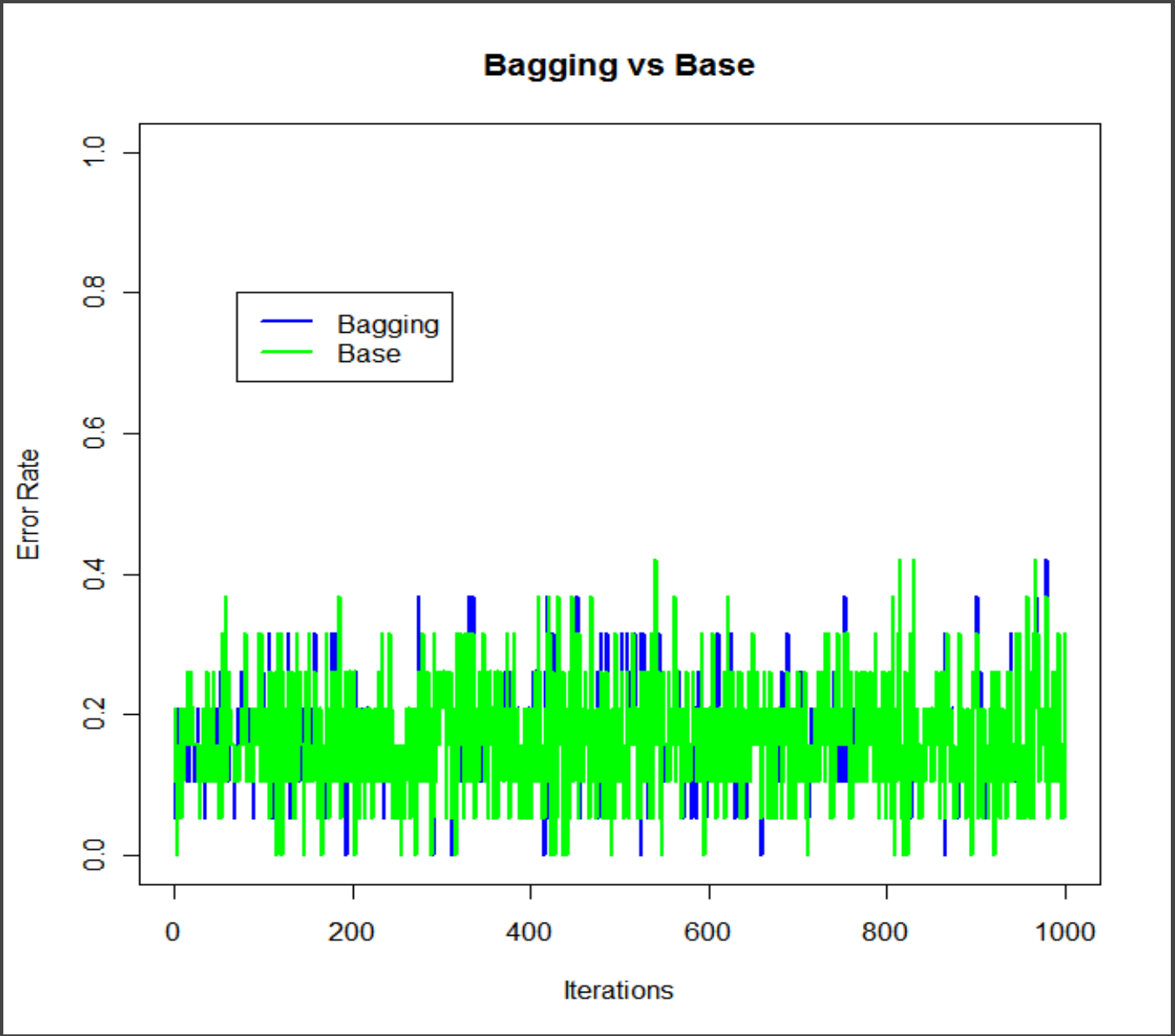
Results we expected



Results over 100 iterations



Results over 1000 iterations



Overall results over 1000 iterations

- Bagging did well 15.7% of the time
- Base predictor did well 15.7% of the time
- Prediction was same for both models 68.6% of them time

Demo Time

Backup - Bootstrap Distribution

Bootstrap distributions usually approximate the shape, spread and bias of the actual sampling distribution

Bootstrap distributions are centered at the value of the statistics from the original sample plus any bias

The sampling distribution is centered at the value of the parameter in the population plus any bias