Gene Expression Analysis

Youqing Xiang

12/5/2016

```
library(dplyr)
library(caTools)
library(randomForest)
```

Introduction

This project is focusing on taking advantage of Principal Components Analysis (PCA) tehnique to reduce the dimensions. Here is the link for the data source. This data has 20531 variables but only has 801 observations. The data set is part of the RNA-Seq (HiSeq) PANCAN data set, it is a random extraction of gene expressions of patients having five different types of tumor: BRCA, KIRC, COAD, LUAD and PRAD. After reducing the number of dimensions, I also tried to build a model to predict the type of tumor.

Loading the data

```
genes <- read.csv('Data/data.csv')
labels <- read.csv('Data/labels.csv')
dim(genes)
```

```
## [1]   801 20532
```

```
head(genes[,1:5])
```

```
##          X gene_0    gene_1   gene_2   gene_3
## 1 sample_0      0 2.0172093 3.265527 5.478487
## 2 sample_1      0 0.5927321 1.588421 7.586157
## 3 sample_2      0 3.5117590 4.327199 6.881787
## 4 sample_3      0 3.6636179 4.507649 6.659068
## 5 sample_4      0 2.6557411 2.821547 6.539454
## 6 sample_5      0 3.4678533 3.581918 6.620243
```

```
unique(labels$Class)
```

```
## [1] PRAD LUAD BRCA KIRC COAD
## Levels: BRCA COAD KIRC LUAD PRAD
```

Cleaning the data

```
genes <- genes[,colSums(genes !=0)>0]
dim(genes)
```

```
## [1]   801 20265
```

There are some of columns with all zero values, which won't contribute any to the classifiction of tumor types. So, here I got ride of those columns.

Reducing the dimensions with PCA

```
A <- as.matrix(select(genes,-X))
df <- select(labels,-X)

i = 1
while (i < 26) {
  pca_i <- princomp(A[,((i-1)*800+1):(i*800)],center=T,scale.=T)
  pca_i <- as.data.frame(pca_i$scores[,1:10])
  df <- cbind(df,pca_i)
  i = i + 1
}

pca_26 <- princomp(A[,20000:20264],center=T,scale.=T)
pca_26 <- as.data.frame(pca_26$scores[,1:10])
df <- cbind(df,pca_26)
colnames(df) <- c('Class',1:260)
colnames(df) <- paste("N", colnames(df), sep = "_")
dim(df)
```

```
## [1] 801 261
```

```
head(df[,1:5])
```

```
##   N_Class       N_1        N_2        N_3         N_4
## 1    PRAD 12.135331  23.769535 -19.884646   0.7613011
## 2    LUAD  3.173851 -19.124716  -8.633262  13.1611382
## 3    PRAD 17.427720   7.800015 -16.297847 -20.5470332
## 4    PRAD 17.017072  21.711485 -15.634307  -9.8953720
## 5    BRCA 18.381579  -2.315122  15.132634  -6.9436837
## 6    PRAD  6.073407  25.829731  -6.641242  -1.2162171
```

For this step, I performed PCA technique for every 800 variables and only kept top 10 components. For the last 264 variable, I also kept top 10 components. Together I did 26 run PCA and got 260 components in total.

Splitting the data into Train/Test sets

```
sample <- sample.split(df$N_Class, SplitRatio = 0.7)
train <- subset(df, sample==T)
test <- subset(df, sample==F)
dim(train)
```

```
## [1] 561 261
```

```
dim(test)
```

```
## [1] 240 261
```

Training the model

```
rf.model <- randomForest(N_Class ~ ., data = train, importance=TRUE)
rf.model$confusion
```

```
##      BRCA COAD KIRC LUAD PRAD class.error
## BRCA  210    0    0    0    0 0.000000000
## COAD    0   55    0    0    0 0.000000000
## KIRC    1    0  101    0    0 0.009803922
## LUAD    2    0    0   97    0 0.020202020
## PRAD    0    0    0    0   95 0.000000000
```

Testing the model

```
rf.preds <- predict(rf.model, test,type='response')
table(rf.preds,test$N_Class)
```

```
##
## rf.preds BRCA COAD KIRC LUAD PRAD
##     BRCA   90    0    0    0    0
##     COAD    0   23    0    0    0
##     KIRC    0    0   44    0    0
##     LUAD    0    0    0   42    0
##     PRAD    0    0    0    0   41
```

Discussion

1) PCA is an effective way to dramatically reduce the dimension of data.
2) randomForest model gives high accuracy prediction.
3) I have 260 PCA components at the end. However, this number is still high, considering there are only 801 rows of data. So, overfitting is a potential problem.
4) For this data set, 20531 genes' expression were measured. In reality, due to the cost of test, we may be able to measure a certain number of gene expression for each patient. So, how to find the genes which are relate to the type of tumor should be the focus in the future study.