BiS335 FinalTerm Project

Contents

Project introduction

- In mid-term project, you found significant features, such as mutation of certain genes and thresholds of gene expression, that can separate patients with the survival time. Now we consider survival time as "censored" variable that contains "time to follow up" and "survival status". Here "survival status" is 1, when patient is died at the time of clinician followed and 0 otherwise. In this setting you need to use special statistics such as Kaplan-Meier survival test (KM-Survival test) to estimate patient survival probability through non-parametric estimation of survival function. (see supplementary tutorial material in Nonparametric testing lecture)
- Suppose that clinicians predict survival of patient based on the predefined classes such as pathological stages, and subtypes. As a genomic consultant, you would like to suggest better classifier of patient survival based on the patient's genomic profile. Answer questions below.

Project description

1. Find the best class label for prediction of survival time among 3 class labels such as stages, subtypes and survival index given in the clinical data set.
   a. Perform Kaplan-Meier test with given class labels.
   b. What is the best class label which maximize p-value of KM-test? ?
   c. Perform KM-test using samples exist in all three data sets (clinical, mutation, and expression data sets). Is there any difference from the result in "b"?
   d. Compare median survival time of this result with the result of question #1 in your mid-term project. (If KM-plot does not cross 50% of survival probability, use 75% instead)

2. Using the best class label you found in question #1, find the best classification model based on genomic feature combination which can maximize classification performance.
   a. Build LDA or QDA, SVM, and one of the tree-based (Decision tree, Bagging, Random forest, Boosting) model utilizing patient genomic features suggested by pre-selected feature table(gex_anova_result.csv, mut_chisq_result.csv).
   b. Perform 5-fold cross validation.

3. Now you would like to suggest patient class that can explain patient survival time equivalently well compared to the best class label you found question #1.
   a. Perform K-means, Hierarchical clustering using gene expression data set. (option: You can apply dimension reduction techniques)

   b. What is the optimal number of cluster? Evaluate cluster quality with “NbClust” package. (Consider Scree plot and Silluette width first, and optionally explain other measures you would like to use. You can specify measure with “index = silhouette” argument)
   c. Calculate cluster similarity between the best class label and the cluster you found.
   d. Perform clustering using the best feature set you found in question #2, check whether cluster similarity is improved or not.

4. Conclude your project
   a. Do the features you found in mid-term project improves performance of your classification model?
      i. How many are they in your final model?
      ii. What is the performance improvement/decrement if you add or neglect the random features versus features that you found in mid-term project to your model? Describe the performance in terms of variance and bias.
   b. Using features that you found in mid-term project, build a classification model. Compare performance of this model and the best model you obtained in question #2.

Description of the data set

?

We have 3 data sets for this term project. Mutation and gene expression data are the same as the MidTerm project. Clinical annotation added survival index. The survival index is a class label related to the survival of the patient and could be regarded as a class label of the same type as the stage or subtype.

- gex_anova_result.csv: p-value was extracted by performing ANOVA test based on the expression level of individual genes and then the FDR value was extracted by adjusting the p-value using the Benjamini–Hochberg(BH) method. It was filtered with FDR 0.05

- mut_chisq_result.csv: p-value was extracted by performing Chisq test based on the mutation frequency of individual genes and then the FDR value was extracted by adjusting the p-value using the Benjamini–Hochberg(BH) method. It was filtered with FDR 0.1

Report format

- Contents [1]
  - Team member’s name
  - Method Objective
  - Mathematical description of the method principle & analysis procedure
  - References (Journal, Web Site, . . .)
- Contents [2]
  - R codes & Outputs with explanation of functions and parameters of code
  - Peer’s contributions (see Peer-review section)
- Upload your work (Content [1], [2]) on KLMS with one zip file.
  - File name: Team1_TermProject.zip, Team1_Content[1].docx, Team1_Content[2].docx

Policies

- This project is a team assignment. Cheating/copying/using other’s work get 0 point.
- Programming language is limited to R.
- If your code does not run, you will get 0 point.
- Late submission will get 20% penalty/hr.
- Due: 6 pm of Dec / 12 (12-12-2018)

Peer-review

1) Include the contributions of colleagues at the end of the report. Follow the guidelines given below.
   - Formal Analysis (30 %): Application of statistical, mathematical, computational, techniques to analyze study data
   - Visualization (20 %): visual presentation of study result
   - Writing - Original Draft Preparation (30 %): writing initial draft
   - Writing - Review & Editing (20 %): finalizing draft
2) Submit your colleagues' ranks individually via email.
3) Individual evaluation may be different depending on participation by question because the difficulty of each question is different.(The difficulty rank of each question: Q4 - Q2 - Q3 - Q1)

Inquiry

- Myeongha Hwang: hmh929@kaist.ac.kr
- Seongyong Park: sypark0215@kaist.ac.kr