Project Proposal: Cancer Classification by Gene Expression Data

Ziyuan Shen

February 14, 2019

1 Problem Description

Over the past few decades, machine learning techniques have been increasingly applied to disease studies [1]. The development of single-cell technology greatly facilitates the availability of multi-omics data [2, 3], including genetic data. Accordingly, delicately designed data mining and machine learning techniques are desired to explore the biological information underlying the large datasets. Relevant studies can aid disease diagnosis and precision medicine. Cancer classification is one of the top promising topics [4]. By far, clinical datasets, genetic datasets [5, 6] as well as integrated datasets have been used for cancer classification studies. Improving the classification performance is undoubtedly meaningful as human beings' overall health conditions will be improved by these efforts.

Although clinical data, such as patient information and tumor appearance, can certainly be employed for cancer classification, a biological intuition is that cancer subtypes are more dependent on patients' genetic components. In fact, patients with similar clinical appearance do have completely different therapy responses [7]. Therefore, genetic datasets should play an important role in cancer class discovery and prediction. Different from other kinds of data, genetic data is cursed by high dimensionality due to the large number of gene types. In addition, cancer classification historically requires prior biological knowledge. However, most machine learning engineers do not have a biological background. In conclusion, the general classification performance with high dimensional data is worth improving. Prediction models with high accuracy are aimed to aid oncologists with diagnosis and prognosis.

With the advancement of biochips, genetic expression data can be obtained from thousands of genes. The goal is to create a good predictor with samples from known classes, and possibly identify hidden gene expression patterns related to each class, without any biological priors. Most existing works develop models based on specific cancer types [5]. Thus, a more general and systematic procedure remains open-ended.

2 Data Description

2.1 Background

The dataset initially comes from the published paper named "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" [5]. Gene expression data is used to classify patients with two types of leukemia: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Through "Neighborhood Analysis", the author validated that more than 1000 genes are highly correlated with the cancer class. Therefore, cancer subtypes have the potential to be classified with only gene expression data, without any prior biological knowledge.

The dataset contains two classes with very high feature dimensions, thus should serve as a good example for conducting classification algorithms.

2.2 Summary Statistics

The data contains two datasets: training data with 38 samples and testing data with 34 samples. 28.9% of the training data and 41.2% of the testing data are from the AML patients. Statistics are summarized in Table 1. The data is split into the training and testing sets because the first 38 points are simultaneously collected bone marrow samples at the time of diagnosis. The next 34 points are independently collected samples consisting of 24 bone marrow samples and 10 blood samples. The original work [5] treats the two datasets as training data and testing data. However, this is not a strict rule to follow and the two datasets can be certainly combined for arbitrarily split for training and testing.

Table 1: Summary Statistics of Gene Expression Datasets.

Statistics \ Datasets	Training Data	Testing Data
#**Total Samples**	38	34
#**AML Patients** (Proportion)	11 (28.9%)	14 (41.2%)
#**ALL Patients** (Proportion)	27 (71.1%)	20 (58.8%)
#**Dimensions**	7129	

2.3 Data Visualization

In order to conduct data visualization, I combine the training and testing datasets together. T-SNE visualization is shown in Figure 1. A 3D plot for three components is provided as well as 2D plots for every two components. As the number of features of this dataset is extremely high, Principle Component Analysis (PCA) is more appropriate for visualizing the data. According to PCA, the first three components only contains 29% of the variance. However, due to the limitation of the dimensions that can be visualized, the first three components can still be used to visualize our data. PCA data visualization is shown in Figure 2.

2.4 Source

The source data can be downloaded from kaggle website [8] in the form of csv files.

3 Anticipated Technical Approach/Method

3.1 Preprocessing

Normalization: z-Scoring may be applied to scale data to zero mean and unit variance.

3.2 Dimension Reduction

As the dataset has greatly high dimensions, an approach for dimension reduction will be important to the prediction performance.
Principle Component Analysis: Principle Component Analysis (PCA) will be a useful tool for dimension reduction.
TruncatedSVD: TruncatedSVD is helpful when data is sparse.
Feature Selection: Apart from PCA, feature selection may also be a good way to reduce dimensions.
Trade-offs: Some anticipated trade-offs are summarized in Table 2.

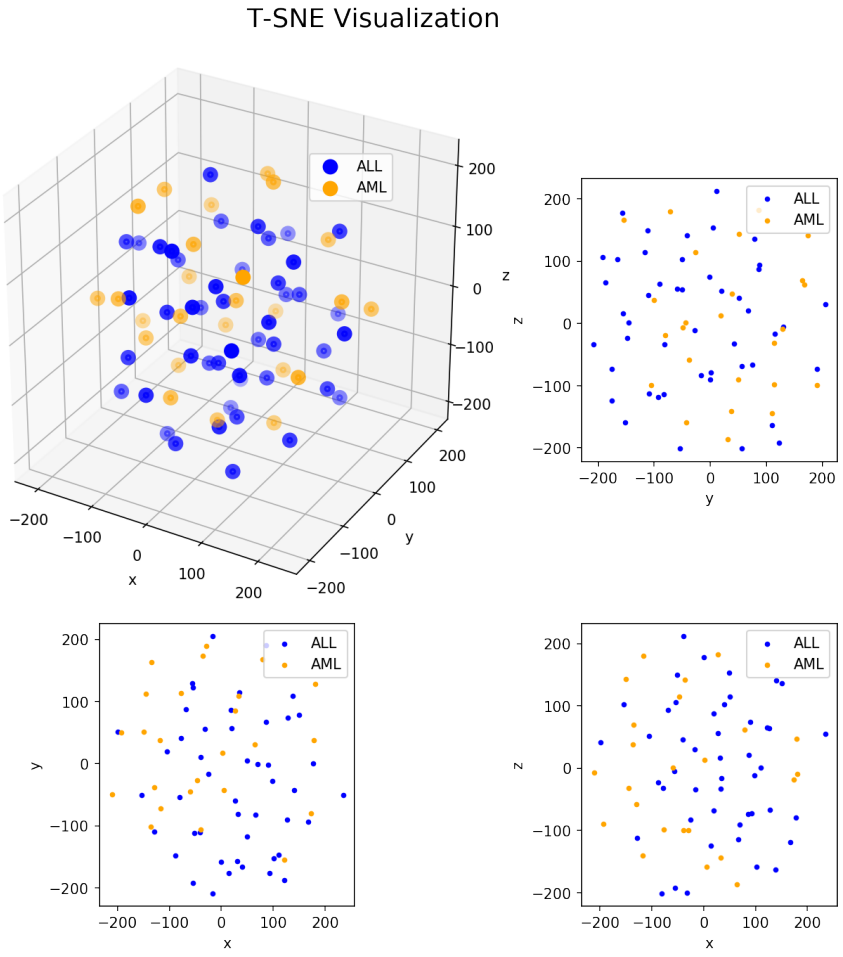

Figure 1: T-SNE visualization.

Table 2: Trade-off Analysis of Dimension Reduction.

Trade-offs	**Fewer Features**	**More Features**
Interpretability	High	Low
Overfitting	Less likely	Likely
Prediction Accuracy	Low	High

3.3 Classifying Algorithms

Defined as a classification problem with two classes, the dataset can be tested by the following algorithms:

1) **Naive Bayes**
2) **Logistic Regression**
3) **Support Vector Machine**
4) **K-Nearest Neighbors**
5) **Decision Tree Classifier**

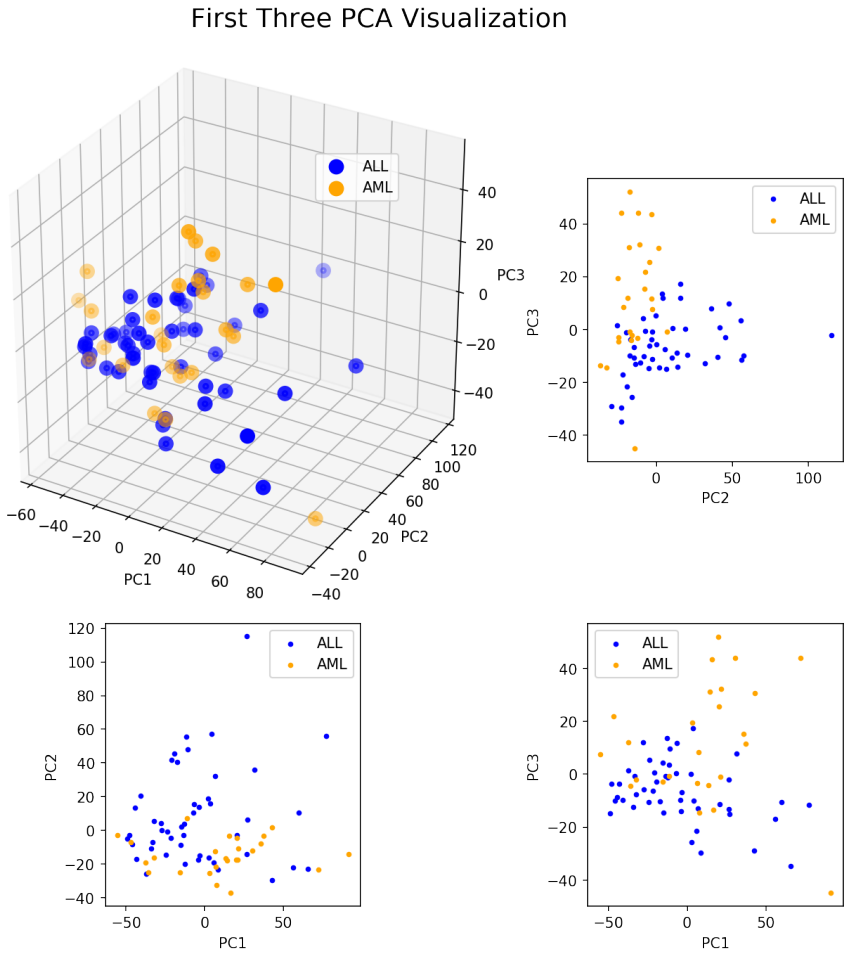

Figure 2: First three principle components visualization.

3.4 Cross-Validation

As the number of samples is limited in the given dataset, cross-validation will be necessary to evaluate performance of the predictor.

References

[1] P. Sajda, "Machine learning for detection and diagnosis of disease," *Annu. Rev. Biomed. Eng.*, vol. 8, pp. 537–565, 2006.

[2] M. Ibrahim and K. Chakrabarty, "Cyber-physical digital-microfluidic biochips: Bridging the gap between microfluidics and microbiology," *Proceedings of the IEEE*, 2017.

[3] M. Ibrahim, K. Chakrabarty, and U. Schlichtmann, "Synthesis of a cyberphysical hybrid microfluidic platform for single-cell analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2018.

[4] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics*, vol. 2, p. 117693510600200030, 2006.

[5] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science*, vol. 286, no. 5439, pp. 531–537, 1999.

[6] X. Huang and W. Pan, "Linear regression and two-class classification with gene expression data," *Bioinformatics*, vol. 19, no. 16, pp. 2072–2078, 2003.

[7] Y. Liu, X. Yin, J. Zhong, N. Guan, Z. Luo, L. Min, X. Yao, X. Bo, L. Dai, and H. Bai, "Systematic identification and assessment of therapeutic targets for breast cancer based on genome-wide rna interference transcriptomes," *Genes*, vol. 8, no. 3, p. 86, 2017.

[8] C. Crawford, "Gene expression dataset (Golub et al.): Molecular Classification of Cancer by Gene Expression Monitoring," https://www.kaggle.com/crawford/gene-expression, 2019.