~~~~
---
title: "random forest"
author: "Yunhe Liu"
date: "12/5/2021"
output: pdf_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r}
data = read.csv("processed_counts.csv")
```

```{r}
label = read.csv("annotation.csv")
```



```{r}
library(sampling)
set.seed(6690)

train_id <- sample(label$ID, round(dim(label)[1]*0.75))

train_data <- data[data$ID %in% train_id, ]
test_data <- data[!(data$ID %in% train_id), ]

train_label <- label[data$ID %in% train_id, ]
test_label <- label[!(data$ID %in% train_id), ]
```

```{r}
total_train = merge(train_data, train_label, by = "ID")
total_test = merge(test_data, test_label, by = "ID")
total_train = total_train[, -1]
total_test = total_test[, -1]
```

```{r}
library(ggplot2)
library(lattice)
library(caret)
~~~~

~~~
control <- trainControl(method = 'repeatedcv', number = 2, repeats = 2)
model <- train(Type~., total_train,
               method = 'rf',
               preProcess = c('center', 'scale'),
               trControl = control)
model
```
~~~

```
Random Forest


5730 samples

2916 predictors

  16 classes: 'BLCA', 'BRCA', 'CESC', 'COAD', 'GBM', 'HNSC', 'LIHC', 'LUAD',
'LUSC', 'Normal', 'PRAD', 'READ', 'SKCM', 'STAD', 'THCA', 'UCEC'


Pre-processing: centered (2916), scaled (2916)

Resampling: Cross-Validated (2 fold, repeated 2 times)

Summary of sample sizes: 2867, 2863, 2866, 2864

Resampling results across tuning parameters:


  mtry  Accuracy   Kappa

     2  0.9137003  0.9065977

    76  0.9368221  0.9316793

  2915  0.9302783  0.9246248


Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 76.



Confusion Matrix and Statistics
```

```
       truth

pred     BLCA BRCA CESC COAD GBM HNSC LIHC LUAD LUSC Normal PRAD READ SKCM STAD
THCA UCEC

  BLCA     92    0    0    0   0    0    0    0    0      2    0    1    0    0
0    0

  BRCA      0  268    0    0   0    0    0    2    0      3    0    0    1    0
0    0

  CESC      2    0   70    0   0    5    0    0    2      0    0    0    0    0
0    0

  COAD      0    0    0  104   0    0    0    0    0      0    0   43    0    1
0    0

  GBM       0    0    0    0  43    0    0    0    0      1    0    0    0    0    0
0

  HNSC      3    2    2    0   0  132    0    0    5      0    0    0    0    1
0    0

  LIHC      0    0    0    0   0    0   82    0    0      0    0    0    0    0
0    1

  LUAD      0    0    0    0   0    0    0  133   10      1    0    0    0    0
0    0

  LUSC      0    0    1    0   0    1    0    3  124      0    0    0    0    0
0    0

  Normal    0    3    0    0   0    0    2    3    0    147    1    0    0    0
1    0

  PRAD      0    0    0    0   0    0    0    0    0      2  126    0    0    0
0    0

  READ      0    0    0    0   0    0    0    0    0      0    0    0    0    0    0
0

  SKCM      0    0    0    0   0    0    1    0    0      0    0    0  109    0
0    0

  STAD      0    0    0    3   0    0    0    0    0      0    0    0    0   94
0    0

  THCA      0    0    0    0   0    0    0    0    0      1    0    0    0    0
124    0
```

```
  UCEC     1   0   4   0  0   0   0   0   0    1   0   0   0   0   0
146


Overall Statistics


               Accuracy : 0.9393

                 95% CI : (0.9276, 0.9496)

    No Information Rate : 0.1429

    P-Value [Acc > NIR] : < 2.2e-16


                  Kappa : 0.9343


 Mcnemar's Test P-Value : NA


Statistics by Class:


                     Class: BLCA Class: BRCA Class: CESC Class: COAD Class: GBM
Class: HNSC

Sensitivity              0.93878     0.9817    0.90909    0.97196
1.00000    0.95652

Specificity              0.99834     0.9963    0.99509    0.97560
0.99946    0.99266

Pos Pred Value           0.96842     0.9781    0.88608    0.70270   0.97727
0.91034

Neg Pred Value           0.99669     0.9969    0.99618    0.99830   1.00000
0.99660

Prevalence               0.05131     0.1429    0.04031    0.05602   0.02251
0.07225

Detection Rate           0.04817     0.1403    0.03665    0.05445
0.02251    0.06911
```

```
Detection Prevalence    0.04974     0.1435    0.04136    0.07749
0.02304    0.07592

Balanced Accuracy      0.96856     0.9890    0.95209    0.97378
0.99973    0.97459

                   Class: LIHC Class: LUAD Class: LUSC Class: Normal Class:
PRAD Class: READ

Sensitivity           0.96471    0.94326    0.87943      0.93038
0.99213    0.00000

Specificity           0.99945    0.99378    0.99717      0.99429
0.99888    1.00000

Pos Pred Value         0.98795    0.92361    0.96124      0.93631
0.98437        NaN

Neg Pred Value         0.99836    0.99547    0.99045      0.99373
0.99944    0.97696

Prevalence            0.04450    0.07382    0.07382      0.08272
0.06649    0.02304

Detection Rate         0.04293    0.06963    0.06492      0.07696
0.06597    0.00000

Detection Prevalence    0.04346    0.07539    0.06754      0.08220
0.06702    0.00000

Balanced Accuracy      0.98208    0.96852    0.93830      0.96234
0.99550    0.50000

                   Class: SKCM Class: STAD Class: THCA Class: UCEC

Sensitivity           0.99091    0.97917    0.99200    0.99320

Specificity           0.99944    0.99835    0.99944    0.99660

Pos Pred Value         0.99091    0.96907    0.99200    0.96053

Neg Pred Value         0.99944    0.99890    0.99944    0.99943

Prevalence            0.05759    0.05026    0.06545    0.07696

Detection Rate         0.05707    0.04921    0.06492    0.07644

Detection Prevalence    0.05759    0.05079    0.06545    0.07958

Balanced Accuracy      0.99518    0.98876    0.99572    0.99490
```