Random forest

This article is about the machine learning technique. For other kinds of random tree, see Random tree (disambiguation).

Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. The algorithm for inducing a random forest was developed by Leo Breiman[1] and Adele Cutler,[2] and "Random Forests" is their trademark. The term came from **random decision forests** that was first proposed by Tin Kam Ho of Bell Labs in 1995. The method combines Breiman's "bagging" idea and the random selection of features, introduced independently by Ho[3][4] and Amit and Geman[5] in order to construct a collection of decision trees with controlled variance.

The selection of a random subset of features is an example of the random subspace method, which, in Ho's formulation, is a way to implement classification proposed by Eugene Kleinberg.[6]

1 History

The early development of random forests was influenced by the work of Amit and Geman[5] who introduced the idea of searching over a random subset of the available decisions when splitting a node, in the context of growing a single tree. The idea of random subspace selection from Ho[4] was also influential in the design of random forests. In this method a forest of trees is grown, and variation among the trees is introduced by projecting the training data into a randomly chosen subspace before fitting each tree. Finally, the idea of randomized node optimization, where the decision at each node is selected by a randomized procedure, rather than a deterministic optimization was first introduced by Dietterich.[7]

The introduction of random forests proper was first made in a paper by Leo Breiman.[1] This paper describes a method of building a forest of uncorrelated trees using a CART like procedure, combined with randomized node optimization and bagging. In addition, this paper combines several ingredients, some previously known and some novel, which form the basis of the modern practice of random forests, in particular:

1. Using out-of-bag error as an estimate of the generalization error.
2. Measuring variable importance through permutation.

The report also offers the first theoretical result for random forests in the form of a bound on the generalization error which depends on the strength of the trees in the forest and their correlation.

More recently several major advances in this area have come from Microsoft Research,[8] which incorporate and extend the earlier work from Breiman.

2 Algorithm

2.1 Preliminaries: decision tree learning

Decision trees are a popular method for various machine learning tasks. Tree learning "come[s] closest to meeting the requirements for serving as an off-the-shelf procedure for data mining", say Hastie *et al.*, because it is invariant under scaling and various other transformations of feature values, is robust to inclusion of irrelavant features, and produces inspectable models. However, they are seldom accurate.[9]:352

In particular, trees that are grown very deep tend to learn highly irregular patterns: they overfit their training sets, because they have low bias, but very high variance. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance.[9]:587–588 This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance of the final model.

2.2 Tree bagging

Main article: Bootstrap aggregating

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1, \ldots, x_n$ with responses $Y = y_1, \ldots, y_n$, bagging repeatedly selects a bootstrap sample of the training set and fits trees to these samples:

For $b = 1, \ldots, B$:

1. Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.
2. Train a decision or regression tree f_b on X_b, Y_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x')$$

or by taking the majority vote in the case of decision trees.

In the above algorithm, B is a free parameter. Typically, a few hundred to several thousand trees are used, depending on the size and nature of the training set. Increasing the number of trees tends to decrease the variance of the model, without increasing the bias. As a result, the training and test error tend to level off after some number of trees have been fit. An optimal number of trees B can be found using cross-validation, or by observing the out-of-bag error: the mean prediction error on each training sample x_i, using only the trees that did not have x_i in their bootstrap sample.[10]

2.3 From bagging to random forests

Main article: Random subspace method

The above procedure describes the original bagging algorithm for trees. Random forests differ in only one way from this general scheme: they use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. This process is sometimes called “feature bagging”. The reason for doing this is the correlation of the trees in an ordinary bootstrap sample: if one or a few features are very strong predictors for the response variable (target output), these features will be selected in many of the B trees, causing them to become correlated.

Typically, for a dataset with p features, √p features are used in each split.

2.4 Extensions

Adding one further step of randomization yields *extremely randomized trees*, or ExtraTrees. These are trained using bagging and the random subspace method, like in an ordinary random forest, but additionally the top-down splitting in the tree learner is randomized. Instead of computing the locally *optimal* feature/split combination (based on, e.g., information gain or the Gini coefficient), for each feature under consideration a random value is selected in the feature's empirical range (in the tree's training set, i.e., the bootstrap sample). The best of these is then chosen as the split.[11]

3 Relationship to Nearest Neighbors

Given a set of training data

$$\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^{n}$$

a weighted neighborhood scheme makes a prediction for a query point X , by computing

$$\hat{Y} = \sum_{i=1}^{n} W_i(X) Y_i$$

for some set of non-negative weights $\{W_i(X)\}_{i=1}^{n}$ which sum to 1. The set of points X_i where $W_i(X) > 0$ are called the neighbors of X . A common example of a weighted neighborhood scheme is the k-NN algorithm which sets $W_i(X) = 1/k$ if X_i is among the k closest points to X in $\mathcal{D}_n$ and 0 otherwise.

Random forests with constant leaf predictors can be interpreted as a weighted neighborhood scheme in the following way. Given a forest of M trees, the prediction that the m -th tree makes for X can be written as

$$T_m(X) = \sum_{i=1}^{n} W_{im}(X) Y_i$$

where $W_{im}(X)$ is equal to $1/k_m$ if X and X_i are in the same leaf in the m -th tree and 0 otherwise, and k_m is the number of training data which fall in the same leaf as X in the m -th tree. The prediction of the whole forest is

$$F(X) = \frac{1}{M} \sum_{m=1}^{M} T_m(X) = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{n} W_{im}(X) Y_i = \sum_{i=1}^{n} \Bigg(\frac{1}{M} \sum_{m=1}^{M} W_i$$

which shows that the random forest prediction is a weighted average of the Y_i 's, with weights

$$W_i(X) = \frac{1}{M} \sum_{m=1}^{M} W_{im}(X)$$

The neighbors of X in this interpretation are the points X_i which fall in the same leaf as X in at least one tree of the forest. In this way, the neighborhood of X depends in a complex way on the structure of the trees, and thus on the structure of the training set.

This connection was first described by Lin and Jeon in a technical report from 2001[12] where they show that the shape of the neighborhood used by a random forest adapts to the local importance of each feature.

4 Variable importance

Random forests can be used to rank the importance of variables in a regression or classification problem in a natural way. The following technique was described in Breiman's original paper[1] and is implemented in the R package randomForest.[2]

The first step in measuring the variable importance in a data set $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ is to fit a random forest to the data. During the fitting process the out-of-bag error for each data point is recorded and averaged over the forest (errors on an independent test set can be substituted if bagging is not used during training).

To measure the importance of the j -th feature after training, the values of the j -th feature are permuted among the training data and the out-of-bag error is again computed on this perturbed data set. The importance score for the j -th feature is computed by averaging the difference in out-of-bag error before and after the permutation over all trees. The score is normalized by the standard deviation of these differences.

Features which produce large values for this score are ranked as more important than features which produce small values.

This method of determining variable importance has some drawbacks. For data including categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels. Methods such as partial permutations[13][14] and growing unbiased trees[15] can be used to solve the problem. If the data contain groups of correlated features of similar relevance for the output, then smaller groups are favored over larger groups.[16]

5 Unsupervised learning with random forests

As part of their construction, RF predictors naturally lead to a dissimilarity measure between the observations. One can also define an RF dissimilarity measure between unlabeled data: the idea is to construct an RF predictor that distinguishes the "observed" data from suitably generated synthetic data.[1][17] The observed data are the original unlabeled data and the synthetic data are drawn from a reference distribution. An RF dissimilarity can be attractive because it handles mixed variable types well, is invariant to monotonic transformations of the input variables, and is robust to outlying observations. The RF dissimilarity easily deals with a large number of semi-continuous variables due to its intrinsic variable selection; for example, the "Addcl 1" RF dissimilarity weighs the contribution of each variable according to how dependent it is on other variables. The RF dissimilarity has been used in a variety application, e.g. to find clusters of patients based on tissue marker data.[18]

6 Variants

Instead of decision trees, linear models have been proposed and evaluated as base estimators in random forests, in particular multinomial logistic regression and naive Bayes classifiers.[19][20]

7 See also

- Decision tree learning
- Gradient boosting
- Randomized algorithm
- Bootstrap aggregating (bagging)
- Ensemble learning
- Boosting
- Non-parametric statistics

8 References

[1] Breiman, Leo (2001). "Random Forests". *Machine Learning* **45** (1): 5–32. doi:10.1023/A:1010933404324.

[2] Liaw, Andy (16 October 2012). "Documentation for R package randomForest". Retrieved 15 March 2013.

[3] Ho, Tin Kam (1995). "Random Decision Forest". Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

[4] Ho, Tin Kam (1998). "The Random Subspace Method for Constructing Decision Forests". *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** (8): 832–844. doi:10.1109/34.709601.

[5] Amit, Yali; Geman, Donald (1997). "Shape quantization and recognition with randomized trees". *Neural Computation* **9** (7): 1545–1588. doi:10.1162/neco.1997.9.7.1545.

[6] Kleinberg, Eugene (1996). "An Overtraining-Resistant Stochastic Modeling Method for Pattern Recognition". *Annals of Statistics* **24** (6): 2319–2349. doi:10.1214/aos/1032181157. MR 1425956.

[7] Dietterich, Thomas (2000). "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization". *Machine Learning*: 139–157.

[8] Criminisi, Antonio; Shotton, Jamie; Konukoglu, Ender (2011). "Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning". *Foundations and Trends in Computer Vision* **7**: 81–227. doi:10.1561/0600000035.

[9] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). *The Elements of Statistical Learning* (2nd ed.). Springer. ISBN 0-387-95284-5.

[10] Gareth James; Daniela Witten; Trevor Hastie; Robert Tibshirani (2013). *An Introduction to Statistical Learning*. Springer. pp. 316–321.

[11] Geurts, P.; Ernst, D.; Wehenkel, L. (2006). "Extremely randomized trees". *Machine Learning* **63**: 3. doi:10.1007/s10994-006-6226-1.

[12] Lin, Yi; Jeon, Yongho (2002), "Random forests and adaptive nearest neighbors", Technical Report No. 1055, University of Wisconsin Missing or empty |title= (help)

[13] Deng,H.; Runger, G.; Tuv, E. (2011). "Bias of importance measures for multi-valued attributes and solutions". Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN). pp. 293–300.

[14] Altmann A, Tolosi L, Sander O, Lengauer T (2010). "Permutation importance:a corrected feature importance measure". *Bioinformatics*. doi:10.1093/bioinformatics/btq134.

[15] Strobl,C.; Boulesteix,A.; Augustin,T. (2007). "Unbiased split selection for classification trees based on the Gini index". *Computational Statistics & Data Analysis*: 483–501.

[16] Tolosi L, Lengauer T (2011). "Classification with correlated features: unreliability of feature ranking and solutions.". *Bioinformatics*. doi:10.1093/bioinformatics/btr300.

[17] Shi, T., Horvath, S. (2006). "Unsupervised Learning with Random Forest Predictors". *Journal of Computational and Graphical Statistics* **15** (1): 118–138. doi:10.1198/106186006X94072.

[18] Shi, T., Seligson D., Belldegrun AS., Palotie A, Horvath, S. (2005). "Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma". *Modern Pathology* **18** (4): 547–557. doi:10.1038/modpathol.3800322. PMID 15529185.

[19] Prinzie, A., Van den Poel, D. (2008). "Random Forests for multiclass classification: Random MultiNomial Logit". *Expert Systems with Applications* **34** (3): 1721–1732. doi:10.1016/j.eswa.2007.01.029.

[20] Prinzie, A., Van den Poel, D. (2007). Random Multiclass Classification: Generalizing Random Forests to Random MNL and Random NB, Dexa 2007, Lecture Notes in Computer Science, 4653, 349–358.

9 External links

- Random Forests classifier description (Site of Leo Breiman)
- Liaw, Andy & Wiener, Matthew "Classification and Regression by randomForest" R News (2002) Vol. 2/3 p. 18 (Discussion of the use of the random forest package for R)
- Ho, Tin Kam (2002). "A Data Complexity Analysis of Comparative Advantages of Decision Forest Constructors". Pattern Analysis and Applications 5, p. 102-112 (Comparison of bagging and random subspace method)
- Prinzie, Anita; Poel, Dirk (2007). "Database and Expert Systems Applications". Lecture Notes in Computer Science **4653**. p. 349. doi:10.1007/978-3-540-74469-6_35. ISBN 978-3-540-74467-2. |chapter= ignored (help)
- C# implementation of random forest algorithm for categorization of text documents supporting reading of documents, making dictionaries, filtering stop words, stemming, counting words, making document-term matrix and its usage for building random forest and further categorization.
- A python implementation of the random forest algorithm working in regression, classification with multi-output support.

10 Text and image sources, contributors, and licenses

10.1 Text

10.2 Images

10.3 Content license