Term Project Proposal
Cancer Prediction using Single Nucleotide Polymorphism Dataset

Tanjung Dion (201899213)
Fawwaz Dzaky Zakiyal (201899213)

Bioinformatics (Fall 2018)

1 Introduction

The

2 Vocabulary

- **Single Nucleotide Polymorphisms (SNP)**: a variation in a single base pair in a DNA sequence.
- **Somatic Mutations**: genetic alteration acquired by a cell that can be passed to the progeny of the mutated cell in the course of cell division. Somatic mutations are frequently caused by environmental factors, such as exposure to ultraviolet radiation or certain chemicals.
- **Gleason Score**: a measure (from 2-10) of the aggression of prostate cancer cells based on clinical pathology of prostate tissue.

3 Methodology

3.1 Dataset

The data used in this paper is publicly available through the National Cancer Institute GDC Data Portal, and can be found online under Project TCGA-PRAD. This data set consists of genomic information and clinical pathology reports belonging to roughly 500 patients who have been diagnosed with prostate cancer (a Gleason Score anywhere between 2 to 10). Of the genomic information in this dataset, we used the SNP profiles of each patient. These SNP profiles specifically contain patient somatic mutations along with the position of each mutation on their genomes. The clinical pathology reports were also critical for our analysis because they contain the Gleason scores for each patient, which allows us to determine whether a patient has HGPCa. With respect to how the samples were actually drawn, all patients were sequenced using the Illumina HiSeq 2000 Sequencing System shortly after they were diagnosed with prostate cancer.

3.2 System Design

It compare three Learned models.

3.3 Preprocessing

Significant amounts of pre-processing needed to be performed in our to make the data usable in any way. First (and most challenging), we had to consolidate our data across hundreds of clinical pathology reports and SNP profiles. After consolidation, we constructed a feature matrix with 500 rows and 19,950 columns in which each row corresponds to a patient and each column corresponds to a single SNP.

We will also applied the adjustments described below to each column in order to normalize our data and using Principal Component Analysis (PCA) as feature selection algorithm to eliminate the useless SNPs and reduce the search space.

3.4 Classifier Algorithm

Perform a classification for cancer prediction :

1. K-Nearest Neighbor (KNN)
2. Support Vector Machine (SVM)
3. Multilayer Perceptron (MLP)

3.5 Testing & Evaluation

We were interested in finding a model that has the maximal classification accuracy, given by the metric 1, where y_i' is the predicted label for instance i, y is the true label for instance i, and $I()$ is an indicator function, in which it returns 1 if the statement within it is true and 0 if the statement within it is false. We used this metric to evaluate our accuracy on the cross-validation and testing sets.

$$accuracy = \frac{1}{n}\sum_{i=1}^{n} I(y_i' = y_i) \tag{1}$$

References

[1] A. I. Baba, H. Lu, T. B. Pedersen, and X. Xie. *Handling false negatives in indoor RFID data.* In MDM, pages 117–126, 2014.

[2] B. Fazzinga, S. Flesca, F. Furfaro, and F. Parisi. *Cleaning trajectory data of RFID-monitored objects through conditioning under integrity constraints.* In EDBT, pages 379–390, 2014.