Università degli Studi di Torino - M.Sc. in Stochastic and Data Science - A.Y. 2021/2022
Final project of Statistical Machine Learning (MAT0043)

Gene selection for cancer type classification

From the DepMap Portal website[1] we downloaded a dataset containing 1032 cancerous cells whose $\sim$ 17000 columns represent the probability that the inhibition of that given gene stops the growth of the cancer. The label of each cell line refers to the cancer type and is a categorical factor which can assume 10 values: Eye, Gastrointestinal, Gynecologic, Musculoskeletal, Neurological, Breast, Head-Neck, Hematologic, Genitourinary, Lung. The project then focuses on the problem of Genes Selection for Cancer Classification, i.e. finding a relatively small number of genes to predict the type of cancer of a given tumorous cell. To this aim we explored in detail three Features Selection algorithms: Random Forests combined with Feature Importance, Lasso-SVM and Neural Networks (NN) combined with Permutation Importance. We repeated these procedures on three distinct classification tasks: blood cancer vs rest, lung cancer vs rest and multiclass. Apart from the cases involving lung data, we achieved satisfying classification accuracies on the whole dataset and we were able to select about 100-200 genes from the starting 17000 ones. We then used these features to fit reduced versions of the previous classifiers and obtained classification accuracies which range from 67% to 98%. We then concluded that the possibility of classifying the type of cancer from an extremely reduced numbers of genes depends on the cancer type itself. In particular, these methodologies worked incredibly well on blood cancer for which we achieved almost 100% accuracy with the reduced classifier while failed miserably on lung cancer.

1 Introduction

Cancer is a complex disease characterized by the uncontrolled growth of abnormal cells anywhere in the body. These abnormal cells are extremely invasive and we usually identify them with the name of their original tissue (for instance, breast cancer, lung cancer, brain cancer, etc.). In recent years, medicine has made a great step forward in finding new and efficient therapies for different diseases and cancer is one of them. In particular, thanks to numerous advances in technology, collecting huge amount of data is no longer an issue, so that one can exploit such information to define personalized treatments for patients. In this regards, the DepMap project and, in particular, the Achilles project aim to use genome-wide screens to collect data regarding mutations of cancerous cells, identify essential genes and report vulnerabilities across hundreds of human cancers.

Many researches are currently using DepMap datasets to identify relatively small sets of genes which are responsible of cancers growth. This procedure is often driven by medical knowledge, which we do not possess, together with some rough measures of importance. Being Maths student, we instead base our research on statistical models only and on the hypothesis that if a given classifier is able to distinguish different types of cancer then the most relevant genes are the most important features for that given classifier (the meaning of "important features" will be clarified later). This finding would have outstanding implications in the medical field such building faster diagnosis tools and synthesizing less toxic drugs that target only these specific genes.

[1] DepMap Portal: https://depmap.org/portal/

2 Dataset

We used two public datasets from the DepMap Portal website[2]

D1 *CRISPR_gene_dependency.csv*, which contains 1.032 cancer cell lines characterised by 17.393 gene scoring results

D2 *sample_info.csv*, which contains cell lines information, such as primary disease and sample collection site

Data were collected from real patients and successively processed, so that element (i, j) of this (1.032×17.393)-data frame is the probability that knocking out gene j has a real depletion effect on the i-th cell. In order to work on a labeled datates and perform classification we associated the primary disease (i.e. the name of the cancer type) to each cell line performing an inner join of D1 and D2 on the DepMap ID. Then, we dedicated ourselves in data cleaning: we removed missing values, which affected only 10 rows, and all instances labeled as "Non Cancerous" and "Engineered".

We then group the various cancer types in 10 classes according to common medical knowledge[3] and we obtain the groupings as reported in Figure 1. "Eye" is the smallest one as there are only 16 observations and 5 of them are labeled while "Gastrointestinal" is the largest class as it comprehends 7 types of cancer, making this group quite heterogeneous. The project was then developed by investigating two binary classification problems, Blood vs Rest and Lung vs Rest, and the multiclass problem. The choice of Lung was dictated by the fact that it is the most numerous class while the choice of Blood was driven by some underlying biological knowledge. In fact, Blood cancer is quite different from other tumors because

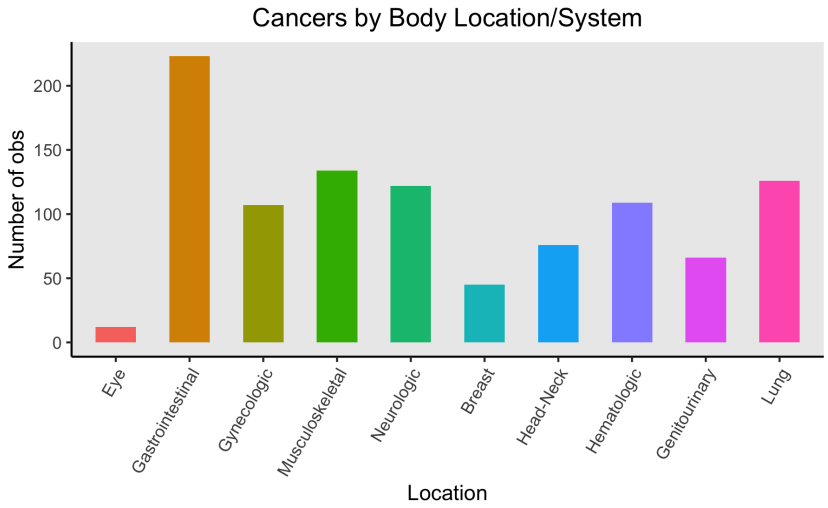

Figure 1: Cancer classes

- blood is in the whole body
- Leukemia, Lymphom and Myeloma are the main kinds of cancer but they all affect white blood cells
- not all blood cancers require a treatment, just periodical monitoring

[2]Download dataset from DepMap: https://depmap.org/portal/download/:
[3]Cancer types grouped by body location: https://www.cancer.gov/types/by-body-location

3 Methods

3.1 Random Forests

3.2 SVM-Lasso

3.3 Neural Networks

4 Results

5 Conclusion and future works