用 R 进行 Titanic 生存预测

lucifercook

2017 年 8 月 22 日星期二

首先把从 kaggle 下载的训练和测试的数据读入 R 中

```
test<-read.csv("test.csv",stringsAsFactors = FALSE)
train<-read.csv("train.csv",stringsAsFactors = FALSE)
```

看下数据的大概情况

```
table(train$Survived)
```

```
##
##   0   1
## 549 342
```

```
test$Survived<-rep(0,418)
test$Survived<-0
test$Survived[test$Sex=='female']<-1
train$Child<-0
train$Child[train$Age<18]<-1
aggregate(Survived~Child + Sex,data = train,FUN=sum)
```

```
##   Child    Sex Survived
## 1     0 female      195
## 2     1 female       38
## 3     0   male       86
## 4     1   male       23
```

```
test$Fare2<-'30+'
train$Fare2<-'30+'
train$Fare2[train$Fare<30&train$Fare>=20]<-'20-30'
train$Fare2[train$Fare<20&train$Fare>=10]<-'10-20'
train$Fare2[train$Fare<10]<-'<10'
str(test)
```

```
## 'data.frame':    418 obs. of  13 variables:
##  $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
##  $ Pclass     : int  3 3 2 3 3 3 3 2 3 3 ...
##  $ Name       : chr  "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen N
eeds)" "Myles, Mr. Thomas Francis" "Wirz, Mr. Albert" ...
##  $ Sex        : chr  "male" "female" "male" "male" ...
##  $ Age        : num  34.5 47 62 27 22 14 30 26 18 21 ...
##  $ SibSp      : int  0 1 0 0 1 0 0 1 0 2 ...
##  $ Parch      : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ Ticket     : chr  "330911" "363272" "240276" "315154" ...
```

```
##  $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
##  $ Cabin      : chr  "" "" "" "" ...
##  $ Embarked   : chr  "Q" "S" "Q" "S" ...
##  $ Survived   : num  0 1 0 0 1 0 1 0 1 0 ...
##  $ Fare2      : chr  "30+" "30+" "30+" "30+" ...

str(train)

## 'data.frame':    891 obs. of  14 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John B
radley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mr
s. Jacques Heath (Lily May Peel)" ...
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "1138
03" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
##  $ Child      : num  0 0 0 0 0 0 0 1 0 1 ...
##  $ Fare2      : chr  "<10" "30+" "<10" "30+" ...

test<-subset(test,select = -Fare2)
test$Survived<-NA
```

把 test 和 train 合并，方便一起处理，因为 test 没有 Survived 这列，需要添加,把数据类型转成随机森林可以用的类型

```
test<-read.csv("test.csv",stringsAsFactors = FALSE)
train<-read.csv("train.csv",stringsAsFactors = FALSE)
test$Survived<-NA
combi<-rbind(train,test)
combi$Name<-as.character(combi$Name)
combi$Name[1]

## [1] "Braund, Mr. Owen Harris"

strsplit(combi$Name[1],split='[,.]')

## [[1]]
## [1] "Braund"       " Mr"          " Owen Harris"

strsplit(combi$Name[1],split='[,.]')[[1]]

## [1] "Braund"       " Mr"          " Owen Harris"

strsplit(combi$Name[1],split='[,.]')[[1]][2]
```

```
## [1] " Mr"

combi$Title<-sapply(combi$Name,FUN=function(x) {strsplit(x,split='[,.]')
[[1]][2]})
combi$Title<-sub(' ','',combi$Title)
table(combi$Title)

##
##         Capt          Col          Don         Dona           Dr
##            1            4            1            1            8
##     Jonkheer         Lady        Major       Master         Miss
##            1            1            2           61          260
##         Mlle          Mme           Mr          Mrs           Ms
##            2            1          757          197            2
##          Rev          Sir the Countess
##            8            1            1

combi$Title[combi$Title %in% c('Mme','Mile')]<-'Mile'
combi$Title[combi$Title %in% c('Capt','Don','Major','Sir')]<-'Sir'
combi$Title[combi$Title %in% c('Dona','Lady','the Countess','Jonkheer')]
<-'Lady'
combi$Title<-factor(combi$Title)
combi$FamilySize<-combi$SibSp+combi$Parch+1
combi$Surname<-sapply(combi$Name,FUN=function(x) {strsplit(x,split='[,.]
')[[1]][1]})
combi$FamilyID<-paste(as.character(combi$FamilySize),combi$Surname,sep=
"")
combi$FamilyID[combi$FamilySize<=2]<-'Small'
table(combi$FamilyID)

##
##            11Sage           3Abbott         3Appleton          3Beckw
ith
##                11                 3                 1
  2
##           3Boulos           3Bourke            3Brown          3Caldw
ell
##                 3                 3                 4
  3
##          3Christy          3Collyer          3Compton           3Corn
ell
##                 2                 3                 3
  1
##           3Coutts           3Crosby           3Danbom            3Dav
ies
##                 3                 3                 3
  5
##            3Dodge          3Douglas            3Drew             3El
ias
##                 3                 1                 3
  3
```

```
##       3Frauenthal        3Frolicher 3Frolicher-Stehli        3Goldsm
ith
##                 1                 1                 2
  3
##       3Gustafsson       3Hamalainen           3Hansen            3H
art
##                 2                 2                 1
  3
##             3Hays          3Hickman         3Hiltunen         3Hirvo
nen
##                 2                 3                 1
  1
##         3Jefferys          3Johnson             3Kink    3Kink-Heilm
ann
##                 2                 3                 2
  2
##           3Klasen         3Lahtinen           3Mallet            3Mc
Coy
##                 3                 2                 3
  3
##          3Minahan         3Moubarek            3Nakid         3Navra
til
##                 1                 3                 3
  3
##          3Newell           3Newsom         3Nicholls          3Peac
ock
##                 1                 1                 1
  3
##           3Peter            3Quick         3Richards          3Rosb
lom
##                 3                 3                 2
  3
##          3Samaan        3Sandstrom           3Silven          3Sped
den
##                 3                 3                 1
  3
##           3Strom          3Taussig           3Thayer           3Tho
mas
##                 1                 3                 3
  1
##           3Touma     3van Billiard         3Van Impe    3Vander Pla
nke
##                 3                 3                 3
  2
##           3Wells            3Wick          3Widener          4Alli
son
##                 3                 3                 3
  4
##        4Backstrom          4Baclini           4Becker           4Car
ter
```

```
##                 1                 4                 4
  4
##         4Davidson            4Dean           4Herman           4Hock
ing
##                 1                 4                 4
  2
##        4Jacobsohn         4Johnston          4Laroche            4Ren
ouf
##                 1                 4                 4
  1
##    4Vander Planke            4West             5Ford           5Hock
ing
##                 1                 4                 5
  1
##    5Kink-Heilmann          5Lefebre          5Palsson           5Ryer
son
##                 1                 5                 5
  5
##          6Fortune           6Panula             6Rice          6Richa
rds
##                 6                 6                 6
  1
##            6Skoog        7Andersson          7Asplund           8Good
win
##                 6                 9                 7
  8
##             Small
##              1025
```

```
famIDS<-data.frame(table(combi$FamilyID))
famIDS<-famIDS[famIDS$Freq<=2,]
combi$FamilyID[combi$FamilyID%in% famIDS$Var1]<-'Small'
combi$FamilyID<-factor(combi$FamilyID)
train <- combi[1:891,]
test <- combi[892:1309,]
library(rpart)
fit<-rpart(Survived~Pclass+Sex+Age+SibSp+Parch+Fare+Embarked+Title+Fami
lySize+FamilyID,data=train,method='class')
Agefit<-rpart(Age~Pclass+Sex+SibSp+Parch+Fare+Embarked+Title+FamilySize,
data=combi[!is.na(combi$Age),],method='anova')
combi$Age[is.na(combi$Age)]<-predict(Agefit,combi[is.na(combi$Age),])
summary(combi)
```

```
##   PassengerId      Survived          Pclass          Name
##  Min.   :   1   Min.   :0.0000   Min.   :1.000   Length:1309
##  1st Qu.: 328   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median : 655   Median :0.0000   Median :3.000   Mode  :character
##  Mean   : 655   Mean   :0.3838   Mean   :2.295
##  3rd Qu.: 982   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :1309   Max.   :1.0000   Max.   :3.000
```

```
##                  NA's   :418
##      Sex                 Age            SibSp            Parch
##  Length:1309        Min.   : 0.17   Min.   :0.0000   Min.   :0.000
##  Class :character   1st Qu.:22.00   1st Qu.:0.0000   1st Qu.:0.000
##  Mode  :character   Median :28.86   Median :0.0000   Median :0.000
##                     Mean   :29.70   Mean   :0.4989   Mean   :0.385
##                     3rd Qu.:36.50   3rd Qu.:1.0000   3rd Qu.:0.000
##                     Max.   :80.00   Max.   :8.0000   Max.   :9.000
##
##     Ticket              Fare            Cabin
##  Length:1309        Min.   :  0.000   Length:1309
##  Class :character   1st Qu.:  7.896   Class :character
##  Mode  :character   Median : 14.454   Mode  :character
##                     Mean   : 33.295
##                     3rd Qu.: 31.275
##                     Max.   :512.329
##                     NA's   :1
##    Embarked            Title       FamilySize       Surname

##  Length:1309        Mr     :757   Min.   : 1.000   Length:1309

##  Class :character   Miss   :260   1st Qu.: 1.000   Class :character

##  Mode  :character   Mrs    :197   Median : 1.000   Mode  :character

##                     Master : 61   Mean   : 1.884

##                     Dr     :  8   3rd Qu.: 2.000

##                     Rev    :  8   Max.   :11.000

##                     (Other): 18

##        FamilyID
##  Small     :1074
##  11Sage    :  11
##  7Andersson:   9
##  8Goodwin  :   8
##  7Asplund  :   7
##  6Fortune  :   6
##  (Other)   : 194
```

```r
summary(combi$Embarked)
```

```
##    Length     Class      Mode
##      1309 character character
```

```r
combi$Embarked[c(62,830)]='S'
```

```
combi$Embarked<-factor(combi$Embarked)
combi$Fare[1044]<-median(combi$Fare,na.rm=TRUE)
combi$FamilyID2<-combi$FamilyID
combi$FamilyID2<-as.character(combi$FamilyID2)
combi$FamilyID2[combi$FamilySize<=3]<-'Small'
combi$FamilyID2<-as.factor(combi$FamilyID2)
library(randomForest)

## randomForest 4.6-12

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

set.seed(9)

combi$FamilyID2<-factor(combi$FamilyID2)

combi$FamilyID2<-combi$FamilyID
combi$FamilyID2<-as.character(combi$FamilyID2)
combi$FamilyID2[combi$FamilySize<=3]<-'Small'
combi$FamilyID2<-factor(combi$FamilyID2)
combi$Sex<-factor(combi$Sex)
```

把转好的数据重新分成训练集合测试集两部分

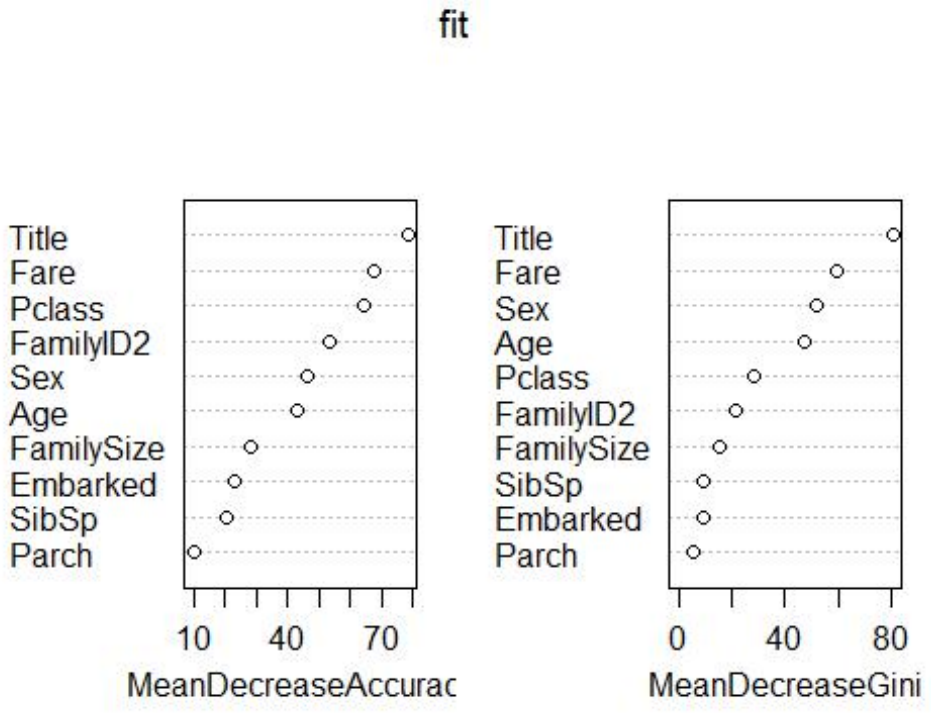
```
train<-combi[1:891,]
test<-combi[892:1309,]
```

用随机森林对训练集训练，建立模型

```
fit<-randomForest(as.factor(Survived) ~ Pclass + Sex + Age + SibSp + Pa
rch + Fare + Embarked + Title +FamilySize+FamilyID2,data=train,importan
ce=TRUE,ntree=2000)
```

查看各特征值重要性排序

```
varImpPlot(fit)
```



对测试机数据进行预测，并将结果保存

```
Prediction<-predict(fit,test)
submit<-data.frame(PassengerId=test$PassengerId,Survived=Prediction)
write.csv(submit,file = "firstforest.csv",row.names = FALSE)
```

把预测结果上传到 Kaggle 后，可以查看得分，我的得分是 **0.76555**