Classification using stochastic ensembles

Linden McBride and Austin Nichols

July 31, 2014

Topics

- Discriminant analysis and classfication
- Classification and Regression Trees
- Stochastic ensemble methods
- Our application: USAID Poverty Assessment Tools
- Other applications

Discriminant analysis and classification

Classification, or predictive discriminant analysis, involves the assignment of observations to classes.

Predictions are based on a model trained in a dataset in which class membership is known (Huberty 1994, Rencher 2002, Hastie et al. 2009).

- Prediction of qualitative response
- With class>2, linear regression methods generally not appropriate
- Methods available in statistics, machine learning, predictive analytics

Our classification problem: identifying poor from nonpoor

To fulfill the terms of a social safety net intervention in a developing country, we wish to classify households as poor or nonpoor based on a set of observable household characteristics. Households classified as poor will recieve a cash transfer.

Our classification problem: identifying poor from nonpoor

To fulfill the terms of a social safety net intervention in a developing country, we wish to classify households as poor or nonpoor based on a set of observable household characteristics. Households classified as poor will recieve a cash transfer.

Our objective is accurate, out-of-sample, prediction: we want to make predictions about a household's poverty status (otherwise unknown) using a model trained by other households (poverty status known) in that population. Assume that we are indifferent between types of misclassification.

Our classification problem: identifying poor from nonpoor

To fulfill the terms of a social safety net intervention in a developing country, we wish to classify households as poor or nonpoor based on a set of observable household characteristics. Households classified as poor will recieve a cash transfer.

Our objective is accurate, out-of-sample, prediction: we want to make predictions about a household's poverty status (otherwise unknown) using a model trained by other households (poverty status known) in that population. Assume that we are indifferent between types of misclassification.

This is a stylized example of the problem faced by USAID, the World Bank, and other institutions attempting to target the poor in developing countries where income status is difficult to assess.

Discriminant methods available in Stata

Many discrimination methods are available, including linear, quadratic, logistic, and nonparametric methods:

MV discrim lda and [MV] candisc.

MV discrim qda provides quadratic discriminant analysis

MV discrim logistic provides logistic discriminant analysis.

MV discrim knn provides kth-nearest-neighbor discrimination.

Discriminant methods available in Stata

- Linear discriminant analysis (LDA)

Discriminant methods available in Stata

- Linear discriminant analysis (LDA)
  - Assumes obs within each class are normally distributed with class specific means but common variance

Discriminant methods available in Stata

- Linear discriminant analysis (LDA)
  - Assumes obs within each class are normally distributed with class specific means but common variance
  - Assigns to class for which estimated posterior probability is greatest

Discriminant methods available in Stata

- Linear discriminant analysis (LDA)
  - Assumes obs within each class are normally distributed with class specific means but common variance
  - Assigns to class for which estimated posterior probability is greatest
- Quadratic discriminant analysis (QDA)

Discriminant methods available in Stata

- Linear discriminant analysis (LDA)
  - Assumes obs within each class are normally distributed with class specific means but common variance
  - Assigns to class for which estimated posterior probability is greatest
- Quadratic discriminant analysis (QDA)
  - Assumes obs within each class are normally distributed with class specific means and variance

Discriminant methods available in Stata

- Linear discriminant analysis (LDA)
  - Assumes obs within each class are normally distributed with class specific means but common variance
  - Assigns to class for which estimated posterior probability is greatest
- Quadratic discriminant analysis (QDA)
  - Assumes obs within each class are normally distributed with class specific means and variance
  - Assigns to class for which estimated posterior probability is greatest

Discriminant methods available in Stata

- Logistic discriminant analysis (LD)

Discriminant methods available in Stata

- Logistic discriminant analysis (LD)
  - Places distributional assumption on the likelihood ratio

Discriminant methods available in Stata

- Logistic discriminant analysis (LD)
  - Places distributional assumption on the likelihood ratio
  - Assigns to class for which estimated posterior probability is greatest

Discriminant methods available in Stata

- Logistic discriminant analysis (LD)
  - Places distributional assumption on the likelihood ratio
  - Assigns to class for which estimated posterior probability is greatest
- K-nearest-neighbor discrimination (KNN)

Discriminant methods available in Stata

- Logistic discriminant analysis (LD)
  - Places distributional assumption on the likelihood ratio
  - Assigns to class for which estimated posterior probability is greatest
- K-nearest-neighbor discrimination (KNN)
  - Nonparametric

Discriminant methods available in Stata

- Logistic discriminant analysis (LD)
  - Places distributional assumption on the likelihood ratio
  - Assigns to class for which estimated posterior probability is greatest
- K-nearest-neighbor discrimination (KNN)
  - Nonparametric
  - Assigns to class based on distance from neighbors belonging to that class

Discriminant methods available in Stata

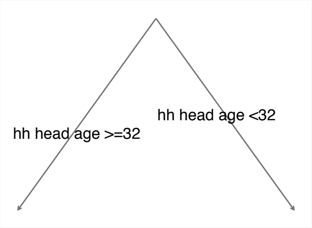
- Logistic discriminant analysis (LD)
  - Places distributional assumption on the likelihood ratio
  - Assigns to class for which estimated posterior probability is greatest
- K-nearest-neighbor discrimination (KNN)
  - Nonparametric
  - Assigns to class based on distance from neighbors belonging to that class

Methods not available in Stata include SVM, CART (limited), boosting, various ensemble methods.

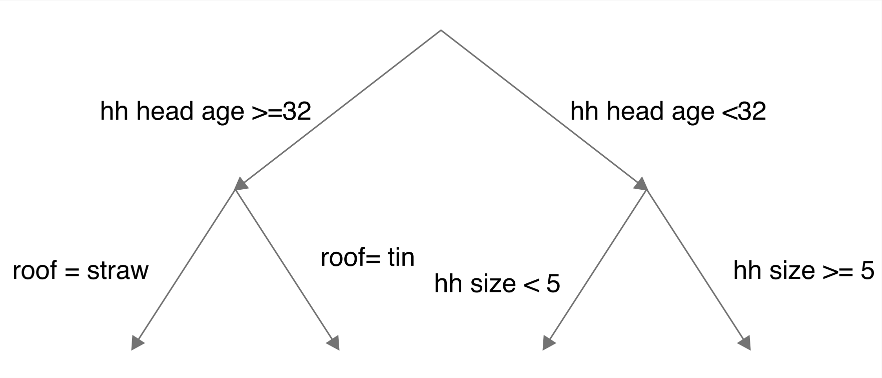
CART

Classification and regression trees recursively partition a feature space to meet some criteria (entropy reduction, minimized prediction error, etc).

Predictions for a given set of features are made based on the relative proportion of classes found in a terminal node (classification) or on the mean response for the data in that partition (regression).



CART


hh head age >=32
hh head age <32
roof = straw
roof= tin
hh size < 5
hh size >= 5

Stochastic ensemble methods

Ensemble methods construct many models on subsets of data (e.g. via resampling with replacement); they then average across these models (or allow them to vote) to obtain a less noisy prediction.

One version of this is known as **bootstrap aggregation**, or bagging (Breiman 1996a).

A stochastic ensemble method adds randomness to the construction of the models. This has the advantage of "de-correlating" models across subsets, which can reduce total variance (Breiman 2001).

Out-of-sample error is estimated by training the model in each randomly selected subset, and using the balance of the data to test the model. This estimated out-of-sample error is an unbiased estimator of the true out-of-sample prediction error (Breiman 1996b).

Stochastic ensemble methods for trees

Stochastic ensemble methods can address the weaknesses of CART models (Breiman 2001).

Stochastic ensemble methods for trees

Stochastic ensemble methods can address the weaknesses of CART models (Breiman 2001).

For example, if we allow CART to grow large, we make a bias for variance trade off: large trees will have high variance but low bias.

Stochastic ensemble methods for trees

Stochastic ensemble methods can address the weaknesses of CART models (Breiman 2001).

For example, if we allow CART to grow large, we make a bias for variance trade off: large trees will have high variance but low bias.

Bagging produces a large number of approximately unbiased and identically distributed trees. Averaging or voting across these trees significantly reduces the variance of the classifier.

Stochastic ensemble methods for trees

Stochastic ensemble methods can address the weaknesses of CART models (Breiman 2001).

For example, if we allow CART to grow large, we make a bias for variance trade off: large trees will have high variance but low bias.

Bagging produces a large number of approximately unbiased and identically distributed trees. Averaging or voting across these trees significantly reduces the variance of the classifier.

The variance of the classifier can be further reduced by de-correlating the trees via the introduction of randomness in the tree building process.

Stochastic ensemble methods for trees

Stochastic ensemble methods can address the weaknesses of CART models (Breiman 2001).

For example, if we allow CART to grow large, we make a bias for variance trade off: large trees will have high variance but low bias.

Bagging produces a large number of approximately unbiased and identically distributed trees. Averaging or voting across these trees significantly reduces the variance of the classifier.

The variance of the classifier can be further reduced by de-correlating the trees via the introduction of randomness in the tree building process.

This combination of bagging and decorrelating ensembles of trees produces classification and regression forests.

Algorithms in R and Stata

In R, classification and regression forests can be generated with **randomForest** (Breiman and Cutler 2001, Liaw and Wiener 2002). Extensions such as quantile regression forests **quantregForest** (Meinshausen 2006) are also available.

- Classification Forest (CF) and Regression Forest (RF)

Algorithms in R and Stata

In R, classification and regression forests can be generated with **randomForest** (Breiman and Cutler 2001, Liaw and Wiener 2002). Extensions such as quantile regression forests **quantregForest** (Meinshausen 2006) are also available.

- Classification Forest (CF) and Regression Forest (RF)
  - Minimizes classification error in each binary partition (CF)

Algorithms in R and Stata

In R, classification and regression forests can be generated with **randomForest** (Breiman and Cutler 2001, Liaw and Wiener 2002). Extensions such as quantile regression forests **quantregForest** (Meinshausen 2006) are also available.

- Classification Forest (CF) and Regression Forest (RF)
  - Minimizes classification error in each binary partition (CF)
  - Minimizes MSE in each binary partition (RF)

Algorithms in R and Stata

In R, classification and regression forests can be generated with **randomForest** (Breiman and Cutler 2001, Liaw and Wiener 2002). Extensions such as quantile regression forests **quantregForest** (Meinshausen 2006) are also available.

- Classification Forest (CF) and Regression Forest (RF)
  - Minimizes classification error in each binary partition (CF)
  - Minimizes MSE in each binary partition (RF)
  - Uses bagging to reduce variance

Algorithms in R and Stata

In R, classification and regression forests can be generated with **randomForest** (Breiman and Cutler 2001, Liaw and Wiener 2002). Extensions such as quantile regression forests **quantregForest** (Meinshausen 2006) are also available.

- Classification Forest (CF) and Regression Forest (RF)
  - Minimizes classification error in each binary partition (CF)
  - Minimizes MSE in each binary partition (RF)
  - Uses bagging to reduce variance
  - Randomizes over variables available at any given split

Algorithms in R and Stata

In R, classification and regression forests can be generated with **randomForest** (Breiman and Cutler 2001, Liaw and Wiener 2002). Extensions such as quantile regression forests **quantregForest** (Meinshausen 2006) are also available.

- Classification Forest (CF) and Regression Forest (RF)
  - Minimizes classification error in each binary partition (CF)
  - Minimizes MSE in each binary partition (RF)
  - Uses bagging to reduce variance
  - Randomizes over variables available at any given split
  - Estimates out-of-sample prediction error in the out-of-bag sample

Algorithms in R and Stata

In R, classification and regression forests can be generated with **randomForest** (Breiman and Cutler 2001, Liaw and Wiener 2002). Extensions such as quantile regression forests **quantregForest** (Meinshausen 2006) are also available.

- Classification Forest (CF) and Regression Forest (RF)
    - Minimizes classification error in each binary partition (CF)
    - Minimizes MSE in each binary partition (RF)
    - Uses bagging to reduce variance
    - Randomizes over variables available at any given split
    - Estimates out-of-sample prediction error in the out-of-bag sample
- Quantile Regression Forest (QRF)

Algorithms in R and Stata

In R, classification and regression forests can be generated with **randomForest** (Breiman and Cutler 2001, Liaw and Wiener 2002). Extensions such as quantile regression forests **quantregForest** (Meinshausen 2006) are also available.

- Classification Forest (CF) and Regression Forest (RF)
  - Minimizes classification error in each binary partition (CF)
  - Minimizes MSE in each binary partition (RF)
  - Uses bagging to reduce variance
  - Randomizes over variables available at any given split
  - Estimates out-of-sample prediction error in the out-of-bag sample
- Quantile Regression Forest (QRF)
  - Similar to RF, but estimates entire conditional distribution of response variable through a weighting function

Algorithms in R and Stata

In R, classification and regression forests can be generated with **randomForest** (Breiman and Cutler 2001, Liaw and Wiener 2002). Extensions such as quantile regression forests **quantregForest** (Meinshausen 2006) are also available.

- Classification Forest (CF) and Regression Forest (RF)
  - Minimizes classification error in each binary partition (CF)
  - Minimizes MSE in each binary partition (RF)
  - Uses bagging to reduce variance
  - Randomizes over variables available at any given split
  - Estimates out-of-sample prediction error in the out-of-bag sample
- Quantile Regression Forest (QRF)
  - Similar to RF, but estimates entire conditional distribution of response variable through a weighting function
  - Regression Forest analog of quantile regression

Algorithms in R and Stata

In Stata, we will use a user-written command **stens** (Nichols 2014) to classify households based on an ensemble of perfect random trees (Cutler and Zhao 2001).

- Ensemble of Perfect Random Trees

Algorithms in R and Stata

In Stata, we will use a user-written command **stens** (Nichols 2014) to classify households based on an ensemble of perfect random trees (Cutler and Zhao 2001).

- Ensemble of Perfect Random Trees
  - Grows trees randomly

Algorithms in R and Stata

In Stata, we will use a user-written command **stens** (Nichols 2014) to classify households based on an ensemble of perfect random trees (Cutler and Zhao 2001).

- Ensemble of Perfect Random Trees
  - Grows trees randomly
  - Averages over the most influential voters

Poverty Assessment Tools

The Poverty Assessment Tools were developed by the University of Maryland IRIS Center for USAID.

The IRIS tool is typically developed via quantile regression in a randomly selected subset of the data. Accuracy (out of sample prediction error) is assessed on the data not used for model develpment.

Our methods

We replicate the IRIS tool development process using the same publicly available nationally representative Living Standards Measurement Survey datasets; we then attempt to improve on their estimates.

Our methods

We replicate the IRIS tool development process using the same publicly available nationally representative Living Standards Measurement Survey datasets; we then attempt to improve on their estimates.

We randomly divide the data into training and testing sets, estimate the model in the training data and then assess accuracy in the testing data. We iterate this process 1000 times. We report the means and the 2.5^{th} and 97.5^{th} percentile confidence intervals.

Our methods

We replicate the IRIS tool development process using the same publicly available nationally representative Living Standards Measurement Survey datasets; we then attempt to improve on their estimates.

We randomly divide the data into training and testing sets, estimate the model in the training data and then assess accuracy in the testing data. We iterate this process 1000 times. We report the means and the 2.5^{th} and 97.5^{th} percentile confidence intervals.

We use the 2005 Bolivia Household Survey, the 2004/5 Malawi Integrated Household Survey, and the 2001 East Timor Living Standards Survey.

Classification error

	$P = 1$	$P = 0$
$\hat{P} = 1$	True Positive (TP)	False Positive (FP)
$\hat{P} = 0$	False Negative (FN)	True Negative (TN)

Classification error

For our application, we're interested in five accuracy measures:

- Total Accuracy (TA) $= \frac{1}{N}(TP + TN) = 1 - \frac{1}{N}(FN + FP) = 1 - MSE$

Classification error

For our application, we're interested in five accuracy measures:

- Total Accuracy (TA) $= \frac{1}{N}(TP + TN) = 1 - \frac{1}{N}(FN + FP) = 1 - MSE$
- Poverty Accuracy (PA) $= TP/(TP + FP)$

Classification error

For our application, we're interested in five accuracy measures:

- Total Accuracy (TA) $= \frac{1}{N}(TP + TN) = 1 - \frac{1}{N}(FN + FP) = 1 - MSE$
- Poverty Accuracy (PA) $= TP/(TP + FP)$
- Undercoverage (UC) $= FN/(TP + FN)$

Classification error

For our application, we're interested in five accuracy measures:

- Total Accuracy (TA) $= \frac{1}{N}(TP + TN) = 1 - \frac{1}{N}(FN + FP) = 1 - MSE$
- Poverty Accuracy (PA) $= TP/(TP + FP)$
- Undercoverage (UC) $= FN/(TP + FN)$
- Leakage (LE) $= FP/(TP + FN)$

Classification error

For our application, we're interested in five accuracy measures:

- Total Accuracy (TA) $= \frac{1}{N}(TP + TN) = 1 - \frac{1}{N}(FN + FP) = 1 - MSE$
- Poverty Accuracy (PA) $= TP/(TP + FP)$
- Undercoverage (UC) $= FN/(TP + FN)$
- Leakage (LE) $= FP/(TP + FN)$
- Balanced Poverty Accuracy Criterion (BPAC) $= TP/(TP + FP) - |FN/(TP + FP) - FP/(TP + FP)|$

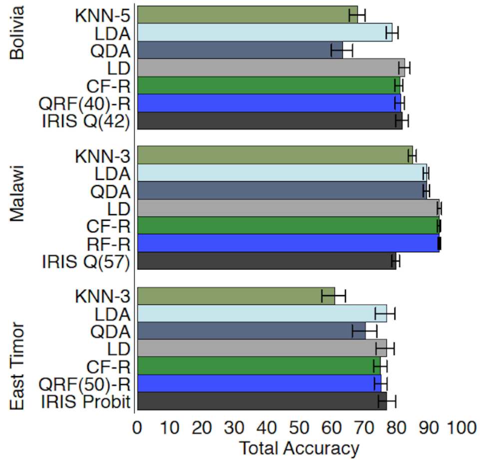
Classification error

For our application, we're interested in five accuracy measures:

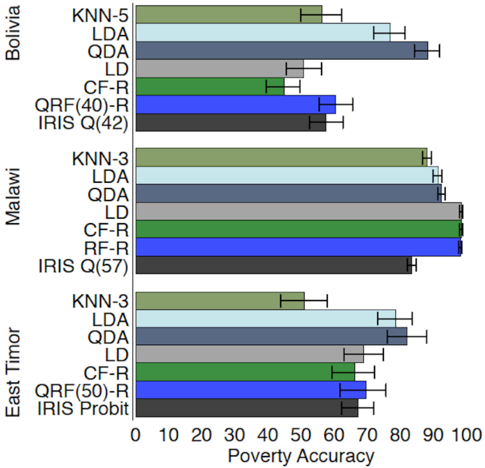
- Total Accuracy (TA) $= \frac{1}{N}(TP + TN) = 1 - \frac{1}{N}(FN + FP) = 1 - MSE$
- Poverty Accuracy (PA) $= TP/(TP + FP)$
- Undercoverage (UC) $= FN/(TP + FN)$
- Leakage (LE) $= FP/(TP + FN)$
- Balanced Poverty Accuracy Criterion (BPAC) $= TP/(TP + FP) - |FN/(TP + FP) - FP/(TP + FP)|$

The next five slides present the comparative out-of-sample accuracy of discriminant analysis and stochastic ensemble methods in these datasets.

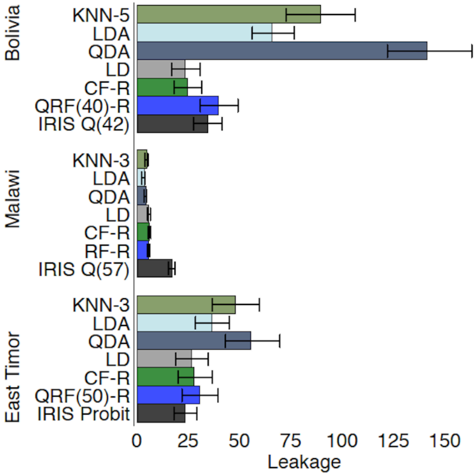
Total Accuracy



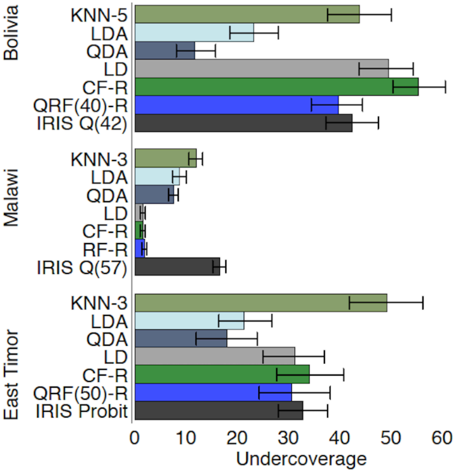
Poverty Accuracy



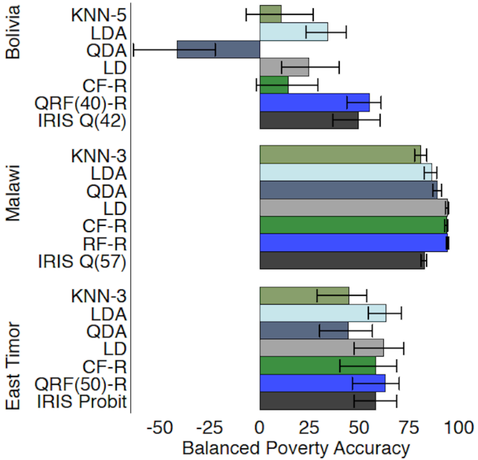
Leakage


Bolivia
KNN-5
LDA
QDA
LD
CF-R
QRF(40)-R
IRIS Q(42)
Malawi
KNN-3
LDA
QDA
LD
CF-R
RF-R
IRIS Q(57)
East Timor
KNN-3
LDA
QDA
LD
CF-R
QRF(50)-R
IRIS Probit
0
25
50
75
100
125
150
Leakage

Undercoverage



Balanced Poverty Accuracy



Other applications

Stochastic ensemble methods in general, and random forests in particular, have become essential tools in a variety of applications.

Bioinformatics: In comparison with DL, KNN, and SVM, Diaz-Uriarte and Alvarez de Andres (2006) conclude, "because of its performance and features, random forest and gene selection using random forest should probably become part of the 'standard tool-box' of methods for class prediction and gene selection with microarray data."

Kaggle and predictive analytics: the following kaggle competitions were won using random forests

- Semi-supervised feature learning (computer science)
- Air quality prediction (environmental science)
- RTA freeway travel time prediction (urban development/economics)

Other applications: remote sensing, diagnostics, spam filters

Conclusion

Stochastic ensemble methods have broad applicability to classification and prediction problems; we find their use promising in poverty assessment tool development.

Such methods would be additional assets in the Stata classification tool kit.

Breiman, L. 1996a. Bagging predictors. *Machine Learning*, 26(2):123-140.

Breiman, L. 1996b. Out of bag estimation.
ftp.stat.berkeley.edu/pub/users/breiman/OOBestimation.ps

Breiman, L. 2001. Random forests. *Machine Learning*, 45: 5-32.

Breiman. L. and A. Cutler. 2007. Random Forests.
www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Cutler, A. and G. Zhao. 2001. PERT-Perfect random tree ensembles. *Computing Science and Statistics*, 33.

Diaz-Uriarte, R., Alvarez de Andres, S. 2006. Gene selection and classification of microarray data using random forest. BMC *Bioinformatics*, 7:3.

Huberty, C. 1994. *Applied Discriminant Analysis*. New York: Wiley.

Hastie, T., R. J. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd Ed. New York: Springer.

Liaw, A. and M. Wiener. 2002. Classification and regression by randomForest. *R News*, 2:18-22.

Meinshausen, N. 2006. Quantile regression forests. *Journal of Machine Learning Research*, 7: 983-999.

Rencher, A. 2002. *Methods of Multivariate Analysis*, 2nd Ed. New York: Wiley.

Classification error

For a two-group classification problem, when misclassification costs are equal,

$$MSE = \frac{1}{N}\sum_{i=0}^{n}(\hat{P}_i - P_i)^2 = \frac{1}{N}(FN + FP)$$

	$P = 1$	$P = 0$
$\hat{P} = 1$	True Positive (TP)	False Positive (FP)
$\hat{P} = 0$	False Negative (FN)	True Negative (TN)