

Random Forests: Finding Quasars

Leo Breiman Michael Last John Rice
Department of Statistics
University of California, Berkeley

0.1 Introduction

The automatic classification of objects from catalogues is a common statistical problem encountered in many surveys. From a list of values of variables (e.g. color, magnitude) associated with an object, it is desired to identify the object's type (e.g. star, galaxy). In the last section of this paper, we discuss an example in which we classify objects as quasars or non-quasars using the combined results of a radio survey and an optical survey. Such classification helps guide the choice of which objects to follow up with relatively expensive spectroscopic measurements.

The last five years of research in the Machine Learning field has produced classification methods with significantly higher accuracies than previous methods. There have been two lines of productive research. One estimates the border between classes by increasing the dimensionality of the input predictor space. The classifiers produced by this method are called Support Vector Machines [2], [3].

The other creates a varied ensemble of classifiers, lets each classifier vote for the class it favors, and then outputs the classification that has the plurality of votes. The most accurate classifier of this type is called Random Forests [1], abbreviated RF. We will describe the construction of RF, and compare its performance with single CART trees. RF can also quantify which variables are important to the class and this procedure is described as well.

0.2 Construction of RF

Recall the steps in constructing an ordinary CART tree: A node is a subset of the data. The root node contains all data. At each node, search through all variables to find the best split into two children nodes. Split all the way down and then prune the tree up to get minimal test set error.

The construction of RF differs:

Table 1. Data Set Descriptions.

Data Set	Training	Test	Variables	Classes
cancer	699	-	9	2
ionosphere	351	-	34	2
diabetes	768	-	8	2
glass	214	-	9	6
soybean	683	-	35	19
letters	15,000	5000	16	26
satellite	4,435	2000	36	6
shuttle	43,500	14,500	9	7
DNA	2,000	1,186	60	3
digit	7,291	2,007	256	10

1. The root node contains a bootstrap sample from the original data. A different bootstrap sample is drawn for each tree to be grown.
2. An integer K is fixed, K is much smaller than the number of variables. K is the only parameter that needs to be specified. The default is the square root of number of variables.
3. At each node, K of the variables are selected at random. Only these variables are searched through for the best split. The largest tree possible is grown and is not pruned.
4. The forest consists of N trees. To classify a new object having coordinates x , put x down each of the N trees. Each tree gives a classification for x .
5. The forest chooses that classification having the most votes out of the N votes cast

Code for random forests is publicly available. ¹

0.3 Accuracy of RF Compared to CART

Accuracy of single trees (CART) to random forests is compared using data sets from the UCI repository (<ftp://ics.uci.edu/pub/MachineLearningDatabases>).

For the five smaller data sets above the line, the test set error was estimated by leaving out a random 10% of the data, then running CART and the forest on the other 90%.

¹<http://www.stat.Berkeley.EDU/users/breiman/>

Table 2. Test Set Misclassification Error (%)

Data Set	Forest	Single Tree
breast cancer	2.9	5.9
ionosphere	5.5	11.2
diabetes	24.2	25.3
glass	22.0	30.4
soybean	5.7	8.6
letters	3.4	12.4
satellite	8.6	14.8
shuttle	7.0	62.0
DNA	3.9	6.2
digit	6.2	17.1

The left-out 10% was run down the tree and the forest and the error on this 10% computed for both. This was repeated 100 times and the errors averaged. The larger data sets below the line came with a separate test set.

The reductions in test set error are dramatic—often over 50%, and almost always, over 30%. RF achieves state-of-the-art accuracy and on the synthetic data sets it has been tested on, where the lowest possible error rate can be analytically computed, gets close to this lower limit.

0.4 RF Byproducts

A wealth of information can be obtained in a single run of Random Forests, including test set error rate and variable importance. This information comes from using the “out-of-bag” cases in the training set that have been left out of the bootstrapped training set.

Each tree is constructed using a different bootstrap sample from the original data. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the k-th tree.

Test Set Error Rate:

Put each case left out in the construction of the k-th tree down the k-th tree to get a classification. In this way, a test set classification is gotten for each case in about one third of the trees. Let the final test set classification of the forest be the class having the most votes. Compare this classification with the classification given in the data to get an estimate of the test set error.

Variable Importance

To estimate the importance of variable #4: In the left out cases for the k-th tree, randomly permute all values of variable #4. Put these new covariate

values down the tree and get classifications. Proceed as though computing a new test set error. The amount by which this new error exceeds the original test set error is defined as the importance of variable #4.

0.5 Application: Automatic Identification of Quasars

In [4] decision trees were used to automatically identify quasaars, combining information from the FIRST survey and from POSS-I plates. The aim was to construct a radio-selected sample of optically bright quasars, and in particular to bridge the gap between radio-loud and radio-quiet quasars. Continuing that effort with an enlarged set of data, we trained classifiers on 2127 objects (1366 quasars) identified from their spectra.

The following variables were used in constructing the classifiers:

1. The result of a star/galaxy classifier for the red plate
2. Another star/galaxy classifier for the red plate
3. Red magnitude
4. A star/galaxy classifier for the blue plate
5. Another star/galaxy classifier for the blue plate
6. Blue magnitude
7. Color (blue magnitude minus red magnitude)
8. Separation between radio and optical sources in arcseconds
9. Another estimate of separation between radio and optical sources
10. Radio peak flux
11. Radio integrated flux

On the basis of their spectra, the objects were classified in the following categories:

1. A: Narrow line Active Galactic Nucleus
2. B: BL Lac (a kind of blazar)
3. G: Galaxy without emission lines
4. H: H/II star forming galaxy
5. Q: Quasar
6. S: Star

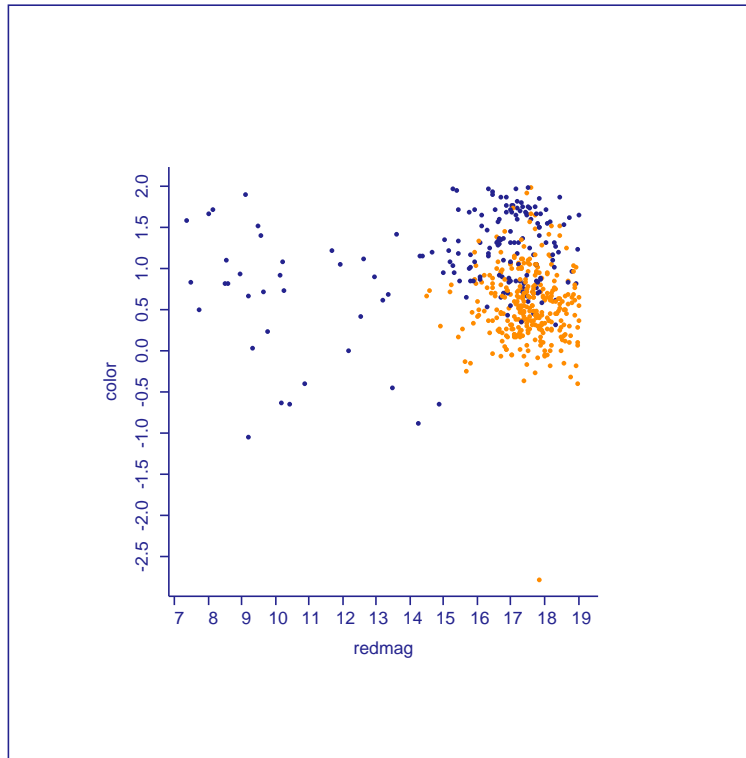


Figure 1. Color versus red magnitude. Quasars are lighter.

The task was thus to use the measurements of the variables listed above to automatically classify objects into these categories and in particular to discriminate quasars from other types of objects. As would be expected, there is a substantial amount of information available from color and magnitude, as shown in Figure 1.

An automatic classifier carves up the 11 dimensional space defined by the variables into regions corresponding to different types of objects. This is illustrated in Figure 2 which shows a projection of the data onto a plane determined by several of the variables. The figure indicates that one should be able to achieve fairly good separation.

When objects were classified as either quasars or non-quasars, random forests had misclassification rate of 14.3%. For baseline comparison, a standard classification tree had an error rate of 19.7%. A support vector machine had an error rate of 13.9%, comparable to that of random forests. It might be thought that basically only color and magnitude are informative, but this is not the case: when only these variables are used as classifiers, the error rate is 19.2%. When the categories of blazars and quasars were merged so that one does not try to distinguish between them,

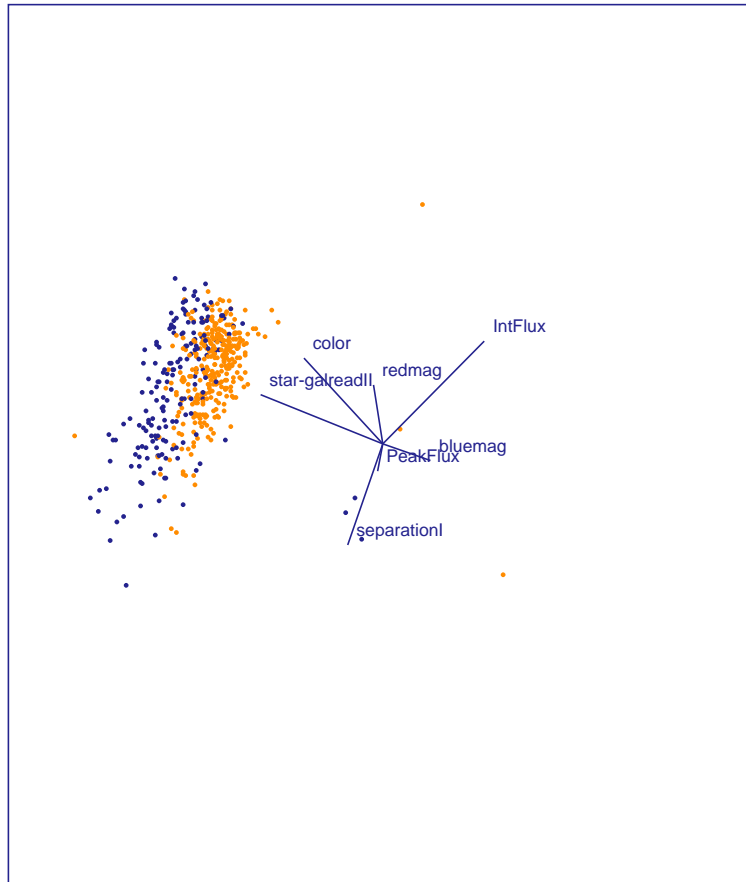


Figure 2. A projection of the data. Quasars are lighter.

the error rate for random forests dropped to 10.5% (Examination of figures like those above makes it clear that it is quite difficult to discriminate quasars from blazars.)

Figure 3 shows that the misidentified quasars tended to be bluer and brighter. The quasar fraction increases for fainter objects (because the number of quasars per square degree rises very rapidly as we go fainter), which makes fainter samples easier to classify.

Variables were scored for importance, as discussed above: Color is thus by far the most important variable for determining the classification, but as we have seen above adding variables other than color and magnitude increases predictive performance.

Random forests produce an estimate of the probability, $P(Q)$, that an object is a quasar. Examination of the results shows that these probabilities are “calibrated,” i.e. of those objects for which the estimated probability

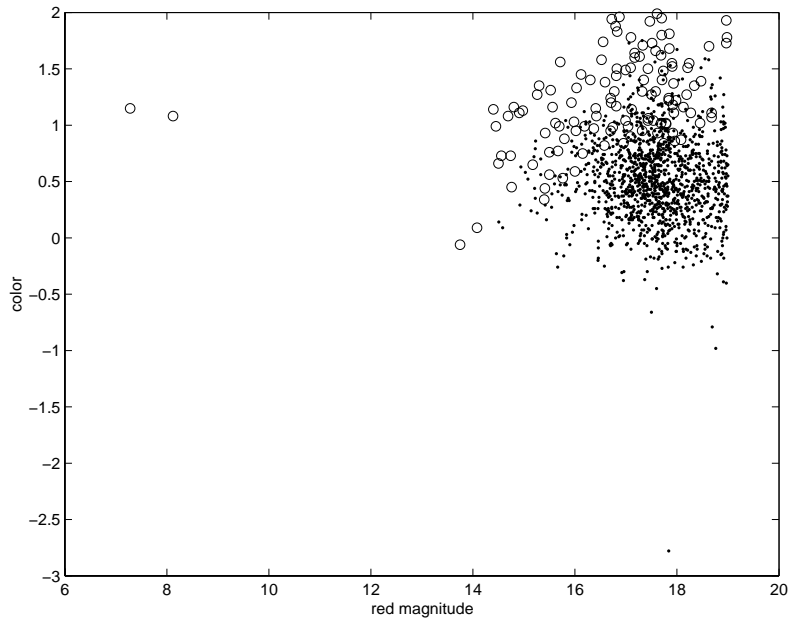


Figure 3. Errors in quasar identification. Misidentified quasars are shown as circles.

Table 3. Variable importance

Variable	Importance
red star/gal classifier 1	.99
red star/gal classifier 2	.33
red magnitude	4.95
blue star/gal classifier 1	.33
blue star/gal classifier 2	2.64
blue magnitude	.9
color	33.0
radio-optical separation 1	5.61
radio-optical separation 2	1.53
radio peak flux	1.98
radio integrated flux	.33

of being a quasar is 90%, about 90% are in fact quasars, etc. The decision of whether to follow up an object with spectroscopic observations can thus be guided by its probability $P(Q)$.

An object can be declared to be a quasar if $P(Q) > p$, for a given p (in the results quoted above we used the threshold $p = 0.5$). Varying p produces a tradeoff between two types of errors – false positives (calling an

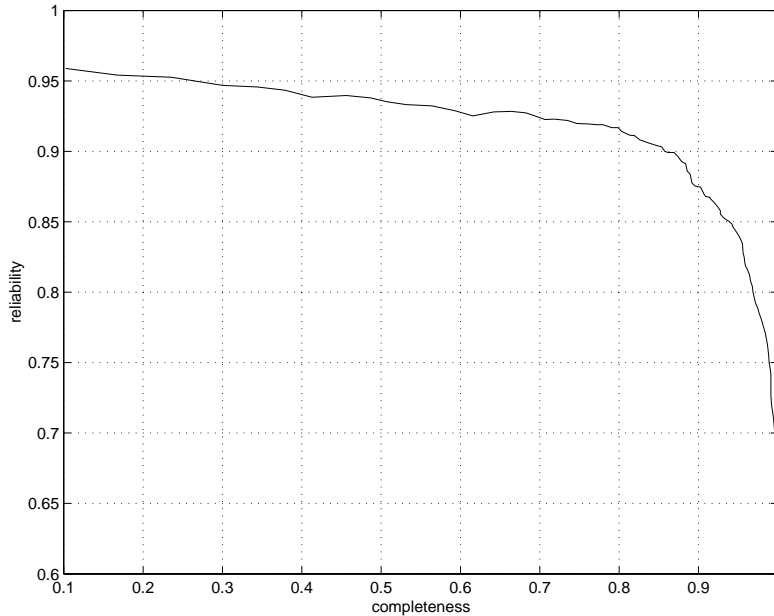


Figure 4. Completeness-reliability curve

object a quasar when it is not) and false negatives (failing to identify a real quasar as such). Equivalently we can define completeness as the fraction of actual quasars included and reliability as the fraction selected that are in fact quasars and view completeness and reliability as functions of p , as shown in Figure 4. From this figure we see that completeness of 90% can be achieved with about 87% reliability.

The classification errors can be examined to find those for which misidentified quasars were badly misidentified, i.e. $P(Q)$ is small. The quasar fraction increases for fainter objects (because the number of quasars per square degree rises very rapidly as we go fainter), which makes fainter samples easier to classify. You can see this effect in your plot that shows the misclassified objects (as large colored dots) in a plot of color vs. red magnitude—misclassifications are much more common for quasars brighter than 16th magnitude. Also, bluer quasars tend to be more likely to be misidentified.

More ambitiously, we attempted to classify each object into each of the categories above, not merely as quasar or non-quasar. The results are shown in the following “confusion matrix,” shown in Table 0.5. The columns give the true classes and the rows give the guessed classes. Thus 59 of the 1366 quasars were misidentified as H , etc. It is interesting that the completeness-reliability curve for classifying quasars when attempting to identify all objects is virtually identical to that when quasars are merely

Table 4. Confusion matrix.

True Class	Assigned Class					
	A	B	G	H	Q	S
A	36	5	15	26	9	7
B	4	10	2	2	5	0
G	10	1	1	9	90	0
H	59	9	28	159	40	18
Q	36	85	3	59	1292	58
S	1	2	2	8	20	88

discriminated from all other objects, so that little is lost in being more ambitious.

References

- [1] L. Breiman. Random forests—random features. Technical Report 567, Department of Statistics, University of California, Berkeley, 1999.
- [2] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.
- [3] V. Vapnik. *Statistical Learning Theory*. Springer, 1998.
- [4] R. L. White, R. H. Becker, M. D. Gregg, S. A. Laurent-Muehleisen, M. S. Brotherton, C. D. Impey, C. E. Petry, C. B. Folz, F. H. Chaffee, G.T. Richards, W. R. Oergerle, D. J. Helfand, R. G. McMahon, and J. E. Cabanela. The *FIRST* bright quasar survey. II. 60 nights and 1200 spectra later. *The Astrophysical Journal Supplement Series*, 126:133–207, 2000.