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

Given a high-dimensional genomics data, the purpose of our project is to find a relatively small number of genes to predict the cancer type of ill cells. This is known as Genes Selection for Cancer Classification and it is in line with many up-to-date problems of applied medicine.

Our dataset contains 1.032 cells and their knock-out probabilities, i.e. the probabilities of stopping the growth of cancer by inhibiting one of the $\sim$ 17.000 genes. Each cell line is characterised by one of 10 possible cancer labels: Eye, Gastrointestinal, Gynecologic, Musculoskeletal, Neurological, Breast, Head-Neck, Hematologic, Genitourinary and, finally, Lung.

We explore in details three Features Selection algorithms: Random Forests (RF) combined with Feature Importance, Lasso-SVM and Neural Networks (NN) combined with Permutation Importance. We have studied two binary classifications (Blood-cancer vs. All, Lung-cancer vs. All) and multiclassification. Besides lung models, we achieves satisfying classification accuracies and we are able to select about 100-200 genes from the initial $\sim$ 17.000 ones. Models fitted on such genes obtain classification accuracies ranging from 67% to 98%. Therefore, it seems that classifying cancer type from an extremely small set of genes depends on the cancer type itself.

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.). Whenever damaged or unrepaired cells do not die and become themselves malignant cells with uncontrolled division and growth, we say that a mass of tumour cells has developed.

In recent years, medicine has made a great step forward in finding new and efficient therapies for different diseases and, specifically, for cancer. 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. Examples of these works are carried by the DepMap project. Here, the Achilles sub-project aims 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 selection of genes is often driven by rough measures of importance together with medical knowledge, which we do not possess. Being Maths student, we instead base our research only on statistical models and on the hypothesis that relevant genes are the most important features for a given classifier, where the meaning of "important" will be clarified later.

Of course, selecting few truly significant genes have outstanding implications in the medical field. Indeed, they can be used not only in building faster diagnosis tools but also in synthesising less toxic drugs that target only these specific genes.

2 Dataset

We use two publicly available datasets, both found on the DepMap Portal website, [1, 2]:

- *CRISPR_gene_dependency.csv*, which contains 1.032 cancer cell lines characterised by 17.393 gene scoring results;
- *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 each element of this (1.032×17.393)-matrix is the probability that knocking out a gene has a real depletion effect on such cell.
First of all, we look for any missing values: only 10 rows have empty columns, specifically either 678 or 1.285 Nas. At first impact, this could seem a big deal, but it is actually the 4% and 7% of the total genes. Moreover, these cells come from different tumours, so we decide to simply remove them all.
Before proceeding with our analysis, we also notice some weird observations: 2 of them are labelled as "Non-Cancerous" and 6 as "Engineered". The first can be reasonably discarded as our goal is classifying cancer cells, whereas the latter requires a little care. Engineered cells are synthetically modified samples in lab and, here, they are manly associated to the Eye sample collection site. We keep them and we associate them the cancer type of the site.

One could ask: why do we focus only on the primary disease and not also on the sample collection site, as done for Engineered observations? If we count the number of cells with respect to cancer types and collection sites, we find many peculiarities. For instance, some Brain-cancer cells have been picked from the abdomen, whereas Lung-cancer cells comes from a variety of different places. This is because of the nature/curse of cancer: metastasis are ill cells identifiable as the original tissue but found on a different site. Therefore, this subdivision would only complicate our task.

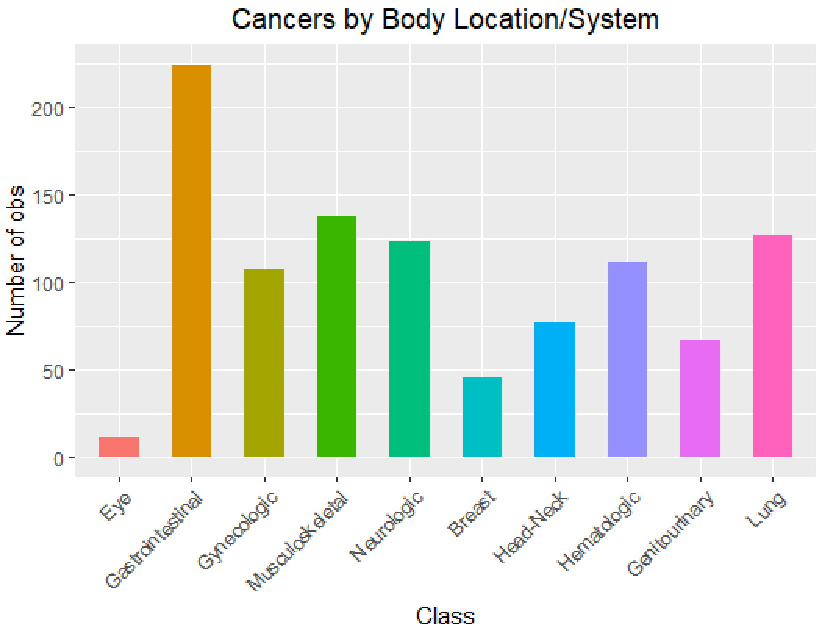

Figure 1: Cancer classes

We group the various cancer types in 10 classes according to common medical knowledge, [4], and we obtain the classes as reported in Figure 1.
We see that class "Eye" is the smallest one: there are only 16 observations and 5 of them are labelled as Enginereed, previously referred as *weird observations*. On the other hand, "Gastrointestinal" is the largest class and it comprehends 7 types of cancer, making this group quite dispersive. We hence grasp that these two classes could cause some problems in classification. Still following our intuition, we decide to focus our One-Vs-All binary classification on the "Lung" class. Since all these samples have been originally labelled as lung tumours, we suppose they do not suffer from noise caused by grouping together different disease.
Finally, we choose to study the "Hematologic" group relying on some underlying biological knowledge. In fact,

Blood cancer is quite different from other tumours because:

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

3 Methods

4 Results

5 Conclusion and future works

References

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

[2] DepMap, Broad (2021): DepMap 21Q3 Public, figshare. Dataset: https://doi.org/10.6084/m9.figshare.15160110.v2

[3] Project Achilles: https://depmap.org/portal/achilles/

[4] Cancer types grouped by body location: https://www.cancer.gov/types/by-body-location