ML4S Project Proposal

Jon, Ray, Shashank, Rohit

Problem Statement

Accurate subtyping of cancer types is crucial for targeted therapy and improving patient outcomes but is hampered by traditional methods' time-consuming nature and variability in interpretation. As cancer's complexity unfolds with advances in genomics, the manual classification of its subtypes becomes increasingly impractical. So our project proposes leveraging machine learning (ML) to overcome the classification problem of subtyping cancer types into more refined subtypes. First, we aim to implement traditional clustering models to cluster the data, then use several classifiers for subtyping the cancer type.

Summary of Related Works

Back in 2004, Au, NHC, et al. studied 284 cases of Non-Small Cell Lung Cancer (NSCLC) by performing unsupervised hierarchical clustering on them based on 18 biomarkers They obtained four subclusters of NSCLC, two of which showed significant proportions of different subtypes of NSCLC (squamous cell and adenocarcinoma), indicating the efficacy of unsupervised clustering in identifying subtypes of cancer.

In 2021, Ferro, Sara, et al. performed unsupervised clustering of Primary Breast Cancer samples by using multiple algorithms. They found hierarchical clustering to be the best. The 712 cancer samples were split into two subgroups with different properties: age, estrogen receptor, and progesterone receptor amounts.

These works indicate that such unsupervised clustering of cancer samples can help identify subgroups within cancer categories and aid in better tumor classification.

Overview of Methods

Clustering

By examining the RNAseq data for differential expression of genes in various cancers, we intend to implement K-means clustering to find distinguishable groups in the high-dimensional space, which could lead to further subtyping of parent groups of cancers. K-means clustering will involve finding the right K meta parameters, for which we can employ the hockey stick method.

PCA

Once we've clustered the data into cancer subtypes, we hope to more easily visualize how distinguishable these subtypes are by employing dimensionality reduction techniques such as principal component analysis to project the high dimensional data down onto a 2D plot, coloring data points according to the clusters found in the prior step.

Classification

Finally, with distinguishable clusters, we intend to build a classifier working off of the clustered data, verified with PCA, so that cancers of the parent type can be subclassed into the distinct groups found in the prior steps. This will likely be done with a kNN classifier, as it can easily handle nonlinear boundaries and multiple classes. We also experiment with implementing a shallow MLP classifier as a comparative approach.

Evaluation Metrics

We will use a number of metrics to evaluate the performance of our model. Specifically, in regard to our classification model, we will implement cross-validation to reduce overfitting and assess the performance of our folds. In addition, we will implement the following metrics: accuracy, precision, recall, and f1-score. We implement all of these metrics, with an emphasis on f1-score, to account for class imbalances. To evaluate our PCA and clustering methods, we will implement scikit-learn's equivalent packages as a benchmark.

References

- Au, N. H., Cheang, M., Huntsman, D. G., Yorida, E., Coldman, A., Elliott, W. M., Bebb, G., Flint, J., English, J., Gilks, C. B., & Grimes, H. L. (2004). Evaluation of immunohistochemical markers in non-small cell lung cancer by unsupervised hierarchical clustering analysis: a tissue microarray study of 284 cases and 18 markers. *The Journal of pathology*, *204*(1), 101–109. https://doi.org/10.1002/path.1612
- Ferro, S., Bottigliengo, D., Gregori, D., Fabricio, A. S., Gion, M., & Baldi, I. (2021). Phenomapping of Patients with Primary Breast Cancer Using Machine Learning-Based Unsupervised Cluster Analysis. *Journal of Personalized Medicine*, *11*(4), 272. https://doi.org/10.3390/jpm11040272