

36

Learn R | Random Forest of Data Mining (下)



Jason

[关注他](#)

36 人赞了该文章

接上文，在对决策树及随机森林有一个基本的了解后，我们继续深入学习随机森林算法。

一、随机森林效果的影响因素

随机森林的分类效果（即错误率）与以下两个因素有关（内容引自博客[\[Machine Learning & Algorithm\] 随机森林 \(Random Forest\)](#)）：

- 森林中任意两棵树的相关性：相关性越大，错误率越大
- 森林中每棵树的分类能力：每棵树的分类能力越强，整个森林的错误率越低

减小特征选择个数 m ，树的相关性和分类能力也会相应的降低；增大 m ，两者也会随之增大。所以关键问题是如何选择最优的 m （或者是范围），这也是随机森林唯一的一个参数。

二、袋外错误率 (oob error)

如何选择最优的特征个数 m ，要解决这个问题，我们主要依据计算得到的袋外错误率 **oob error** (out-of-bag error)。

随机森林有一个重要的优点就是，没有必要对它进行交叉验证或者用一个独立的测试集来获得误差的一个无偏估计。它可以在内部进行评估，也就是说在生成的过程中就可以对误差建立一个无偏估计。



样本。

袋外错误率 (**oob error**) 计算方式如下：

1. 对每个样本计算它作为oob样本的树对它的分类情况
2. 以简单多数投票作为该样本的分类结果
3. 最后用误分个数占样本总数的比率作为随机森林的oob误分率

三、随机森林的特点

- 在当前所有算法中，具有极好的准确率
- 能够有效地运行在大数据集上
- 能够处理具有高维特征的输入样本，而且不需要降维
- 能够评估各个特征在分类问题上的重要性
- 在生成过程中，能够获取到内部生成误差的一种无偏估计
- 对于缺省值问题也能够获得很好得结果
-

实际上，随机森林的特点不只有这六点，它就相当于机器学习领域的Leatherman（多面手），你几乎可以把任何东西扔进去，它基本上都是可供使用的。

（该节内容同样引自博客[\[Machine Learning & Algorithm\] 随机森林 \(Random Forest \)](https://zhuanlan.zhihu.com/p/24416833)）

四、随机森林算法的R实现

在R语言中，我们调用randomForest包中的randomForest()函数来实现随机森林算法，该函数中的决策树基于基尼指数（Gini index）构建，即CART分类决策树。不过该函数有两点不足：第一，它不能处理缺失值，如果数据集有缺失值的话，我们必须在使用该函数之前填补；第二，每个分类属性的最大数量不能超过32个，如果属性超过32个，那么在使用randomForest()之前那些属性必须被转化。

```
> install.packages("randomForest")  
> library(randomForest)
```



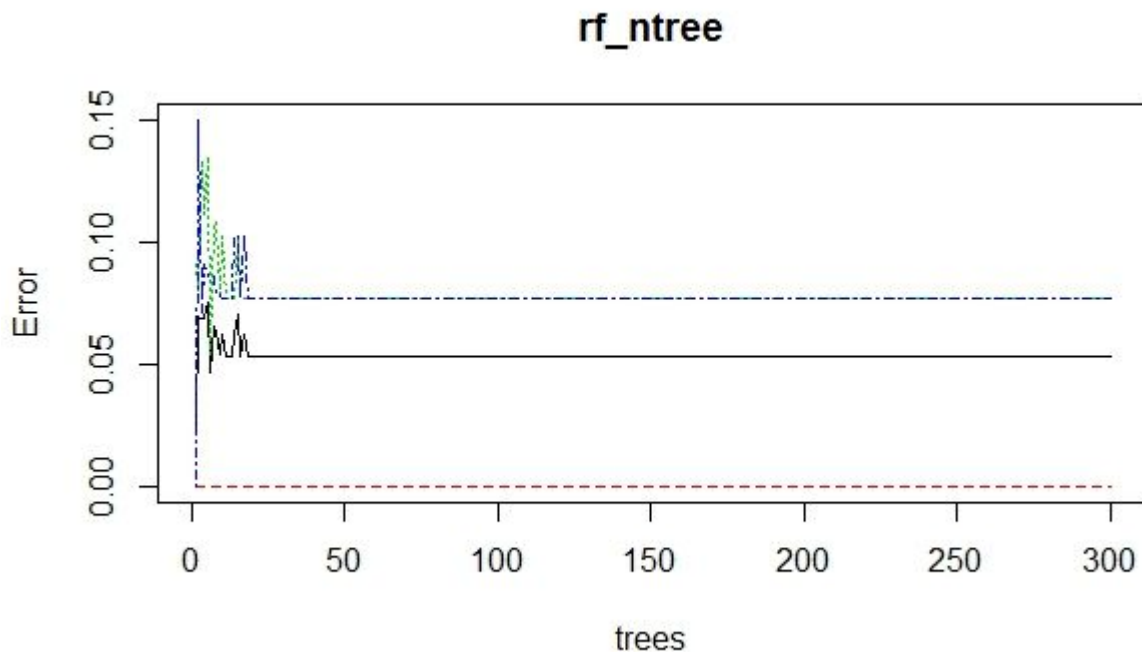


```
> index <- sample(2,nrow(iris),replace = TRUE,prob=c(0.7,0.3))
traindata <- iris[index==1,]
testdata <- iris[index==2,]
```

在使用模型前，我们需要确定，要生成多少棵树构建森林，即模型中ntree参数的具体值，可通过图形大致判断模型内误差稳定时的值。

分享

```
> set.seed(1234)
> rf_ntree <- randomForest(Species~.,data=traindata,ntree=300)
> plot(rf_ntree)
```



从图中可以看到，当ntree=50时，模型内的误差就基本稳定了，出于更保险的考虑，我们确定ntree值为100。

```
> iris_rf <- randomForest(Species ~ ., data=traindata, ntree=100,
+ proximity=TRUE)
# ntree: 指定随机森林所包含的决策树数目，默认为500
# proximity: 逻辑参数，是否计算模型的临近矩阵，主要结合MDSplot()函数使用
```

```
# 模型解读
```

```
> iris_rf
```

36

19 条评论

分享

收藏

...

Type of random forest: classification

Number of trees: 100

No. of variables tried at each split: 2

36 OOB estimate of error rate: 7.22%

Confusion matrix:

	setosa	versicolor	virginica	class.error
setosa	31	0	0	0.00000000
versicolor	0	32	3	0.08571429
virginica	0	4	27	0.12903226

从上面对模型的解释中，我们可以看到该模型的oob error为7.22%，下面的矩阵表明第二类和第三类存在误判。

```
# 绘制每一棵树的误判率
> plot(iris_rf)
```

```
# importance()函数用于计算模型变量的重要性
> importance(iris_rf)
```

	MeanDecreaseGini
Sepal.Length	7.445145
Sepal.Width	1.692992



```
> varImpPlot(iris_rf)
```

36

从返回的数据和图形可知，在四个变量中，Petal.Width和Petal.Length最为重要，其次分别是Sepal.Length和Sepal.Width。

```
# treesize: 计算随机森林中每棵树的节点个数
> head(treesize(iris_rf,terminal = TRUE))
[1] 6 6 7 7 5 9
> count_data <- as.data.frame(plyr::count(treesize(iris_rf,
+ terminal = TRUE)))
> head(count_data,5)
  x freq
1  4    2
2  5   10
3  6   21
4  7   27
5  8   18

# 使用rCharts包rPlot()函数进行交互可视化展示
> rPlot(x='bin(x,1)',y='freq',data=count_data,type='bar')
```



36

接着，我们使用已建立好的随机森林模型进行新数据集的测试

```
> iris_pred <- predict(iris_rf,newdata=testdata)
> table(iris_pred,testdata$Species)
```

iris_pred	setosa	versicolor	virginica
setosa	19	0	0
versicolor	0	14	1
virginica	0	1	18

绘制测试数据的预测边距图，数据点的边距为正确归类后的比例减去被归到其他类别的最大比例。一般来说，边距为正数说明该数据点划分正确。

```
> plot(margin(iris_rf, testdata$Species))
```



36

至此，一个简单的随机森林模型的R实现就完成了。不过最后还需要补充一点，在第一节随机森林效果的影响因素中，我们提到在构建随机森林模型时，模型的准确性如何，关键在于选择最优的变量个数 m （参数'mtry'），在上面的例子中，我们使用的iris数据集使用了函数中该参数的默认情况（数据集变量个数的二次方根（分类模型）或三分之一（预测模型））。但一般情况下需要进行人为的逐次挑选，以确定最佳的 m 值。

这里借鉴博客基于R语言的随机森林算法运用的做法：

```
> n <- length(names(traindata))
> set.seed(1234)
> for (i in 1:(n-1)){
+   model <- randomForest(Species~., data = traindata, mtry = i)
+   err <- mean(model$err.rate)
+   print(err)
+ }
[1] 0.05342314
[1] 0.0513293
[1] 0.05110437
[1] 0.05122018
```

观察输出结果，当mtry=3时，模型内平均误差最小。

```
# 指定mtry=3，重新构建模型
> iris_rf_1 <- randomForest(Species ~ ., data=traindata,mtry=3, ntree=100, proxim
> iris_rf_1
```

Call:

```
randomForest(formula = Species ~ ., data = traindata, mtry = 3,          ntree = 100
              Type of random forest: classification
              Number of trees: 100
              No. of variables tried at each split: 3
```



	setosa	versicolor	virginica	class.error
setosa	31	0	0	0.00000000
versicolor	0	32	3	0.08571429
virginica	0	3	28	0.09677419

36

我们发现，当mtry=3时，随机森林的袋外错误率由7.22%下降至6.19%，错判数减少1个。这说明我们通过指定合适的mtry值实现了模型准确度的提升。

References :

- 1. [R Random Forest](#)
- 2. [\[Machine Learning & Algorithm\] 随机森林 \(Random Forest \)](#)
- 3. [Random Forest Variable Importance](#)
- 4. [R语言随机森林 - R语言教程](#)
- 5. [决策树和随机森林](#)
- 6. [Random forests - classification description](#)
- 7. [基于R语言的随机森林算法运用](#)
- 8. [分类回归树与随机森林 \(R语言 \)](#)

编辑于 2017-01-01

「真诚赞赏，手留余香」

赞赏

1 人已赞赏



R (编程语言) 机器学习 数据挖掘

文章被以下专栏收录



推荐阅读

36



Learn R | Random Forest of Data Mining (上)

Jason 发表于数据科学笔...



Learn R | Association Rules of Data Mining (二)

Jason 发表于数据科学笔...



L
o
Ji

19 条评论

切换为时间排序

写下你的评论...



卢莫科

1 年前

oob和obb ~

赞



Jason (作者) 回复 卢莫科

1 年前

谢谢提醒，已修改~

赞 查看对话



Mingjian Tian

1 年前

Iris_rf那张图有4条曲线，没看明白那是什么东西，或什么error，以及所说的第二类第三类是什么意思？谢谢。

赞



Jason (作者) 回复 Mingjian Tian

1 年前

这张图我在学习的时候就有些困惑，我的理解是应该指的是setosa、versicolor 和virginica三类数据的误判情况，而且class.error的值正好与其中的三条线相吻合，但为什么会是四条线我暂时也




👍 赞 💬 查看对话

 宇文

1 年前

我用自己的数据模仿里面的代码，挑选最优的mtry，不行呀

36 👍 赞

 任逍遥

12 个月前

请问有没有matlab的算法？谢谢！

👍 赞

 Jason (作者) 回复 任逍遥

12 个月前

不好意思啊，我没有用过matlab～

👍 赞 💬 查看对话

 xiaokekd 回复 Jason (作者)

11 个月前

我猜三条虚线是三个类的error，实线应该是一个综合的error，可能取得是三者的均值，我用两个分类结果的模型，就有两条虚线和一条直线。

👍 赞 💬 查看对话

 任逍遥 回复 Jason (作者)

10 个月前

仔细看了一下，那四条线，三条虚线为三个类别的分类错误率，实线为总体的错误率。另外，运行的时候出现“> count_data <- as.data.frame(ply::count(treesize(iris_rf,terminal = TRUE)))
Error in loadNamespace(name): 不存在叫‘plyr’这个名字的程辑包”，初学求指教

👍 赞 💬 查看对话

 Jason (作者) 回复 任逍遥

10 个月前

关于错误率图，我和你的观点差不多，应该就是总体的错误率情况～
另外，plyr包并不是R自带的，你需要去安装一下才能使用～


👍 赞 💬 查看对话

 任逍遥 回复 Jason (作者)

10 个月前

感谢您的回答，另外有个问题想请教，对于这些聚类算法，如果有一个变量的时间序列数据（60000个点），或者说波动的曲线，该提取什么量作为其特征作为输入。您有什么建议？再次感谢

👍 赞 💬 查看对话

 张雪婷

10 个月



👍 赞



vira85

10 个月前

想问一下 用plot函数绘图的时候 只要输入代码plot(随机森林的名称)就可以吗 我试了好几次 没有图
36 像输出哎..

👍 赞



王俊婷

9 个月前

你好，非常感谢你的分享，我想问一下回归怎么做，一直显示more than 53 categories，但是我的因变量是连续性数据啊，不是可以直接就做回归的么？单目前看来r一直将他看作分类变量处理

👍 赞



Skylar

8 个月前

二，每个分类属性的最大数量不能超过32个，如果属性超过32个，那么在使用randomForest()之前那些属性必须被转化。

请问这步转化怎么做

👍 赞



林晓池Smile 回复 Skylar

5 个月前

降维啦

👍 赞 💬 查看对话



闻道暮东 回复 Jason (作者)

4 个月前

iris_rf\$err.rate是图形的数据源。可以给图形价格legend就明白了

👍 赞 💬 查看对话



闻道暮东 回复 vira85

4 个月前

iris_rf\$err.rate看下有没有数据。单独plot就出结果的

👍 赞 💬 查看对话



辉常尊是呆瓜

28 天前

请问有人能理解为什么 predict 模型+训练数据 结果正确率总是100%吗？我想计算训练集、验证集与测试集的ROC，我先用 predict (type="prob") 计算出分类的概率，但是只要用训练集去 predict结果就是100%，有人能讲解一下吗。

▲ 生生

