COMP4730 Project Proposal

Topic 1: Breast Cancer Subtypes

Kolby Sarson - 104232327

Brandon Ferrari - 104396553

Abstract

The project we have decided to work on is “Topic 1: Breast Cancer Subtypes.” The problem we will be focusing on solving is feature reduction and selection as there is a massive number of features in this dataset (13,582). This many features should be reducible to a much more manageable and favourable set of features. We will look at reducing the number of features used when diagnosing patients by eliminating the features with the least, or no information gain while keeping the prediction accuracy as high as possible. Doing so will provide many benefits to both performance and efficiency.

The Problem

The problem presented to us is breast cancer classification. Currently, there are 13,582 features presented in the dataset when classifying breast cancer by type (Normal, LumA, LumB, Her2, Basal). With this many features, it is likely that the information gain for a considerable number will be minimal at best. Also, using such a large number of features could have many negative implications to the classification of the data. These negative implications include long processing times, hardware constraints, accuracy loss from irrelevant data, false patterns being presented, and others. For these reasons, we wish to investigate the possibility of feature

reduction for this dataset. In a perfect world, the end goal would be to minimize the number of features used in our machine learning algorithms, while achieving as close to 100% accuracy as possible. The purpose of this feature reduction is to help lower the possibility and probability of these negative impacts on the classification. Feature reduction could reduce processing times and prevent hardware constraints from becoming an issue, as well as remove irrelevant data and false patterns from the dataset.

Solution

Using the Scikit-Learn Python library, we will examine possible solutions to this feature selection problem. First of all, there are several methods in the realm of feature selection that would help us identify the genes that carry the most weight within this dataset. This means we would be able to generate scores for each gene, and minimize the dataset to the top k features, with k being a chosen number which maximizes prediction accuracy. For example, the paper given to us chose k=20, or the top 20 features to represent the dataset. Many tests will be done to identify the optimal k value as well as the features that provide the most information gain. On the topic of information gain, we may also utilize decision trees to solve our problem. While a decision tree itself would not provide any new information, we may use our knowledge learned in class to calculate the information gain of each feature and optimize the decision tree. This would provide the same effect as our feature selection problems and eliminate genes that provide minimal to no information gain. Other methods outside of our realm of knowledge may also be explored as we do more research on the problem itself. Once the dataset is cleaned up, we will use a portion of the dataset as the training data, and the remaining portion as the testing data (usually a 70/30 split, respectively). The training data will be used to train a classification algorithm of our choosing, and the test data will be used in predictions and accuracy calculation.