Status report for molecular recommendation system

Zhouyang Shen 2359009s

December 16, 2020

1 Project description

1.1 Motivation

Over the last decade, the extraordinary understanding of the inner functioning of tumours has been developed by the enormous scale endeavours of molecular profiling of human cancer. The data that has been utilized in those experiments has reflected the massive complexity of the changes across the genome, proteome of tumours, and the transcriptome. Consequently, the conventional statistical modelling of making useful predictions will be limited due to the fact stated above. Therefore, another innovation of using the data-driving method to analyze the information is applied to help with overcoming the difficulty. Besides, with the information perceived in the article "Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer", the idea of utilizing the features from 5 platforms and mutational signatures from TCGA platform to classify different cancer is gained.

1.2 Aims

In this project, we aim to investigate what a tumour is? And We will answer this question by building a recommendation system that can estimate the molecular characteristics of tumours given incomplete data. To be more specific, The project contains two-part of the job. The base section is to find the appropriate features from 5 platform and mutational signatures and classify the cancer of the tumour based on them. Hopefully, we will find those features that are useful when it comes to classifying the cancers accurately, and we will assume that they are the patterns for the specific cancers. The second part of the work is to predict the cancer type of the tumour given the sample id using the pattern we found. Furthermore, if the time is sufficient enough, the convolutional neural network will also be applied to help to improve the accuracy of recommendation and reduce the possible problems that might be elicited from the data sparsity.

2 Progress report

- Read the article“Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer”and comprehend the methods and concepts it used to classify cancer. The idea of classifying cancers based on the features from different platforms and mutational signatures are also developed after the researching.
- Read the article ”HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures” and understand the approach it utilized for classification and realized that the features it utilized would be used classify the different type of tumours.
- Statistically researched the features and the sample id of the files from TCGA platform and learned the process of how the mutational signatures could be used for classification.
- Learned the random forest algorithm and logistic regression algorithm to help with classifying.
- Learned statistical hypothesis testing,regulation method for improving the classification process.
- Extract the features using the algorithm provided.
- Wrote the code to classify the tumours based on the single based substitution signature part.
- Wrote the classification algorithm and framework for classifying the tumor based on the mutational signatures(single based substitution signature and rearrangement signatures).

3 Problems and risks

3.1 Problems

The following issues were encountered in the project so far.

- The tools that extract the features requires the specific type file as input where i do not have the authority and the application of the access of those data requires lots of time and procedures,the file is hard to obtain.
- The data set is enormous, and it requires massive memory for processing data, a server is probably needed in the future development.
- Some of the samples does not have the indel and rearrangement related signatures(incomplete data).
- some features such as rearrangements which are found using matrix factorization are approximated values. Thus, the accuracy of the model could be affected by them.
- Some of the cancer types of somatic mutation data is hard to find on platform and missing.
- The classification accuracy will be reduced after the involving of the incomplete data.

3.2 Risks

- The classification is performed on the whole cancer types, some of the features extracted might be missing due to the incomplete sample data of each cancer type, thus,if some of the features do share heavier weight in classification,the accuracy of the classification might be affected **Mitigation**: will try to extract the features as complete as possible, if can't, The classification will be mainly focused on those features that are available after extraction.
- The accuracy of the classification might not be optimal, and it might cause false diagnosis. **Mitigation**: will try different classification algorithm to improve accuracy.
- Unclear how to evaluate the success of the project. **Mitigation**: will do background research to investigate how the success of classification based on molecular features has been performed in the research literature.

4 Plan of work

4.1 Semester 2

- Week 1-3: Do research on the article and try to append other features(indel and hrd index) into classification and adding 5 other platform's feature and writing dissertation **Deliverable:** The improved recommendation system that use required molecular features to predict the sample's cancer type
- Week 4-5: Investigate on the related articles about the convolutional neural network that is used on classification and writing dissertation **Deliverable:** construct the idea of how to apply the convolutional neural network on the project to improve the accuracy of the recommendation system
- Week 6: ensure the method that is going to be applied to the recommendation system, and implement it and writing dissertation **Deliverable:** The molecular recommendation system with a convolutional neural network applied to help with improving accuracy.
- Week 7-9: final implementation and improvements to molecular recommendation system and writing dissertation **Deliverable: polished software ready, basic passing tests, ready for evaluation stage.**
- Week 9: evaluation experiments run and writing dissertation **Deliverable: quantitative measures of usability and qualitative measures of effectiveness for at least ten users.**
- Week 8-10: Write up. and polishing dissertation **Deliverable: first draft submitted to supervisor two weeks before the final deadline.**

5 Ethics

This project will involve tests and personal data(cancer type, etc.) from human users. I have verified that the experiments I plan to do comply with the Ethics Checklist and I will sign and complete the checklist. and the data protection impact assessment will be performed during the process.