RE6071020 徐翊展

Gene of affecting Leukemia

Motivation

Everything of life is compose of gene. All phenotype is determined by your own gene expression. For example, leuemia patient will have different expression in some gene compare to ordinary people. Our motivation is that if we could find which gene has different expression in case and control patient then it can infer which gene may affected symptoms of leukemia.

Implement

- Data preprocess

  The table come form bioinformation has n << p usually, where n is subject and p is gene. Formal start up for analyse microarray table is transpose it. There are two advantages about doing transpose, time of read whole table will reduce and keep from overfitting cause from sample size is smaller than number of feature.

- Analyse method

  We use two sample t-test to check all gene whether two group $\bar{x}$ is different between case and control because our goal is to find the winner list of gene that has different gene expression. Finally, we use random forest fit features of winner list then to predict who may get leukemia.

Implement 2

- Precision medicine

  The term of precision medicine is fancy, the main concept of it is to treat patients different remedy by their gene. Reason of do the thing is that efficiency of mediciene is different for everyone because everything of life is compose of gene. Overall, precision medicine is not only make diagnostic become more efficient but also save a lot of cost on unnecessary remedy.

- Attention mechanism

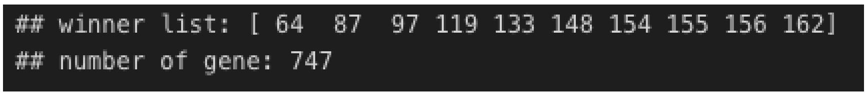
  THe term of attention is come from domain of computer vision, it's clear that we only focus on some point when our see something. For example, we can distinguish a car or not only see the number of wheel. There are some advantages of doing attention, such as make the accuracy more exact compare to use all feature for fitting.

Conclusion

- Winner list

  Setting $\alpha$ as 0.01, we reject $H_0$ if p-value < $\alpha$

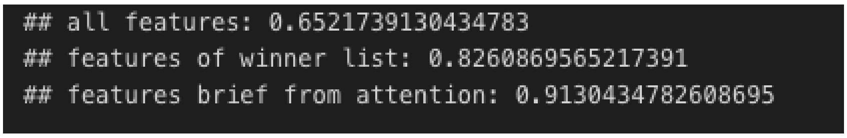
  ```
  ALPHA = 0.01
  winnerIndex = get_winnerList(x, y, 0.01)
  ```

  ```
  ## winner list: [ 64  87  97 119 133 148 154 155 156 162]
  ## number of gene: 747
  ```

- Accuracy Using sklearn::RandomForestClassifier to predict the label

  ```
  RF_all = RandomForestClassifier(random_state=1)
  RF_all.fit(x_train, y_train)
  RF_winner = RandomForestClassifier(random_state=1)
  RF_winner.fit(x_winner_train, y_winner_train)
  RF_attention = RandomForestClassifier(random_state=1)
  RF_attention.fit(x_train_att, y_train)
  ```

  ```
  ## all features: 0.6521739130434783
  ## features of winner list: 0.8260869565217391
  ## features brief from attention: 0.9130434782608695
  ```