用R实现随机森林的分类与回归
Applications of Random Forest using R Classification and Regression

李欣海
中科院动物所

邮件：lixh@ioz.ac.cn
主页：http://people.gucas.ac.cn/~LiXinhai
博客：http://blog.sciencenet.cn/u/lixinhai
微博：http://weibo.com/lixinhaiblog
Random Forest

- Random Forest is an ensemble classifier that consists of many decision trees.
- It outputs the class that is the mode of the class's output by individual trees (Breiman 2001).
- It deals with “small n large p”-problems, high-order interactions, correlated predictor variables.

History

The algorithm for inducing a random forest was developed by Leo Breiman (2001) and Adele Cutler, 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 (1995) and Amit and Geman (1997) in order to construct a collection of decision trees with controlled variation.
Tree models

\[ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \]

Regression tree

(Classification tree)

(Crawley 2007 *The R Book* p691)

(Crawley 2007 *The R Book* p694)
Statistical Modeling: The Two Cultures
Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.
Ensemble classifiers

Tree models are simple, often produce noisy (bushy) or weak (stunted) classifiers.

• Bagging (Breiman, 1996): Fit many large trees to bootstrap-resampled versions of the training data, and classify by majority vote.

• Boosting (Freund & Shapire, 1996): Fit many large or small trees to reweighted versions of the training data. Classify by weighted majority vote.

• Random Forests (Breiman 1999): Fancier version of bagging.

In general Boosting > Random Forests > Bagging > Single Tree (Trevor Hastie).
How Random Forest Works

• At each tree split, a random sample of $m$ features is drawn, and only those $m$ features are considered for splitting. Typically $m = \sqrt{p}$ or $\log(p)$, where $p$ is the number of features.

• For each tree grown on a bootstrap sample, the error rate for observations left out of the bootstrap sample is monitored. This is called the out-of-bag (OOB) error rate.

• Random forests tries to improve on bagging by “de-correlating” the trees. Each tree has the same expectation.

(Trevor Hastie, p21 in *Trees, Bagging, Random Forests and Boosting*)
R Packages

**randomForest** \texttt{randomForest()}
Title: Breiman and Cutler’s random forests for classification and regression
Version: 4.6-6
Date: 2012-01-06
Author: Fortran original by Leo Breiman and Adele Cutler, R port by Andy Liaw and Matthew Wiener.

Implementation based on CART trees for variables of different types.
Biased in favor of continuous variables and variables with many categories.

**party** \texttt{cforest()}
Based on unbiased conditional inference trees.
For variables of different types: unbiased when subsampling.
朱鹮的分布
### Data

<table>
<thead>
<tr>
<th>use</th>
<th>x</th>
<th>y</th>
<th>Elev</th>
<th>Aspect</th>
<th>Slope</th>
<th>Land cover</th>
<th>Pop</th>
<th>Footprint</th>
<th>GDP</th>
<th>prec_ann</th>
<th>prec_jan</th>
<th>prec_july</th>
<th>t_ann</th>
<th>t_jan</th>
<th>t_july</th>
<th>year</th>
<th>Nest site</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>107.505</td>
<td>33.392</td>
<td>984</td>
<td>0.67</td>
<td>29.6</td>
<td>21</td>
<td>42.0</td>
<td>20</td>
<td>2.95</td>
<td>845</td>
<td>6</td>
<td>153</td>
<td>12.4</td>
<td>0.3</td>
<td>24.0</td>
<td>1981</td>
<td>金家沟</td>
</tr>
<tr>
<td>1</td>
<td>107.548</td>
<td>33.409</td>
<td>1315</td>
<td>0.90</td>
<td>19.0</td>
<td>14</td>
<td>22.5</td>
<td>26</td>
<td>1.97</td>
<td>869</td>
<td>6</td>
<td>157</td>
<td>11.3</td>
<td>-0.6</td>
<td>22.7</td>
<td>1981</td>
<td>金家沟</td>
</tr>
<tr>
<td>1</td>
<td>107.505</td>
<td>33.392</td>
<td>984</td>
<td>0.67</td>
<td>29.6</td>
<td>21</td>
<td>42.0</td>
<td>20</td>
<td>2.95</td>
<td>845</td>
<td>6</td>
<td>153</td>
<td>12.4</td>
<td>0.3</td>
<td>24.0</td>
<td>1982</td>
<td>金家沟</td>
</tr>
<tr>
<td>1</td>
<td>107.548</td>
<td>33.409</td>
<td>1315</td>
<td>0.90</td>
<td>19.0</td>
<td>14</td>
<td>22.5</td>
<td>26</td>
<td>1.97</td>
<td>869</td>
<td>6</td>
<td>157</td>
<td>11.3</td>
<td>-0.6</td>
<td>22.7</td>
<td>1982</td>
<td>姚家沟</td>
</tr>
<tr>
<td>1</td>
<td>107.505</td>
<td>33.392</td>
<td>984</td>
<td>0.67</td>
<td>29.6</td>
<td>21</td>
<td>42.0</td>
<td>20</td>
<td>2.95</td>
<td>845</td>
<td>6</td>
<td>153</td>
<td>12.4</td>
<td>0.3</td>
<td>24.0</td>
<td>1983</td>
<td>金家沟</td>
</tr>
<tr>
<td>1</td>
<td>107.548</td>
<td>33.409</td>
<td>1315</td>
<td>0.90</td>
<td>19.0</td>
<td>14</td>
<td>22.5</td>
<td>26</td>
<td>1.97</td>
<td>869</td>
<td>6</td>
<td>157</td>
<td>11.3</td>
<td>-0.6</td>
<td>22.7</td>
<td>1983</td>
<td>姚家沟</td>
</tr>
<tr>
<td>1</td>
<td>107.505</td>
<td>33.392</td>
<td>984</td>
<td>0.67</td>
<td>29.6</td>
<td>21</td>
<td>42.0</td>
<td>20</td>
<td>2.95</td>
<td>845</td>
<td>6</td>
<td>153</td>
<td>12.4</td>
<td>0.3</td>
<td>24.0</td>
<td>1984</td>
<td>金家沟</td>
</tr>
<tr>
<td>1</td>
<td>107.548</td>
<td>33.409</td>
<td>1315</td>
<td>0.90</td>
<td>19.0</td>
<td>14</td>
<td>22.5</td>
<td>26</td>
<td>1.97</td>
<td>869</td>
<td>6</td>
<td>157</td>
<td>11.3</td>
<td>-0.6</td>
<td>22.7</td>
<td>1984</td>
<td>姚家沟</td>
</tr>
<tr>
<td>1</td>
<td>107.405</td>
<td>33.406</td>
<td>1056</td>
<td>0.54</td>
<td>11.4</td>
<td>21</td>
<td>0.0</td>
<td>20</td>
<td>0.98</td>
<td>892</td>
<td>7</td>
<td>161</td>
<td>11.4</td>
<td>-0.5</td>
<td>22.9</td>
<td>1984</td>
<td>三岔河</td>
</tr>
<tr>
<td>1</td>
<td>107.405</td>
<td>33.406</td>
<td>1056</td>
<td>0.54</td>
<td>11.4</td>
<td>21</td>
<td>0.0</td>
<td>20</td>
<td>0.98</td>
<td>892</td>
<td>7</td>
<td>161</td>
<td>11.4</td>
<td>-0.5</td>
<td>22.9</td>
<td>1985</td>
<td>三岔河</td>
</tr>
<tr>
<td>1</td>
<td>107.548</td>
<td>33.409</td>
<td>1315</td>
<td>0.90</td>
<td>19.0</td>
<td>14</td>
<td>22.5</td>
<td>26</td>
<td>1.97</td>
<td>869</td>
<td>6</td>
<td>157</td>
<td>11.3</td>
<td>-0.6</td>
<td>22.7</td>
<td>1985</td>
<td>姚家沟</td>
</tr>
</tbody>
</table>

```
> table(ibis$use)
   0   1
2538 560
```
Multicollinearity is a pain
Variables in the two-principal-component space

biplot(princomp(ibis[,2:16], cor=T))
Variance inflation factors (VIF) to detect multicollinearity

```r
library(car)
vif(lm(pop ~ x+y+elevation+slope+aspect+footprint+GDP+
    prec_ann+prec_jan+prec_july+t_ann+t_jan+t_july,
    data=ibis))
```

<table>
<thead>
<tr>
<th></th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>6.86</td>
</tr>
<tr>
<td>y</td>
<td>82.67</td>
</tr>
<tr>
<td>elevation</td>
<td>9.82</td>
</tr>
<tr>
<td>aspect</td>
<td>1.01</td>
</tr>
<tr>
<td>footprint</td>
<td>1.38</td>
</tr>
<tr>
<td>GDP</td>
<td>1.32</td>
</tr>
<tr>
<td>prec_ann</td>
<td>207.97</td>
</tr>
<tr>
<td>prec_jan</td>
<td>6.24</td>
</tr>
<tr>
<td>slope</td>
<td>1.50</td>
</tr>
<tr>
<td>t_ann</td>
<td>15170.35</td>
</tr>
<tr>
<td>prec_july</td>
<td>170.89</td>
</tr>
<tr>
<td>t_jan</td>
<td>4309.17</td>
</tr>
<tr>
<td>t_july</td>
<td>5132.72</td>
</tr>
</tbody>
</table>
Logistic regression

fit <- glm(use ~ ., data=ibis, family=binomial())
summary(fit)
step(fit)

Step: AIC=1710.33
use ~ x + y + elevation + aspect + slope +
    landcover + pop + footprint + GDP + prec_ann +
    prec_jan + prec_july + t_ann + t_jan

library(epicalc)
lroc(fit, title=TRUE,
auc.coords=c(.5,.1))

Area under the curve = 0.913
randomForest()

library(randomForest)
ibis <- ibis[-c(17,18)]
RF <- randomForest(ibis[-1], ibis[,1],
                   proximity=TRUE, importance=TRUE)
summary(RF)

type     | predicted   | err.rate   | confusion | votes   | oob.times | classes | importance | importanceSD | ntree   | mtry    | forest | y      | test | inbag
--- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | ---
call    | 3098         | 1500       | 6        | 6196    | 3098      | 2       | 60          | 45          | 1       | 1       | 14     | 3098   | 0    | 0
character | factor       | numeric    | numeric  | numeric  | numeric    | character | numeric    | numeric    | numeric  | numeric  | list   | numeric  | NULL | NULL | NULL | NULL
imp <- importance(RF)
impvar <- imp[order(imp[, 3], decreasing=TRUE), ]; impvar

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>MeanDecreaseAccuracy</th>
<th>MeanDecreaseGini</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>0.544507</td>
<td>1.3917066</td>
<td>0.6081245</td>
<td>179.819331</td>
</tr>
<tr>
<td>x</td>
<td>0.455705</td>
<td>1.3866419</td>
<td>0.594095</td>
<td>126.056688</td>
</tr>
<tr>
<td>slope</td>
<td>0.169391</td>
<td>1.2966797</td>
<td>0.5587788</td>
<td>39.304633</td>
</tr>
<tr>
<td>prec_ann</td>
<td>0.324172</td>
<td>1.2131164</td>
<td>0.5293486</td>
<td>55.937209</td>
</tr>
<tr>
<td>GDP</td>
<td>0.393354</td>
<td>1.1838283</td>
<td>0.5233282</td>
<td>42.083964</td>
</tr>
<tr>
<td>prec_july</td>
<td>0.366197</td>
<td>1.166394</td>
<td>0.521239</td>
<td>51.248009</td>
</tr>
<tr>
<td>landcover</td>
<td>0.250204</td>
<td>1.1888627</td>
<td>0.5125181</td>
<td>84.391889</td>
</tr>
<tr>
<td>t_jan</td>
<td>0.406161</td>
<td>1.1651366</td>
<td>0.5096002</td>
<td>61.646713</td>
</tr>
<tr>
<td>elevation</td>
<td>0.380496</td>
<td>1.1895061</td>
<td>0.5053193</td>
<td>69.867269</td>
</tr>
<tr>
<td>t_ann</td>
<td>0.383974</td>
<td>1.150548</td>
<td>0.5048529</td>
<td>63.471821</td>
</tr>
<tr>
<td>pop</td>
<td>0.423507</td>
<td>1.0874984</td>
<td>0.4981203</td>
<td>28.297228</td>
</tr>
<tr>
<td>aspect</td>
<td>-0.0799</td>
<td>1.1987937</td>
<td>0.4809901</td>
<td>22.272998</td>
</tr>
<tr>
<td>t_july</td>
<td>0.352301</td>
<td>1.1058594</td>
<td>0.4796287</td>
<td>59.880941</td>
</tr>
<tr>
<td>footprint</td>
<td>-0.05375</td>
<td>0.9224068</td>
<td>0.3773902</td>
<td>15.344756</td>
</tr>
<tr>
<td>prec_jan</td>
<td>0.228913</td>
<td>0.8092592</td>
<td>0.344224</td>
<td>8.766087</td>
</tr>
</tbody>
</table>
varImpPlot(RF)

MeanDecreaseAccuracy

MeanDecreaseGini

- y
- x
- slope
- prec_ann
- GDP
- prec_july
- landcover
- t_jan
- elevation
- t_ann
- t_jan
- t_july
- prec_ann
- prec_july
- GDP
- slope
- pop
- aspect
- t_july
- footprint
- prec_jan

16/25 随机森林：分类 第五届中国R语言会议 北京2012
partialPlot: partial dependence of elevation

\begin{verbatim}
ibis=ibis[order(ibis$x),]
plot(ibis$elevation,
    col=4-as.numeric(ibis$use),
    cex=0.5, pch=4)

ibis=ibis[order(ibis$elevation),]
plot(ibis$elevation,
    col=4-as.numeric(ibis$use),
    cex=0.5, pch=4)

partialPlot(RF, ibis, elevation, "0",
    main='Absence',xlab='Elevation')

partialPlot(RF, ibis, elevation, "1",
    main='Presence',xlab='Elevation')
\end{verbatim}
Model comparison

Predicted current suitable habitat of crested ibis using the models in BIOMOD
(The warm color areas are the suitable areas)
Predicted current suitable habitat of black snub-nose monkey using BIOMOD (The warm color areas are the suitable areas)
Historical decline of Asian elephant

Year
- 211
- 212 - 500
- 501 - 600
- 601 - 700
- 701 - 800
- 801 - 900
- 901 - 1000
- 1001 - 1100
- 1101 - 1200
- 1201 - 1300
- 1301 - 1400
- 1401 - 1500
- 1501 - 1600
- 1601 - 1700
- 1701 - 1800
- 1801 - 1900
- 1901 - 1940
- 1941 - 1960
- 1961 - 1980
- 1981 - 2000
Variables associated with species range

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>34.42</td>
<td>-0.52</td>
<td>-0.41</td>
<td>-0.49</td>
<td>-0.59</td>
<td>-0.22</td>
<td>-0.07</td>
<td>-0.10</td>
<td>-5.32</td>
<td>16.44</td>
</tr>
<tr>
<td>31.57</td>
<td>-0.41</td>
<td>-0.21</td>
<td>-0.35</td>
<td>0.00</td>
<td>-0.02</td>
<td>-0.35</td>
<td>-2.12</td>
<td>-6.37</td>
<td>16.44</td>
</tr>
<tr>
<td>29.53</td>
<td>0.28</td>
<td>-0.23</td>
<td>-0.25</td>
<td>-0.21</td>
<td>-0.11</td>
<td>-0.78</td>
<td>-3.76</td>
<td>-2.19</td>
<td>17.74</td>
</tr>
<tr>
<td>31.87</td>
<td>0.35</td>
<td>-0.25</td>
<td>-0.29</td>
<td>-0.27</td>
<td>0.02</td>
<td>-0.34</td>
<td>-3.24</td>
<td>-5.07</td>
<td>16.31</td>
</tr>
<tr>
<td>28.93</td>
<td>-0.13</td>
<td>-0.09</td>
<td>-0.31</td>
<td>-0.62</td>
<td>-0.16</td>
<td>0.11</td>
<td>-2.38</td>
<td>-1.79</td>
<td>16.93</td>
</tr>
<tr>
<td>34.79</td>
<td>0.19</td>
<td>0.09</td>
<td>-0.13</td>
<td>-0.38</td>
<td>0.14</td>
<td>0.67</td>
<td>7.30</td>
<td>22.57</td>
<td>16.60</td>
</tr>
</tbody>
</table>

RF <- randomForest(lat.max ~ ., data=dd, ntree=1000, importance=TRUE)
imp <- importance(RF)
impvar <- rownames(imp)[order(imp[, 1], decreasing=TRUE)]  #sort importance

# Plot partial effects
op <- par(mfrow=c(3, 3), mar=c(4,4,2,2))
for (i in seq_along(impvar)) {
  partialPlot(RF, dd, impvar[i], xlab=impvar[i], #Partial effects
              ylab='Longitude', ylim=c(26,30), main="")
}
Partial effect of variables on maximum latitude

- Population
- Temp. Mann. eiv. cru
- Temp. Mann. cps. cru
- Flood
- Temp. Ljungqvist
- Precipitation
- Temp. Yang
- Drought
- Temp. Moberg
### Popular species distribution models, history, complexity levels, popularity in climate change studies, types of species data, and reference papers

<table>
<thead>
<tr>
<th>Models</th>
<th>History</th>
<th>Complexity</th>
<th>Popularity</th>
<th>Species data</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generalized linear model (GLM)</td>
<td>1972</td>
<td>low</td>
<td>33/66</td>
<td>p/a or abundance</td>
<td>(Nelder &amp; Wedderburn 1972)</td>
</tr>
<tr>
<td>Generalized additive model (GAM)</td>
<td>1986</td>
<td>medium</td>
<td>28/86</td>
<td>p/a or abundance</td>
<td>(Hastie &amp; Tibshirani 1986)</td>
</tr>
<tr>
<td>Multivariate Adaptive Regression Splines (MARS)</td>
<td>1991</td>
<td>medium</td>
<td>13/56*</td>
<td>p/a</td>
<td>(Friedman 1991)</td>
</tr>
<tr>
<td>Mixture discriminant analysis (MDA)</td>
<td>1996</td>
<td>medium</td>
<td>4/9</td>
<td>p/a</td>
<td>(Hastie &amp; Tibshirani 1996)</td>
</tr>
<tr>
<td>Classification and Regression Tree (CART)</td>
<td>1984</td>
<td>medium</td>
<td>16/23</td>
<td>p/a</td>
<td>(Breiman et al. 1984)</td>
</tr>
<tr>
<td>Generalized Boosting Models (GBM)</td>
<td>1999</td>
<td>medium</td>
<td>0/14</td>
<td>p/a</td>
<td>(Friedman et al. 2000)</td>
</tr>
<tr>
<td>Random Forest</td>
<td>1995</td>
<td>high</td>
<td>26</td>
<td>p/a</td>
<td>(Breiman 2001a)</td>
</tr>
<tr>
<td>Artificial neural networks (ANN)</td>
<td>1943</td>
<td>high</td>
<td>96/75</td>
<td>p/a</td>
<td>(Hopfield 1982)</td>
</tr>
<tr>
<td>Maximum entropy method (Maxent)</td>
<td>2006</td>
<td>high</td>
<td>5/125</td>
<td>p/a or abundance</td>
<td>(Phillips et al. 2006)</td>
</tr>
<tr>
<td>Hierarchical modeling</td>
<td>1996</td>
<td>low</td>
<td>13</td>
<td>p/a or abundance</td>
<td>(Wikle 2003)</td>
</tr>
</tbody>
</table>
The advantages of Random Forest

• For many data sets, it produces a highly accurate classifier
• It handles a very large number of input variables
• It estimates the importance of variables in determining classification
• It generates an internal unbiased estimate of the generalization error as the forest building progresses
• It includes a good method for estimating missing data and maintains accuracy when a large proportion of the data are missing
• It provides an experimental way to detect variable interactions
• It can balance error in class population unbalanced data sets
• It computes proximities between cases, useful for clustering, detecting outliers, and (by scaling) visualizing the data
• Using the above, it can be extended to unlabeled data, leading to unsupervised clustering, outlier detection and data views
• Learning is fast
The disadvantages of Random Forest

• Random forests are prone to overfitting for some datasets. This is even more pronounced in noisy classification/regression tasks.

• Random forests do not handle large numbers of irrelevant features as well as ensembles of entropy-reducing decision trees.

• It is more efficient to select a random decision boundary than an entropy-reducing decision boundary, thus making larger ensembles more feasible. Although this may seem to be an advantage at first, it has the effect of shifting the computation from training time to evaluation time, which is actually a disadvantage for most applications.
Try Random Forest!

Phone took at Hailuogou, Sichuan province in June 2007 by Xinhai Li