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 cell lines whose ~ 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 9 values: 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 purpose we explored in details 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 fit the reduced versions of those classifier 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 we observed that blood cancer is the one which separates the most from the other and indeed we achieved almost 100% accuracy using only 100 genes while lung cancer is the worst behaved so this reduced classification task was unfeasible. In the multiclass task we achieved good classification results with few variables as well and our suspects about lung and blood cancer were confirmed.

1 Introduction

Cancer is the uncontrolled growth of abnormal cells anywhere in the body. These abnormal cells are termed cancer cells, malignant cells, or tumor cells. These cells can infiltrate normal body tissues. Many cancers and the abnormal cells that compose the cancer tissue are further identified by the name of the tissue that the abnormal cells originated from (for example, breast cancer, lung cancer and brain cancer). When damaged or unrepaired cells do not die and become cancer cells and show uncontrolled division and growth - a mass of cancer cells develop. 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 them to define personalized treatments for patients. An example of this is the DepMap project and, in particular, the Achilles project whose aim is to collect data regarding mutations of cancerous cells using genome-wide screens, identify essential genes and report vulnerabilities across hundreds of human cancers.
Many researches have been done using DepMap datasets and their goal was generally to identify a relatively small number of genes which are responsible for the cancer growth. The choice of genes is often driven by medical knowledge, which we do not possess, together with some rough measure of importance. Being Maths student we instead based our research only on statistical models and we hypothesized that important genes are the most important features for a given classifier, where the meaning of "important" will be clarified later. The implication of selecting a few truly important genes would have outstanding implications in the medical field.

It can be used for example to build faster diagnosis tools or to synthesize less toxic drugs which target a small number of specific genes.

2 Dataset

We use two publicly available datasets, both found on the DepMap Portal website:

- *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 the 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 sample in lab and, here, they are manly associated to the Eye sample collection site. We keep them and we associate the cancer type according to the sample.

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 type and collections site, 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.

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

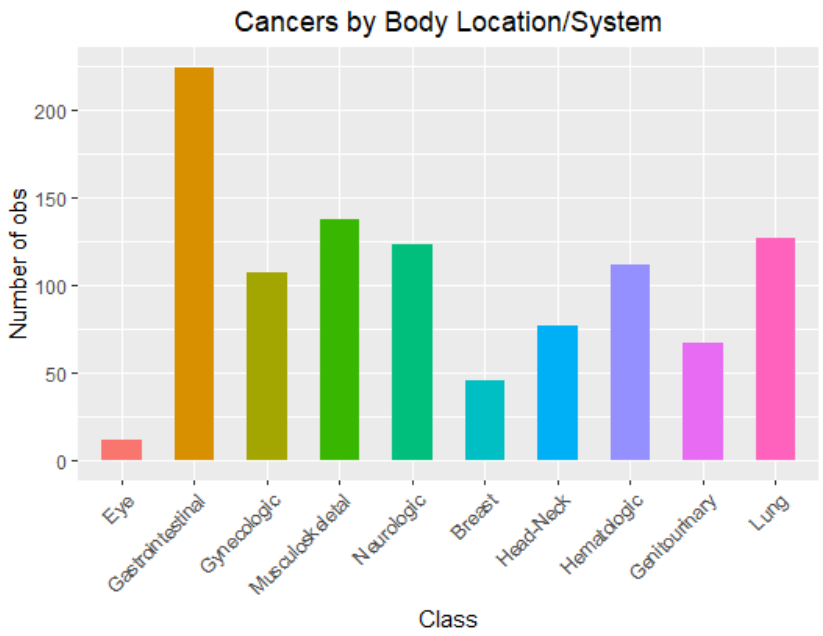

Figure 1: Cancer classes

samples have been originally la-
belled as lung tumours, we suppose they do not suffer from noise caused by grouping together different disease. On the other hand, 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