Discipline Project 2

Ricky Cai (440242475)

Topic

As the technology for sequencing genes becomes more readily available, more powerful and cheaper, the information about genes that is captured becomes more detailed. This becomes a problem when using machine learning techniques to analyse the high dimensional data. This problem is known as the "big p, little n" problem where p refers to the number of measurements and n refers to the subjects. Consequences of high dimension data include multiple hypothesis testing and high computation time. In this report, we analyse the GSE107509 gene expression data set that contains 659 samples and over 54,000 genes. This data set was chosen for its high number of samples leading to a higher diversity of samples.

Dimension reduction techniques such as Principal Component Analysis (PCA) is a way to transform the high dimension data into low dimension data while attempting to preserve the original information as much as possible. PCA is used to help alleviate the problems associated with "big p, little n".

The topic that is addressed in this report analyses how dimension reduction techniques such as PCA affect the accuracy of random forest models. An accompanying Shiny App was developed to provide a way to interactively visualise the effect of how PCA (Principal Component Analysis) affects the accuracy of a random forest model on gene expression.

Approach

The approach to answering the question uses Principal Component Analysis to reduce the number of dimensions in the data, Random Forests to predict the outcome and cross validation to predict the accuracy of the models. The approach includes preselecting a number of principal components since it was not feasible to run in real time. Random Forest models were trained using the reduced number of principal components and cross validation was used with boxplots to show the effect of how the number of principal components affects the accuracy of the Random Forest model.

Shortcomings

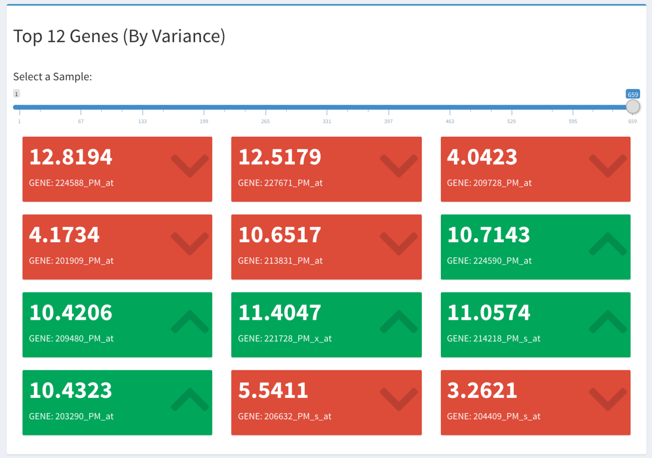
A shortcoming to the data analytics was the size of the data which had an effect on how the data was stored on Github (There is a 100MB limit on Github). To host the Shiny application on Github, the gene expression data was split into several parts and recombined in the Shiny app when it is run.

Another shortcoming was the computation time for running PCA and Random Forest. To address this issue, the data was preprocessed and saved. Some of the data used in the Shiny app uses preprocessed data to reduce the time it takes to run models. This prevents the user in having full control over the parameters but a selection of preprocessed values is provided for comparison.

Graphics

There are 3 sections to the Shiny App. The sections include Calculator, PCA and Random Forest.

The first section makes use of colour to visualise the top 12 genes (measured by the highest variance) and visualises how an individuals gene expression compares with the mean in the data set (Fig. 1). Red indicates the individual had a gene expression lower than the mean while green indicates the individual had a gene expression higher than the mean for that particular gene.



Top 12 Genes (By Variance)

Select a Sample:

12.8194
GENE: 224588_PM_at

12.5179
GENE: 227671_PM_at

4.0423
GENE: 209728_PM_at

4.1734
GENE: 201909_PM_at

10.6517
GENE: 213831_PM_at

10.7143
GENE: 224590_PM_at

10.4206
GENE: 209480_PM_at

11.4047
GENE: 221728_PM_x_at

11.0574
GENE: 214218_PM_s_at

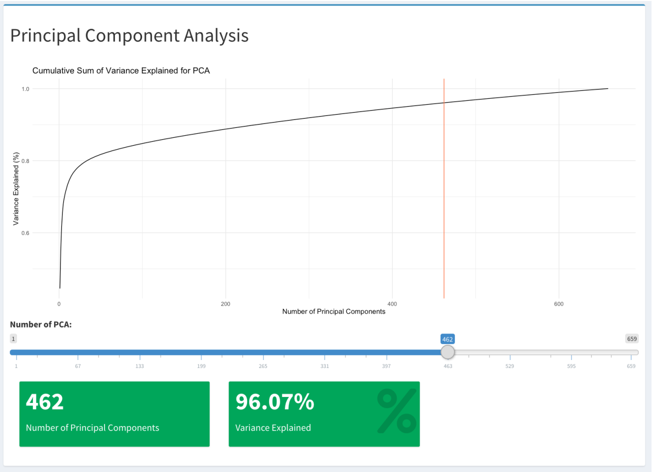
10.4323
GENE: 203290_PM_at

5.5411
GENE: 206632_PM_s_at

3.2621
GENE: 204409_PM_s_at

Figure 1: Comparing individual gene expression with mean

The second section makes use of a screeplot of the PCA to show the cululative variance explained by the number of Principal Components selected (Fig. 2). The user can specify the number of Principal Components by a slider which dynamically shows the percentage of variance explained by the specified number of Principal Components.


Principal Component Analysis

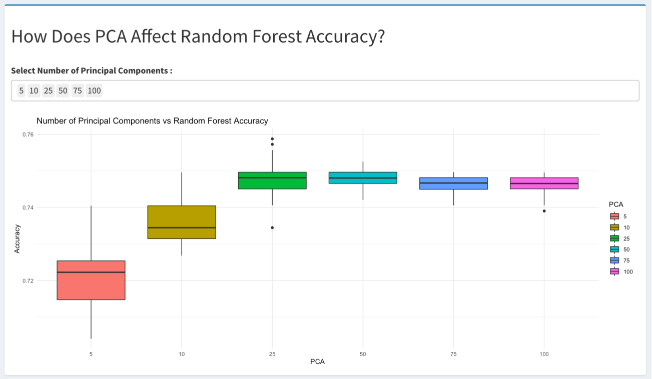
Number of PCA: 462

462 Number of Principal Components

96.07% Variance Explained

Figure 2: Cumulative screeplot of PCA

The third section makes use of boxplots to show how the number of Principal Components affects the accuracy of a Random Forest model (Fig. 3). The user can compare between 5, 10 25, 50, 75, 100, 150 and 200 components using a selection box which dynamically updates the boxplots.


How Does PCA Affect Random Forest Accuracy?

Select Number of Principal Components : 5 10 25 50 75 100

Figure 3: Boxplots showing effect of principal components on accuracy

Shiny App

The Github link to the Shiny App is provided below:

Github Link

The Shiny App can be accessed by the following lines of code in R (May take a few minutes to download):

```r
library(shiny)
runGitHub("DATA3888_Discipline_Project_2", "iRick92")
```