random forest

Yunhe Liu

12/5/2021

```
data = read.csv("processed_counts.csv")
```

```
label = read.csv("annotation.csv")
```

```
label$Type[which(label$Type == "Normal")] <- 0
label$Type[which(label$Type != 0)] <- 1
```

```
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), ]
```

```
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]
```

```
library(ggplot2)
library(lattice)
library(caret)
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:sampling':
##
##     cluster
```

```
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
```

```
##    2 classes: '0', '1'
##
## Pre-processing: centered (2916), scaled (2916)
## Resampling: Cross-Validated (2 fold, repeated 2 times)
## Summary of sample sizes: 2866, 2864, 2865, 2865
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##      2  0.9714657  0.7636145
##     76  0.9836826  0.8771432
##   2915  0.9842932  0.8815635
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2915.
```

```
truth <- total_test$Type
pred <- predict(model, newdata = total_test)
confusionMatrix(table(pred, truth))
```

```
## Confusion Matrix and Statistics
##
##     truth
## pred    0    1
##    0  142    6
##    1   16 1746
##
##                Accuracy : 0.9885
##                  95% CI : (0.9826, 0.9928)
##     No Information Rate : 0.9173
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.9219
##
##  Mcnemar's Test P-Value : 0.05501
##
##             Sensitivity : 0.89873
##             Specificity : 0.99658
##          Pos Pred Value : 0.95946
##          Neg Pred Value : 0.99092
##              Prevalence : 0.08272
##          Detection Rate : 0.07435
##    Detection Prevalence : 0.07749
##       Balanced Accuracy : 0.94765
##
##        'Positive' Class : 0
##
```