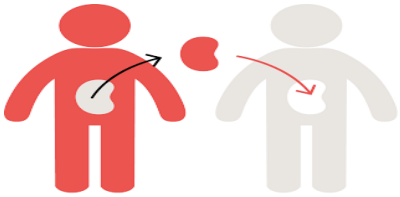


Kidney Transplant Outcome Prediction

Executive Summary

470539309

Objectives

Acute rejection is still common after kidney transplant. Since there are much more kidney recipients than donors, it is preferable to allocate kidneys to the recipients who are unlikely to have acute rejection after kidney transplant. Thus, it is important to make a kidney transplant outcome prediction before the surgery. This project aims to provide doctors a shiny app, which could predict either the patient will have acute rejection or be stable after kidney transplant. The shiny app has a predictive model embedded, which uses random forest to classify the kidney transplant outcome based on RNA-seq.

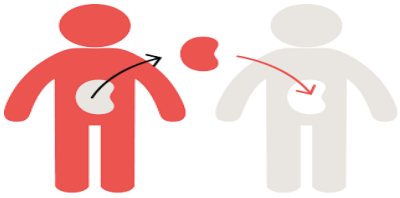
Approach

Gene expressions data GSE131179 from Gene Expression Omnibus (GEO), National Center for Biotechnology Information (NCBI) is used to build the predictive model. This dataset includes 34 patients who had kidney transplant, with an outcome (Acute Cellular Rejection (ACR) or Normal/Non Specific) and 60466 different RNA-seqs for each patient. RNA-seqs with top 10% largest variance among patients are used as explanatory variables to predict transplant outcome. 5-fold cross validation is repeated for 25 times for three classifiers, which are k-nearest neighbours, support vector machine and random forest, and average accuracy for each cross validation is recorded. After the 25 repeats, it shows that random forest has the best accuracy distribution, with a mean accuracy around 0.87. Therefore, random forest is chosen as the classifier of the predictive model.

Random forest is an ensemble learning method which constructs many decision trees, and takes the majority of the results from decision trees as the final result. Since there are only 34 patients in the dataset, random forest is a suitable classifier which could reduce the model instability by averaging multiple decision trees.

Discussion

The major potential shortcoming is feature selection issue before random forest modelling. If an RNA-seq is invariant among different patients with different kidney transplant outcomes, it is more likely that this RNA-seq has low or zero correlation with the outcomes. For this reason, and to reduce computational time for modelling, only the top 10% of RNA-seqs with relatively large variance are selected as explanatory variables. However, larger variance does not mean larger correlation. It is possible that the left 90% less variant RNA-



seqs have high correlation with the outcomes. By removing 90% less variant RNA-seqs, there could be many important features missing.

To solve this problem, and to limit the computation time of modelling as short as possible, t-test could be conducted for each RNA-seq, and only significant features with low p-value are selected to be explanatory variables for modelling. By doing this, only RNA-seqs which are correlated with the transplant outcomes are selected, which can improve the random forest modelling.

Shiny App

A shiny app is created with random forest classifier embedded. This shiny app can produce prediction of the outcome of kidney transplant. Doctors could make the decision of whether the patient should conduct a kidney transplant surgery based on the prediction result.

The shiny app contains a sidebar where users can choose a patient to predict. It has a database with 34 patients which allows users to choose one patient to have a look. It also allows users to upload their own patient's details. The patient's details should be in a csv file which only include one patient.

The shiny app has three tabs in its main panel, which are Prediction Result, Data and Predictive Model: Random Forest. In the Prediction Result tab, the prediction result from the input patient's data is printed. This will tell doctors either the patient is predicted to have acute rejection or not. In the Data tab, the number of RNA-seqs of the input data is printed, as well as the data itself. Doctors who understand RNA-seq well can scroll down the whole RNA-seqs to check the values. In the Predictive Model tab, there is a general introduction of the random forest classifier which is used for the prediction.

The random forest model is trained with the patient database which has 34 patients' 6047 RNA-seqs. It is essential that the input data has all these 6047 RNA-seqs to get a prediction. Excess RNA-seqs will be ignored automatically. The Prediction Result tab will print an insufficient data message when the input data does not contain all of the 6047 RNA-seqs.

Shiny app GitHub link: https://github.com/sherry-fan01/DATA3888_DiscplineProject2
Users should read README.md first.