Term Project Report
High-Grade Prostate Cancer Prediction using DNA Methylation Data

Fawwaz Dzaky Zakiyal (201899213)
Tanjung Dion (201883621)

Bioinformatics (Fall 2018)

1 Introduction

Prostate cancer affects a significant proportion of the male population, but in most cases, the disease is harmless. However, for the small portion of the population with high-grade prostate cancer, the disease can be extremely debilitating, causing painful symptoms and even death [1]. Over the past years, rapid advances in sequencing technology have led to The Cancer Genome Atlas (TCGA) project which provides the most comprehensive genomic data for various kinds of cancer. In the previous research, to indicates the classification of patient samples was done by using TCGA Exon expression or SNP datasets. However, the recent study shows that DNA methylation act as better bio-markers and help in improving the cancer prognosis [2]. Beside that, in the recent years, machine learning become a useful tool for prediction and regression tasks. It could bring benefit to the cancer risk prediction. Thus, this project implements machine learning techniques to classify high-grade prostate cancer risk using DNA methylation data.

2 Methodology

2.1 Concepts

- **DNA methylation**: the process of adding methyl groups to DNA, in this process modification of covalent nucleotides in the human genome, namely cytosine and also guanine. It is one of epigenetic modification which takes an important role in the development of cancer.
- **Gleason Score**: a measure (from 2-10) of the aggression of prostate cancer cells based on clinical pathology of prostate tissue [1].
- **High-Grade Prostate Cancer (HGPCa)**: A Gleason score from 8-10 is indicative of High-Grade prostate cancer, while scores below 8 are considered not as severe, HGPCa generally results in poor patient outcomes mortality, complications, and long-term disease-free survival) [1].
- **Cancer Gene Catalogues**: a catalogue those genes which contain mutations that have been causally implicated in cancer.

2.2 Dataset

The dataset comes from the National Cancer Institute GDC Data Portal and can be found online under Project TCGA-PRAD (The Cancer Genome Atlas Prostate Adenocarcinoma) [3]. This dataset consists of genomic information and clinical pathology reports belonging to 549 patients who have been diagnosed with prostate cancer (a Gleason Score anywhere between 2 to 10). It contains both clinical (recurrence, survival & treatment resistance) and molecular profiles (Exon (mRNA) expression, DNA methylation, Copy Number Variations (CNV) & Single Nucleotide Polymorphism (SNP)) for both tumor samples and normal controls. The file that we used are:

- **PRAD.meth.by_mean.data**: the mean signal values among each gene corresponding to the patient barcode. The size is 20111 genes x 549 patients (52 MB). This data become our learning features.
- **All_CDEs**: clinical data elements corresponding to the patient barcode. It contains many related information, but we only consider the gleason score.

2.3 System Design

The process consists of preprocessing of TCGA-PRAD dataset, consist of feature selection, missing data imputation & adding the label. After that, split the dataset into training & testing data, next we train & evaluate the classifier models. The system design can be seen in Figure 1. The system specification where this project running on:

- **Programming language**: Python
- **Support library**: Scikit-Learn & Pandas
- **PC**: Windows 10 64-bit, Intel i5-7500, RAM 4 GB

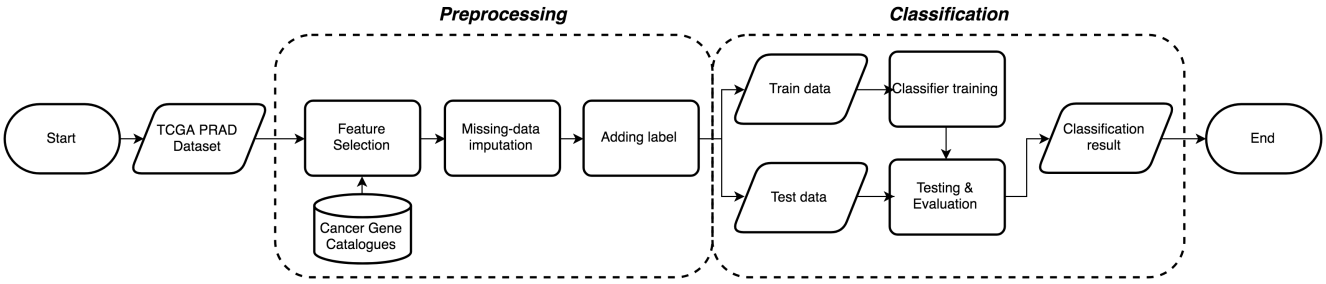

Figure 1: System Design

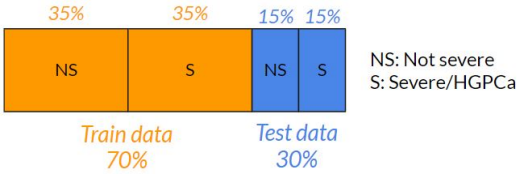

Figure 2: Dataset split

2.4 Preprocessing

- **Feature selection**: To reduce the feature space of the dataset, only those genes for which mutations have been causally implicated in cancer are considered, these are obtained through resources like *COSMIC* (Catalogue of Somatic Mutations in Cancer) that consist of 3946 genes information [4], and *CIVIC* (Clinical Interpretation of Variants in Cancer) that consist of 3946 genes information [5]. Originally, our dataset has 20111 genes, after feature selection it has only 821 genes.
- **Missing data imputation**: It is the process of replacing missing data of dataset features with substituted values. This project using mean imputation.
- **Adding the label**: We categorized into 2 classes, between score of 2 to 7 (not severe) and 8 to 10 (HGPCa / severe) from All_CDEs data. We added class labels to the dataset with correspond to the patient id.

2.5 Classification

- **Dataset split**: We split the dataset, 70% for train data & 30% for test data of each data group (not severe and severe/HGPCa that illustrated in Figure 2.
- **Training and evaluation**: This project perform on several machine learning algorithms, namely *Logistic Regression*, *K-Nearest Neighbor*, *Support Vector Machine*, *Multilayer Perceptron*. In the experiments, we evaluate parameters for each machine learning method to find the one that gives highest accuracy.

3 Experiments

- **Evaluation on Logistic Regression (LR)**: The experiment cases are LR using L1 or L2 penalty. The result is LR+L1 accuracy (76.36%) outperform LR+L2. The result shown at Figure 3.
- **Evaluation on K-Nearest Neighbor (KNN)** : The experiment cases are KNN with number of neighbors 2, 3, 4, or 5. The result is KNN with 3 neighbors accuracy (68.48%) outperform the other cases. The result shown at Figure 4.
- **Evaluation on Support Vector Machine (SVM)**: The experiment cases are SVM using linear kernel, RBF kernel, or Poly kernel (degree 2, 3, or 4). The result is SVM using linear kernel accuracy (76.97%) outperform the other cases. The result shown at Figure 5.
- **Evaluation on Multilayer Perceptron (MLP)**: The experiment cases are MLP using Sigmoid/Logistic, Tanh or ReLU activation function with number of hidden layers 1 (100 neurons) or 2 (100 neurons & 100 neurons). The result is MLP using Tanh activation function with 2 hidden layer accuracy (76.97%) outperform the other cases. The result shown at Figure 6.

4 Conclusion

This project succeeds in implementing machine learning techniques to classify high-grade prostate cancer risk using DNA methylation data. From the experiments, the result shows that SVM and MLP give the highest accuracy at the same value, 76.97%.

Future work to increase the performance accuracy are further explore in tuning the other parameters of each machine learning algorithms and try in combining the other dataset features, such as SNP, Exon, the other clinical data information (hereditary factors, alcohol consumption, sexual activity, race, etc.).

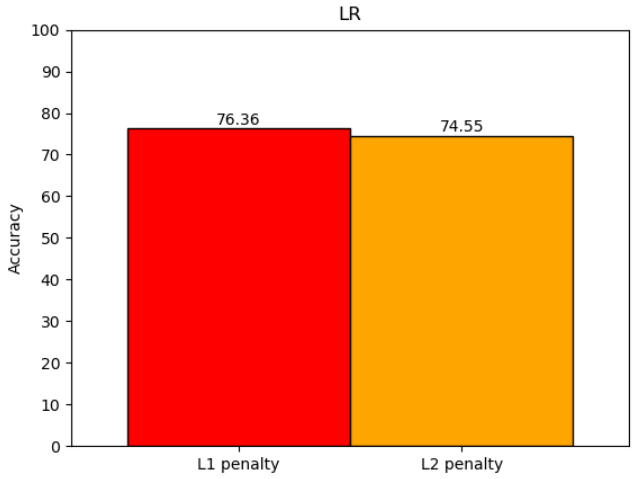

Figure 3: Evaluation on Logistic Regression

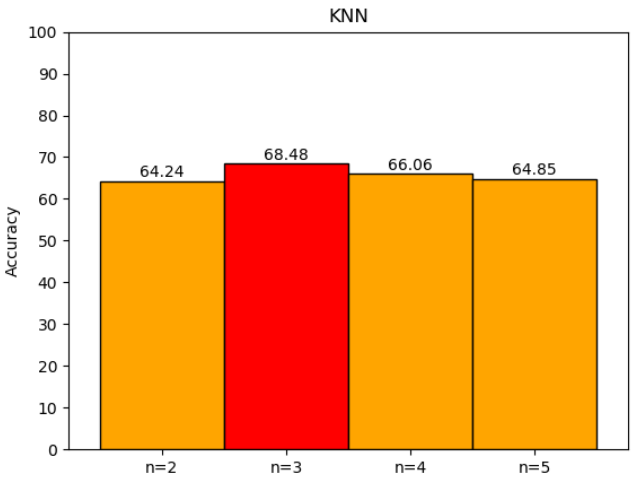

Figure 4: Evaluation on K-Nearest Neighbor

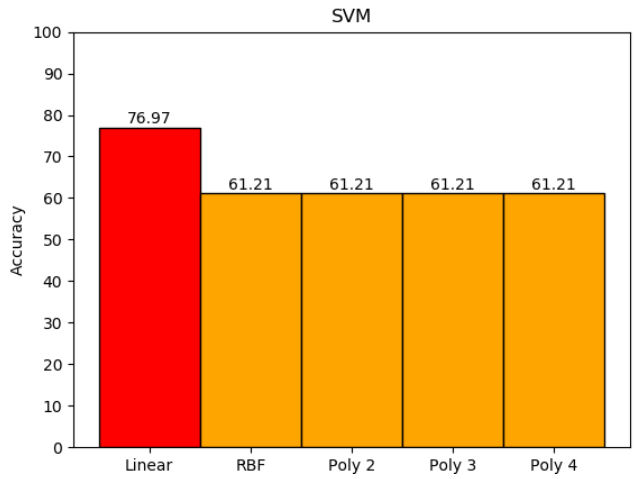

Figure 5: Evaluation on Support Vector Machine

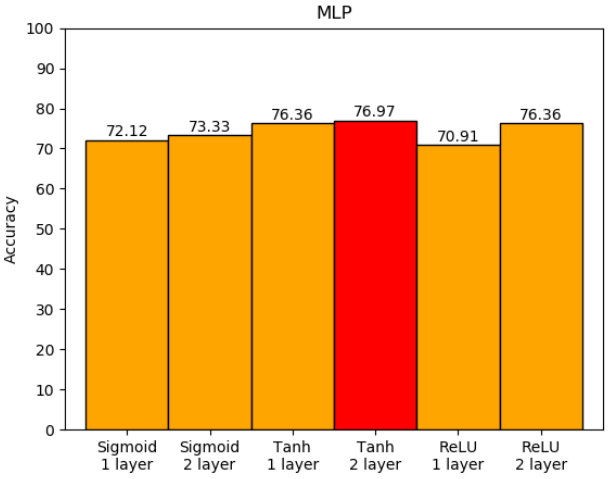

Figure 6: Evaluation on Multilayer Perceptron

References

[1] Ilir Agalliu, Robert Gern, Suzanne Leanza, and Robert D. Burk. 2009. *Associations of High-Grade Prostate Cancer with BRCA1 and BRCA2 Founder Mutations.* Clinical Cancer Research. Vol. 15, No. 3.

[2] Hao, Xiaoke, et al. *DNA methylation markers for diagnosis and prognosis of common cancers.* Proceedings of the National Academy of Sciences 114.28 (2017): 7414-7419.

[3] Genome.ifmo.ru. (2018). TCGA PRAD Dataset. [online] Available at: *https://genome.ifmo.ru/files/software/phantasus/tcga/PRAD/* [Accessed 29 Nov. 2018].

[4] Cancer.sanger.ac.uk. (2018). Cancer Gene Census. [online] Available at: *https://cancer.sanger.ac.uk/census* [Accessed 29 Nov. 2018].

[5] Civicdb.org. (2018). CIViC - Clinical Interpretations of Variants in Cancer. [online] Available at: *https://civicdb.org/releases* [Accessed 29 Nov. 2018].