Team Members: Andrew Wei, nowei@cs.washington.edu

Project: Leukemia classification using RNA gene expression levels

Paper Survey: *Read 3 or more papers on your project topic. Give detailed summaries of the papers and compare how they differ from your approach.*

My project was on Leukemia classification using RNA gene expression, so I surveyed papers that performed some form of cancer or cancer-related classification using gene data.

Cascianelli et. al used PCA on the PAM50 dataset to examine whether the subtypes of breast cancer were separable from each other [1]. They found that they could not separate the subtypes by the first two principle components and hypothesized that due to this non-separability, the boundary between subtypes changes depending on the mixed traits. They introduce the idea of Average of Within Class Averages (AWCA) to get a single number that represents a subtype. They find that this method produced decent results, ~90% accuracy. They then went on to try different machine learning approaches and found that regularized multiclass logistic regression performed the best.

> While they are planning on also doing regularized multiclass logistic regression, we differ in how we will process the data. while I plan to preprocess the data in a way that makes it easier to find how individual genes expression values from other groups differ from the genes expression values of healthy individuals. I also am not planning to use an AWCA approach to examine the data.

Danaee et. al used stacked denoising autoencoders to deal with the high dimensions and noisy inputs of gene expression data [2]. They wanted to use the encoder to try to find the genes that play an important role in breast cancer. They multiply the learned autoencoder weights together to get a rough sense of how the gene expression data interacts, saying that the most heavily weighted ones have the most significant impact/were the most predictive genes. They use the heavily weighted genes and use those as the inputs to DL models and achieve slightly less accuracy than with the SDAE features but argue that it is more readily interpretable.

> They approach the problem by using an autoencoder to reduce the dimensionality of the data and pipe the encoded input into a DL model. I did not originally plan on exploring dimensionality reduction schemes, opting to use Lasso regression, but I am open to exploring different dimensionality reduction schemes, time willing. I am also using multi-class logistic regression, but I'm also willing to look into DL approaches if I finish earlier.

Mei et. al trained a stacked autoencoder to provide input to a neural network trained on a subset of the TCGA dataset to identify survival duration for patients with Acute Myeloid Leukemia [3]. They include features like age, cytogenetics, and mutations and achieve results of around 81% accuracy for predicting whether patients survived for more of less than 730 days. They mention that the main limitation to their model was the amount of training data available.

> This approach was similar to the previous paper we discussed, but they include features like age that were not present in the previous paper. While they are working on classifying something related to Leukemia, the goal is still different. I’m still not planning on utilizing DL approaches, as I’m trying to focus on the interpretability of my results, but there are some ideas that I’m going to try if I have time.

Castillo et. al say that they are one of the first studies in this space to try to identify between 5 classes, 4 with leukemia and one healthy class, since most research only classifies between two classes [4]. They use the minimum-Redundancy Maximum-Relevance feature selection algorithm and train SVM, Random Forests, K-Nearest Neighbors, and Naive Bayes classifiers. They then extract Differentially Expressed Genes (genes that help differentiate between classes) by testing p-values, log-fold change (measures change in gene expression level), and coverage (whether it helps identify between some number of classes). They found that K-Nearest Neighbors worked the best, followed by SVM and Naive Bayes, with the worst being Random Forests. We can likely compare our results to the results found in the study.

> This study is very closely related to what we’re attempting, with the only difference being the classification schemes and the way that we’re processing the data. They also do some form of dimensionality reduction, but their approach is dissimilar to previously mentioned ones. They use statistical significance tests to essentially narrow down the genes that they need to pay attention to. They also try a handful of different classification algorithms. Them saying that K-Nearest Neighbors is surprising though, since it’s likely one of the simpler learning algorithms and it may be worth trying. Although it’ll likely perform worse due to the high number of dimensions if I do not decide to reduce the dimensions somehow.

Krivtsov et. al outlines procedures on how to extract the gene expression data and utilize some form of either hierarchical or K-means clustering and use permutation analysis to determine significance for the purpose of profiling Leukemia Stem Cells [5]. They use the GenePattern software to distinguish marker genes that would distinguish the two groups, comparing means, using a signal-to-noise statistic, and evaluating significance. This was the only study read that seemed to have extracted and recorded the data themselves instead of taking the data from a database.

> I am not planning on isolating a leukemia stem cell population or amplify RNA. They do some form of dimensionality reduction as well, utilizing some sort of significance testing, then use those genes in K-means. K-means is also an explorable option if I have time, but I will see where the data takes me.

Progress report: *Describe your detailed plan on improving your method and analyzing data with it. Explain how you will evaluate your method and compare with baseline methods. Provide preliminary results if available.*

We are planning to leave out 10% of the data by class as testing data and doing 10-fold cross validation on the remaining data to evaluate the approach of standardizing with respect to the data from healthy patients. Thus, each fold will hold out 10% of the remaining data for each class to make sure that the data for a class will be present for classification, i.e. we don't try to identify something that we have no training data for. Standardizing with respect to the healthy patients will be compared to standardizing the data with respect to each gene in the training data and no standardization. We plan on testing Logistic Regression and testing different regularization terms. We hypothesize that the results with L1 regression will likely be the most informative, since it incentivizes the model to give features a weight of 0.

The baseline for comparison will be the best Logistic Regression model that has no standardization with the best regularization terms that we could find.

We plan to measure top-1 and top-5 accuracy, since there are 18 sub-types of Leukemia, which makes classification potentially noisy, so the top-5 accuracy may give us some insight on how close the model was to predicting the correct classification.

We will analyze the weights learned from the logistic regressions to get a better understanding of which genes play a crucial role in classifying each sub-type of Leukemia. We can then compare these results to previous studies on different sub-types of Leukemia to see if the features we have picked up matched results from previous studies.

If we have time, we can explore reducing the dimensionality by keeping only statistically significant gene expression data. We can find these genes by performing t-tests on the training data and comparing them with the data from healthy individuals, correcting for multiple comparisons with the Bonferroni correction, separate for every class. Then if any gene is statistically significant for any of the 18 subtypes of Leukemia, we can use that to train the final model. We are doing this later, as we want to explore the robustness of logistic regression on a large number of features. If we have extra time after that, we will explore other classification schemes like K-Nearest Neighbors or utilize Deep Learning models.

References

[1] Cascianelli, S., Molineris, I., Isella, C. *et al.* Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer. *Sci Rep* **10,** 14071 (2020). https://doi.org/10.1038/s41598-020-70832-2

[2] Danaee, P., Ghaeini, R., Hendrix, D. A Deep Learning Approach for Cancer Detection and Relevant Gene Identification. PSB (2017).

[3] Mei Lin, Vanya Jaitly, Iris Wang, Zhihong Hu, Lei Chen, Md. Amer Wahed, Zeyad Kanaan, Adan Rios, & Andy N. D. Nguyen. (2018). Application of Deep Learning on Predicting Prognosis of Acute Myeloid Leukemia with Cytogenetics, Age, and Mutations.

[4] Castillo, Daniel et al. "Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level." *PloS one* vol. 14,2 e0212127. 12 Feb. 2019, doi:10.1371/journal.pone.0212127

[5] Krivtsov, Andrei V et al. "Gene expression profiling of leukemia stem cells." *Methods in molecular biology (Clifton, N.J.)* vol. 538 (2009): 231-46. doi:10.1007/978-1-59745-418-6_11