RandomForest 应用于 iris 数据集

lucifercook

2017 年 8 月 8 日

机器学习：

Machine learning，机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能。算法包括神经网络，决策树，随机森林，支持向量机，朴素贝叶斯，逻辑回归等方法

决策树：

用决策树来划分物体的类属，树中每一内部节点对应一个物体属性，而每一边对应于这些属性的可选值，树的叶节点则对应于物体的每个基本分类。

随机森林：

Random Forest，是一种机器学习的算法，主要通过对训练集进行不放回抽样，建立多课决策树，然后对分类结果采用投票机制，以得票数最多的结果为准。是一种对决策树的改进算法，可避免过拟合问题。

导入系统自带的 iris 数据集

```
data(iris)
iris.data=iris
```

查看数据结构

```
str(iris.data)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1
1 1 1 1 1 1 1 ...
```

查看首尾数据

```
str(iris.data)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1
1 1 1 1 1 1 1 ...
```

```
head(iris.data)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

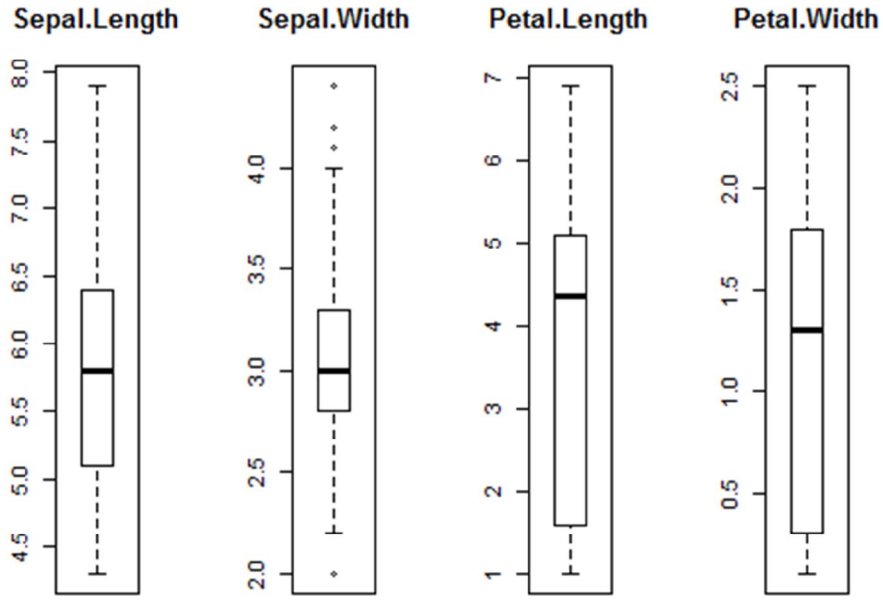
数据集摘要

```
summary(iris.data)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

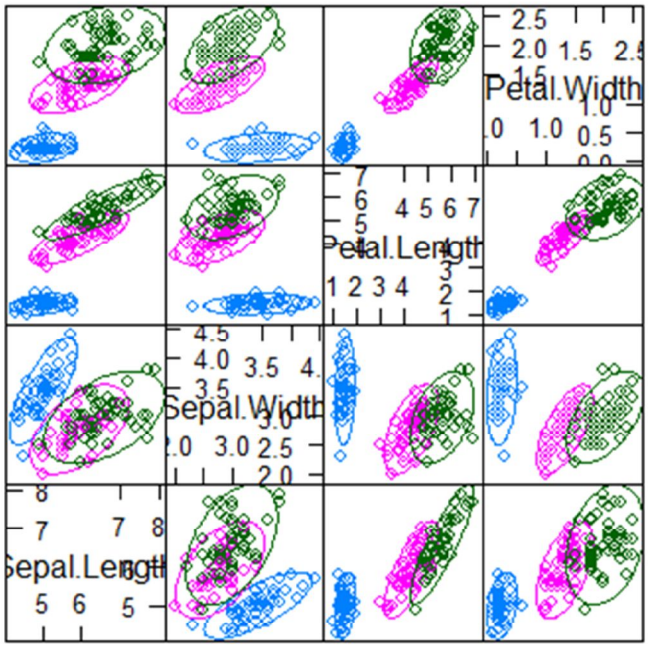
单变量可视化

```
input.val <- iris.data[,1:4]
par(mfrow=c(1,4))
for(i in 1:4) {
  boxplot(input.val[,i], main=names(iris.data)[i])
}
```



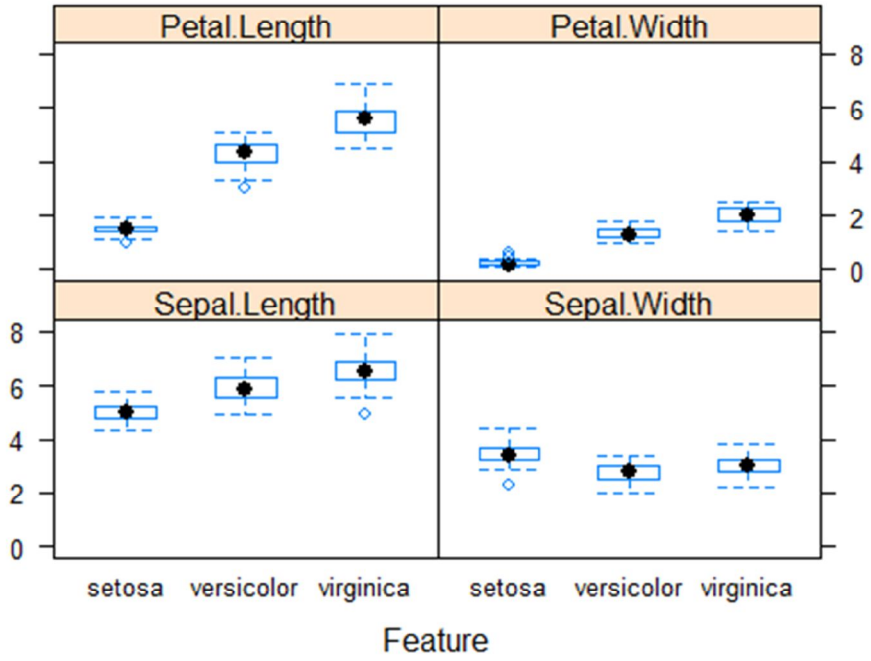
多变量可视化

```
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(ellipse)
output.val <- iris.data[,5]
featurePlot(x=input.val,y=output.val,plot='ellipse')
```



Scatter Plot Matrix

```
featurePlot(x=input.val,y=output.val,plot='box')
```



将样本划分训练集和数据集

```
validation.index<-createDataPartition(iris.data$Species,p=0.8,list=FALSE)
validation.data<-iris.data[-validation.index,]
train.data<-iris.data[validation.index,]
```

验证方式选择 10-折交叉验证

```
control<-trainControl(method="cv",number=10)
metric<-"Accuracy"
library(randomForest)

## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

library(e1071)

set.seed(7)
```

分别使用 LDA,CART,RF 算法建立模型

```
library(rpart)
rf.model<-train(Species~.,data=train.data,method="rf",metric=metric,trControl=control)
lda.model<-train(Species~.,data=train.data,method="lda",metric=metric,trControl=control)

## Loading required package: MASS

cart.model<-train(Species~.,data=train.data,method="rpart",metric=metric,trControl=control)
```

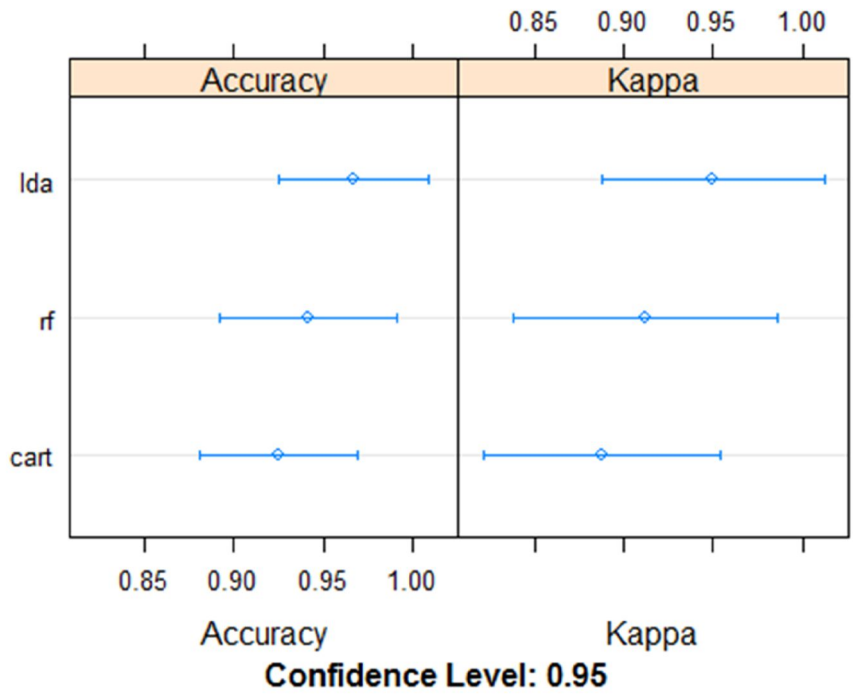
模型评价

```
results.model<-resamples(list(lda=lda.model,cart=cart.model,rf=rf.model))

summary(results.model)

##
## Call:
## summary.resamples(object = results.model)
##
## Models: lda, cart, rf
## Number of resamples: 10
```

```
##
## Accuracy
##           Min.   1st Qu.    Median      Mean   3rd Qu. Max. NA's
## lda  0.8333333 0.9375000 1.0000000 0.9666667 1.0000000    1    0
## cart 0.8333333 0.9166667 0.9166667 0.9250000 0.9791667    1    0
## rf   0.8333333 0.9166667 0.9583333 0.9416667 1.0000000    1    0
##
## Kappa
##      Min. 1st Qu. Median   Mean 3rd Qu. Max. NA's
## lda  0.75 0.90625 1.0000 0.9500 1.00000    1    0
## cart 0.75 0.87500 0.8750 0.8875 0.96875    1    0
## rf   0.75 0.87500 0.9375 0.9125 1.00000    1    0
```

结果可视化

```
dotplot(results.model)
```



结论：针对 iris 数据集，使用 lda 算法预测的准确率最高，RandomForest 居中

模型应用

```
pred.result<-predict(rf.model,validation.data)
confusionMatrix(pred.result,validation.data$Species)
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   setosa versicolor virginica
##   setosa         10          0         0
##   versicolor      0         10         1
##   virginica       0          0         9
##
## Overall Statistics
##
##                Accuracy : 0.9667
##                  95% CI : (0.8278, 0.9992)
##     No Information Rate : 0.3333
##     P-Value [Acc > NIR] : 2.963e-13
##
##                   Kappa : 0.95
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: setosa Class: versicolor Class: virginic
a
## Sensitivity                 1.0000            1.0000           0.900
0
## Specificity                 1.0000            0.9500           1.000
0
## Pos Pred Value              1.0000            0.9091           1.000
0
## Neg Pred Value              1.0000            1.0000           0.952
4
## Prevalence                  0.3333            0.3333           0.333
3
## Detection Rate              0.3333            0.3333           0.300
0
## Detection Prevalence        0.3333            0.3667           0.300
0
## Balanced Accuracy           1.0000            0.9750           0.950
0
```