Strength, Correlation, and Random Features in Random Forests

Zachary Weller

Department of Statistics, Colorado State University
wellerz@stat.colostate.edu

October 23, 2014

Leo Breiman

- Was a faculty member at UC-Berkeley
- Credited with the development of bagging and the random forests algorithms
- Main author of “Classification and Regression Trees” book: cited 25,643 times on Google Scholar (10/20/14)
- “Random Forests” paper: cited 13,772 times on Google Scholar (10/20/14)

Reminder: bagging

Main idea: reduce variance by inducing some bias and averaging over many trees.

We noted that variance reduction is limited by correlation between the random variables (trees).

However, introducing too much bias (e.g. trees that are poor classifiers) creates poor predictions.

Question: how can we create trees that have low correlation, but still exhibit high strength (low bias)?

Random Forests

Definition (random forests)

A **random forest** is a classifier consisting of a collection of tree-structured classifiers, $\{h(\mathbf{x}, \Theta_k), k = 1, \ldots\}$ where the $\{\Theta_k\}$ are independent, identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\mathbf{x}$.

note: modifying the mechanism Θ_k produces different types of random forests, e.g. bagging

Outline:

1. Characterizing random forests: strength and correlation
2. Using random features
   - random feature selection (choice of m)
   - linear combinations of features
3. Some empirical results

Characterizing random forests

Definition (margin function)

The **margin function** for a random forest is:
$mr(\mathbf{X}, Y) = \mathbf{P}_{\Theta}(h(\mathbf{X}, \Theta) = Y) - max_{j \neq Y} \mathbf{P}_{\Theta}(h(\mathbf{X}, \Theta) = j)$

Given an ensemble of classifiers $h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_k(\mathbf{x})$ with training data drawn randomly from $(\mathbf{X}, Y)$, we can estimate the margin function via:

$$\widehat{mr}(\mathbf{X}, Y) = av_k I(h_k(\mathbf{X}) = Y) - max_{j \neq Y} av_k I(h_k(\mathbf{X}) = j) \tag{1}$$

Breimen: “The margin measures the extent to which the average number of votes at $(\mathbf{X}, Y)$ for the right class exceeds the average vote for any other class. The larger [positive] the margin, the more confidence in the classification.”

Characterizing random forests

Definition (generalization error)

For an ensemble of classifiers,the **generalization error**, GE^*, is:
$GE^* = \mathbf{P}_{\mathbf{X},Y}(\widehat{mr}(\mathbf{X}, Y) < 0)$

Theorem (SLLN for RFs)

As the number of trees increases ($k \to \infty$), for almost surely all sequences $\Theta_1, \Theta_2, \ldots$; $GE^* = \mathbf{P}_{\mathbf{X},Y}(\widehat{mr}(\mathbf{X}, Y) < 0) \to \mathbf{P}_{\mathbf{X},Y}(mr(\mathbf{X}, Y) < 0)$

Proof: see paper

Breiman:

- "This result explains why random forests do not overfit as more trees are added, but produce a limiting value of the generalization error."
- We can derive an upper bound for the generalization error in terms of two parameters which measure accuracy (strength) of classifiers and correlation between classifiers.

Characterizing random forests

Definition (strength)

The **strength**, s, for a set of classifiers $\{h(\mathbf{x}, \Theta)\}$ is:
$s = E_{\mathbf{X},Y} mr(\mathbf{X}, Y) = E_{\mathbf{X},Y}[\mathbf{P}_{\Theta}(h(\mathbf{X}, \Theta) = Y) - max_{j \neq Y} \mathbf{P}_{\Theta}(h(\mathbf{X}, \Theta) = j)]$.

Defining $\hat{j}(\mathbf{X}, Y) = argmax_{j \neq Y} \mathbf{P}_{\Theta}(h(\mathbf{X}, \Theta) = j)$, we define the raw margin function.

Definition (raw margin function)

The **raw margin function** is:
$rmg(\Theta, \mathbf{X}, Y) = I(h(\mathbf{X}, \Theta) = Y) - I(h(\mathbf{X}, \Theta) = \hat{j}(\mathbf{X}, Y))$

note:

$$\begin{aligned} mr(\mathbf{X}, Y) &= \mathbf{P}_{\Theta}(h(\mathbf{X}, \Theta) = Y) - max_{j \neq Y} \mathbf{P}_{\Theta}(h(\mathbf{X}, \Theta) = j) \\ &= \mathbf{P}_{\Theta}(h(\mathbf{X}, \Theta) = Y) - \mathbf{P}_{\Theta}(h(\mathbf{X}, \Theta) = \hat{j}(\mathbf{X}, Y)) \\ &= E_{\Theta}[I(h(\mathbf{X}, \Theta) = Y) - I(h(\mathbf{X}, \Theta) = \hat{j}(\mathbf{X}, Y))] \\ &= E_{\Theta}[rmg(\Theta, \mathbf{X}, Y)] \end{aligned} \tag{2}$$

Characterizing random forests

Define:

$\rho(\Theta, \Theta') = CORR(rmg(\Theta, \mathbf{X}, Y), rmg(\Theta', \mathbf{X}, Y))$, and

$\bar{\rho} = av_{\{(\Theta,\Theta')\}}\rho(\Theta, \Theta')$

Using Chebyshev's inequality (and a number of algebraic manipulations), we get a bound for the generalization error:

Theorem

$$GE^* \leq \frac{\bar{\rho}}{s^2}\ (1 - s^2)$$

Goal: minimize the correlation, $\bar{\rho}$, and maximize the strength, s.

How? → Breiman's idea: use randomness that maintains strength while minimizing $\bar{\rho}$

Random Features

Breiman proposed two methods of random features for injecting "useful" randomness. At each node of the tree, split the tree on:

1. randomly selected features.
2. random linear combinations of features.

Random Features

Breiman proposed two methods of random features for injecting "useful" randomness. At each node of the tree, split the tree on:

1. randomly selected features.
2. random linear combinations of features.

Claimed desirable characteristics of using random features:

- accuracy as good or better than Adaboost
- relatively robust to outliers and noise
- faster than bagging or boosting
- provides estimates of error, strength, correlation, and variable importance
- simple and easily parallelized

Random Features

Breiman advocates for using random features in conjunction with bagging, since:

1. bagging seems to enhance accuracy when used in conjunction with random features.
2. bagging provides estimates of generalization error (GE^*), strength, and correlation.
3. out of bag estimates remove the need for a test set.

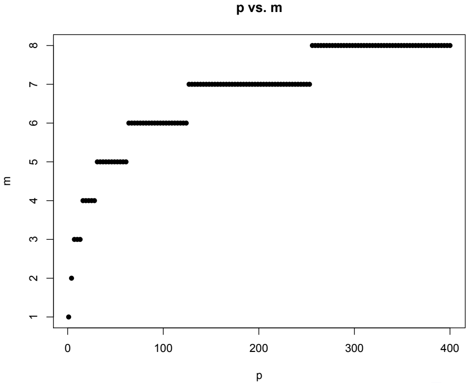
Random Features: random inputs

Idea: at each node, select a small, random group of features to split on.

Grow the tree using CART methodology to maximum size without pruning.

Choice of number of features, m:

1. $m = 1$
2. $m = \lfloor log_2(p+1) \rfloor$



Data Sets

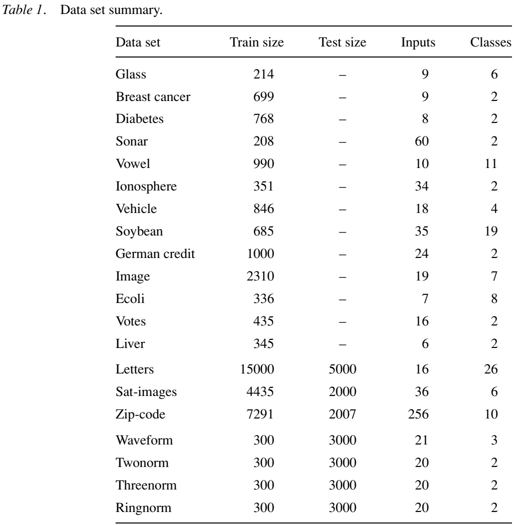


Table 1. Data set summary.

Data set	Train size	Test size	Inputs	Classes
Glass	214	–	9	6
Breast cancer	699	–	9	2
Diabetes	768	–	8	2
Sonar	208	–	60	2
Vowel	990	–	10	11
Ionosphere	351	–	34	2
Vehicle	846	–	18	4
Soybean	685	–	35	19
German credit	1000	–	24	2
Image	2310	–	19	7
Ecoli	336	–	7	8
Votes	435	–	16	2
Liver	345	–	6	2
Letters	15000	5000	16	26
Sat-images	4435	2000	36	6
Zip-code	7291	2007	256	10
Waveform	300	3000	21	3
Twonorm	300	3000	20	2
Threenorm	300	3000	20	2
Ringnorm	300	3000	20	2

Empirical Results: random inputs

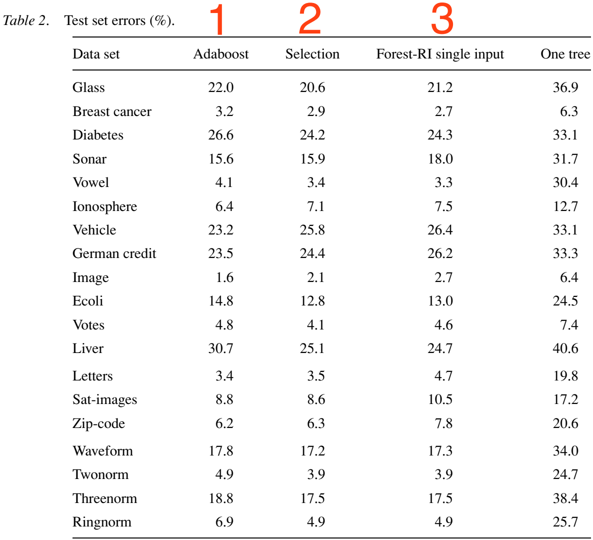


Table 2. Test set errors (%).

	1	2	3	
Data set	Adaboost	Selection	Forest-RI single input	One tree
Glass	22.0	20.6	21.2	36.9
Breast cancer	3.2	2.9	2.7	6.3
Diabetes	26.6	24.2	24.3	33.1
Sonar	15.6	15.9	18.0	31.7
Vowel	4.1	3.4	3.3	30.4
Ionosphere	6.4	7.1	7.5	12.7
Vehicle	23.2	25.8	26.4	33.1
German credit	23.5	24.4	26.2	33.3
Image	1.6	2.1	2.7	6.4
Ecoli	14.8	12.8	13.0	24.5
Votes	4.8	4.1	4.6	7.4
Liver	30.7	25.1	24.7	40.6
Letters	3.4	3.5	4.7	19.8
Sat-images	8.8	8.6	10.5	17.2
Zip-code	6.2	6.3	7.8	20.6
Waveform	17.8	17.2	17.3	34.0
Twonorm	4.9	3.9	3.9	24.7
Threenorm	18.8	17.5	17.5	38.4
Ringnorm	6.9	4.9	4.9	25.7

Empirical Results: random inputs

- Columns:
  1. Average Adaboost test error for $k = 100$ trees and $nrep = 100$.
  2. Lowest average out of bag error between the two m values for $k = 100$ trees and $nrep = 100$. (typically)
  3. Average test set error with $m = 1$ for $k = 100$ trees and $nrep = 100$.
- "It was surprising that using a single randomly chosen variable to split on at each node could produce good accuracy (relative to choosing several)."
- "Random input selection can be much faster than Adaboost or Bagging."

Random Features: random linear combinations

Idea: If p is small, using a large value of m might not decrease correlation. Create new features as follows:

1. Standardize original features.
2. Specify L, the number of features to be combined.
3. Randomly select L standardized features.
4. Choose F, the number of new features.
5. For each f in F, generate random coefficients from $[-1, 1]$ for each feature.
6. Create F new features using linear combinations of the features and random coefficients.
7. Search for and split on the best new feature f

Empirical results for $L = 3$ and $F = 2$ (other values of F were tried):

Empirical Results: random linear combinations

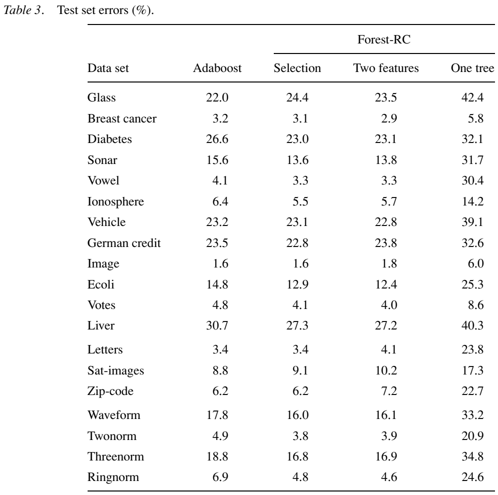


Table 3. Test set errors (%).

		Forest-RC		
Data set	Adaboost	Selection	Two features	One tree
Glass	22.0	24.4	23.5	42.4
Breast cancer	3.2	3.1	2.9	5.8
Diabetes	26.6	23.0	23.1	32.1
Sonar	15.6	13.6	13.8	31.7
Vowel	4.1	3.3	3.3	30.4
Ionosphere	6.4	5.5	5.7	14.2
Vehicle	23.2	23.1	22.8	39.1
German credit	23.5	22.8	23.8	32.6
Image	1.6	1.6	1.8	6.0
Ecoli	14.8	12.9	12.4	25.3
Votes	4.8	4.1	4.0	8.6
Liver	30.7	27.3	27.2	40.3
Letters	3.4	3.4	4.1	23.8
Sat-images	8.8	9.1	10.2	17.3
Zip-code	6.2	6.2	7.2	22.7
Waveform	17.8	16.0	16.1	33.2
Twonorm	4.9	3.8	3.9	20.9
Threenorm	18.8	16.8	16.9	34.8
Ringnorm	6.9	4.8	4.6	24.6

Empirical Results: random linear combinations

- Overall, compares more favorably to Adaboost than random inputs
- For smaller data sets, $F = 2$ gave best test error, while $F = 8$ performed better for large data sets.
- Some experimenting with larger values of F for large data sets. (see paper)

Note: see paper for treatment of categorical feature variables.

Empirical Results: strength and correlation

- Recall: we wanted to inject randomness with the goal of minimizing correlation $\bar{\rho}$ and maximizing strength, s.
- Question: how do the values of $\bar{\rho}$ and s change with the choice of m?
- Empirical study: for different data sets, create plots of strength, correlation, test error, and out of bag error.

Empirical Results: strength and correlation

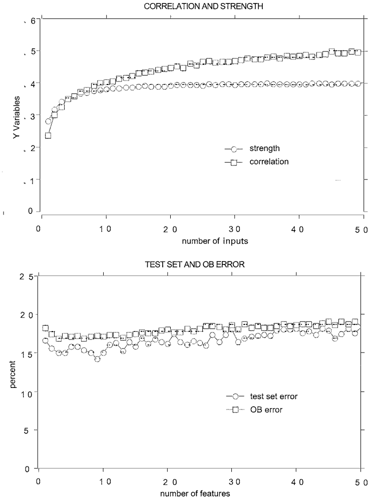


Figure 1. Effect of number of inputs on sonar data.

Empirical Results: strength and correlation

- Results from two other data sets given in paper
- Different behavior for large vs. small data sets
- Results for regression also provided in paper

The End

Psuedo-references:

- Breiman, Leo. Random Forests. Machine Learning, 2001.
- Bernard, Heutte, and Adam. A Study of Strength and Correlation in Random Forests. International Conference on Intelligent Computing, 2010.

Questions?