Acute Myeloid Leukemia Project

Yizhou Pan
pany25@uw.edu

Shihao Qiu
sqiu2@uw.edu

Xingzhi Niu
nxz18@uw.edu

1. ABSTRACT

Acute myeloid leukemia (AML)is a cancer of the blood and bone marrow. A prior project that was initiated during NCBI Hackathon in February 2019 focused on applying automated feature selection and machine learning models to predict the risk groups and patient survival time. Our project is a further development on testing different machine learning models and modifying clustering of AML phenotypes based on the data and results produced in the prior one.

2. INTRODUCTION

Acute Myelogenous Leukemia (AML) is type of leukemia impacting the myeloblast stem cells. According to Tyner *et al.* [1] and Yi *et al.* [2], AML arises approximately 20,000 cases per year with 27.4% 5-year survival. It is molecularly heterogeneous, with many subtypes defined by factors including sequencing assays.

Event free survival time (EFST) of AML patients varies in a broad range based on subtypes, gene mutation types, age of patient and many other factors. Gene sequencing expression level may work as a good evidence for the factors of subtypes or mutation types. In our work, gene expression and the result of sequencing assays are used to predict EFST of AML patients. We use clinical and assay data from pediatric cancer patients from the Therapeutically Applicable Research To Generate Effective Treatments (TARGET) initiative (https://ocg.cancer.gov/programs/target/).

There are 6 members in the prior AML project: Vikas Peddu, Ryan Shean, David Lee, Jenny Smith and Ronald Buie. They mainly focus on three tasks: using gene expressions to predict patient survival time as well as risk groups based on different machine learning models and automated feature selection and clustering of AML phenotypes. Our further exploration is from building up models to predict survival time and modifying clustering of phenotypes.

We extend the work of Ryan Shean [3] in NCBI Hackathon 2019. Shean used the TARGET assay data with 108 observations. EFST of the observations are split into two groups, longer than median (331 days) and shorter than median. A binary classification model was built to predict EFST of one patient is whether longer or shorter than median with the expression level of 46 genes. Lasso regression [4] is built and 67% precision reached.

We consider Shean's work to be trivial and coarse. To perform a fine-grained prediction, we build models to predict the concrete number of EFST in days instead of a binary classification. We define the input of each observation as the gene expression level G, binary features B.

$$G = \{g_1, g_2, .., g_m\} \tag{1}$$

$$B = \{b_1, b_2, .., b_n\} \tag{2}$$

In which g_k denotes the gene expression level of the k^{th} gene and b_k is -1 for the k^{th} mutation not detected, 1 for detected and 0 for unknown binary feature. The output is simply an integer t, representing the predicted EFST of the patient in days.

For the selection of genes and binary features, a simple method is applied. Since both gene expression and EFST is numeric or nearly continuous value, correlation coefficient between particular gene and EFST of the 108 observations can be calculated. Genes are further ordered decreasingly by the absolute value of the correlation coefficient. For binary features, signal to noise, or S/N are calculated by

$$S/N = \frac{|\bar{t_1} - \bar{t_{-1}}|}{\sigma_1 + \sigma_{-1}} \tag{3}$$

in which $\bar{t_1}$ is the mean of EFST of observations with the binary feature, $\bar{t_{-1}}$ is the mean of EFST of observations without the feature, and σ_1 and σ_{-1} is the standard deviation. Binary features are ordered decreasingly by S/N too.

We propose three different approaches to predict the EFST by gene expression and mutation data.

3. METHODS

3.1 Clustering of AML Phenotypes

In the prior project, Buie applied k-means unsupervised classification to group the available phenotype data and finally set number of clusters equal to 4 as his optimum result. In our project, we applied two hierarchical clustering approaches with complete linkage and Ward's method respectively. At last, We compare the clustering plots generated by these two approaches and Buie's.

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

3.1.1 Complete Linkage Hierarchy Clustering

Complete Linkage Hierarchy clustering splits each cluster by the shortest distance. The mathematical definition of the linkage function is defined as

$$D(X,Y) = max_{x \in X, y \in Y} d(x,y)$$

3.1.2 Ward's Hierarchy Clustering

Ward's hierarchy clustering splits each cluster based on the optimal value of an objective function. The function we apply here is the squared Euclidean distance between points, which is defined as

$$d_{ij} = d(X_i, X_j) = ||X_i - X_j||^2$$

3.2 Multi-Classification on Event Free Survival Time of AML Patients

3.2.1 Summary

In WT patientsâĂŹ clinical assay, I try to find relation between RNA-seq gene expressions. After studying previous work from Ryan Shean on this field, I want to find other models to better describe the relation. In RyanâĂŹs project, he/she uses t-test to do feature selection and uses boosted generalized linear model to give prediction. In my plan, I decide to do feature selection with correlation coefficient and to divide Event Free Survival Time into four classifications but not numeric data. In my case, I compare two different multi-classification methods. These two methods are both based on binary classifier. But different strategy will cause different result.

3.2.2 Feature Selection

Feature selection method in this case is Pearson product-moment correlation coefficient. Pearson correlation coefficient is usually used to rate the relation between two variables. Given a pair of random variables (X, Y), the formula of Pearson correlation coefficient is:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{4}$$

cov is the covariance
$\sigma_X \quad and \quad \sigma_Y$ are standard deviations of X and Y
The covariance can be calculated like following:

$$cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] \tag{5}$$

$\mu_X \quad and \quad \mu_Y$ are the means of X and Y E is the expectation

No matter continuous variable or discrete variable, there is always formula to calculate the expectation.

The main idea of feature selection is to calculate the Pearson correlation coefficient between the Even Free Survival Time and each RNA-seq gene expression. Then we want to remove those coefficients which have p-value lower than certain threshold. In my case, I want to compare the results with different thresholds.

3.2.3 Classifier

The basic binary classifier used here is k-nearest neighbors algorithm. In kNN classification, the classifier receives training data together with its classification information first. According to these existing correct classification, when receiving an unknown instance, classifier finds the k nearest correct training instances first. Then it finds the class with the most frequency among the k instances and assigns this class to the unknown instance. There are some ways to define *nearest*. The most common method is Euclidean distance. We can treat each feature as a dimension. If we got N features, we have a N-dimension space. Distance between instances is the distance of two points in the N-dimension space.

We don't know importance of each feature, so I don't add weight to the classifier.

3.2.4 Multi-Classification

There are two multi-classification methods in this case. One is *One vs. One* and another one is *One vs. Rest*. These two methods are both based on binary classification. In this case, target classification is Even Free Survival Time which is numeric classification. In previous work, EFST is divided into two parts, young and old, by its median. For more detailed classification, I divide EFST into four parts by quadrisection. *Level1* indicates 0% 25% survival time, *Level2* indicates 25% 50% survival time and so on. Then here are two methods.

3.2.5 One vs. One

OvO makes a pair between every two classifications. Therefore, if we have N classifications, OvO generates N*(N-1)/2 binary classification tasks. After this, we need to complete these binary classification tasks. In our case, we have four classifications. So we need to complete six binary classification tasks. That is six classifiers. After training, when receiving one instance, it needs to go through six classifiers. The six classifiers vote for the final result of this instance. There may be situations that two classification have the same number of ballot. When this happens, I will choose one of them randomly.

3.2.6 One vs. Rest

OvR is less simple than OvO. Suggest we have N classifications, OvR generates N binary classification tasks. Each task solves the problem that whether one instance belongs to a classification. So the result of each task will be True or False. Expected situation is each instance only have one classifier gives True result. But things are not always what we think. If there is no True result or there are more than one True results, I will select one from all classifications or select one from True classifications randomly.

3.3 Predictive Model on AML Patient Survival Time

3.3.1 Linear Regression

By selecting the genes with top correlation coefficient, we can assume the linearity between the top genes with EFST. Linear regression, which optimizes the residual square error, is a traditional method to build regression model. Baltagi [5] introduced stepwise regression model to eliminate useless variables to avoid overfit for regression models. We choose a backward stepwise model with selected features to build full model, and use Akaike information criterion, shortly AIC [6] to eliminate. For benchmarking the performance of different dimensions of features, no more than 10 features will be eliminated from the full model. Binary features B is concatenated to gene expression G as the input of a vector with $m + n$ dimensions.

3.3.2 Support Vector Regression

Cortes and Vapnik [7] proposed a machine learning method, support vector machine, shortly SVM as a powerful binary classifier. SVM recognizes the observations located near the border between two categories as support vectors. The norm of two support vectors in two categories is defined as the margin. The optimization target is to minimize the margin. Based on the support vector, the classifier can be fine tuned with a slack variable to fit the training samples better. Druker *et al.* [8] further proposed a method using SVM in regression tasks. The kernel function together with the fine tuned slack variable will draw a border between categories, and distance of the support vector to the border is considered as the margin. For each sample located out of the margin, there is a punishment. The optimization of SVM regression (SVR) is to minimize the punishment. After training, the feature vector can be input to the kernel function of SVR model and get the predicted value.

3.3.3 One versus Rest SVM

Practically, EFST accurate to days is beyond the requirement. Actually we can predict EFST with a acceptable range of error of tens or hundreds days with a higher accuracy. In our work, we transform EFST in days to EFST level with 100 days. For example, EFST shorter than 100 days is level 0, longer than 100 days and shorter than 200 days will be level 1. Regression models can still be applied after the transformation. Although accuracy of EFST levels grows, the error in days does not improve significantly.

We turn to a multi label classification work instead of regression. Notice that after transforming EFST in days to EFST levels, each observation belongs to one level. To avoid the upper bound of EFST, EFST longer than 3000 could be categorized into level 30. That changes the prediction of a nearly continuous number to the prediction of several categories. Most classifiers are designed for binary classification and there are some tricks for multi label classification. Kim *et al.* [9] proposes a multi layer method to compare the result of different classifiers. Thomas and Sael [10] build a hybrid one versus one model with multiple classifiers and combine the result by comparing the output of each pair of classifiers of one entry.

We proposes a one versus rest [11] method, which build multiple classifiers. Each classifier represents the border of two adjacent EFST levels. For the k^{th} classifier, it will classify by the border of the $k-1^{th}$ level and k^{th} level. The samples of $k-1^{th}$ level or lower are labelled as negative samples, and the samples of k^{th} level are labelled as positive samples. Suppose there are l levels to be classified, there will be $l-1$ classifiers. After training all $l-1$ classifiers, each entry will be predicted in sequence begin with the 1^{st} classifier, until the q^{th} classifier predicts the entry as negative. That means EFST of the entry should be of $q-1^{th}$ level.

3.3.4 Experiment

We take five experiments in our work. They are the stepwise linear regression, support vector regression, support vector regression with level of 100, SVM multi classification with level of 100, and logistic regression multi classification with level of 100. For the first three experiments, we split the 108 observations into 9 batches, each with size of 12. One batch for test and 8 for training. For the last two experiments, we take the 80 observations whose EFSTs are shorter than 1000 days and remove the observations whose EFSTs are longer than 1000 days because of the sparsity of EFST in the range greater than 1000. We split 80 observations into 10 batches, each with size of 8. One batch for test and 9 for training. For the last two classification tasks, we take the average of two bounds of each level as the predicted EFST in days. For example, all entries of level 6 will be predicted as 650. In each experiment, we tried the combination of m and n, i.e. the number of genes and binary features. There are 57 genes with correlation coefficient greater than 4.0 and 20 binary features with S/N greater than 0.1. We tried the combination of the top 10, 20, 30, 40, 50 and 60 genes with the top 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10 binary features (mutations). We use residual standard error on the test set to evaluate the performance of the methods. Every experiment is repeated 9 times and the average is taken. The following table shows the residual standard error and the combination of m and n for the best case. Detailed output is appended in files.

Table 1: RSE on test set

Method	*RSE*	*m genes*	*n mutations*
Li Regr	17.06	60	4
SVR	15.55	20	9
Discrete SVR	14.99	60	9
Multi SVM	14.57	60	10
Sum	14.53	30	3

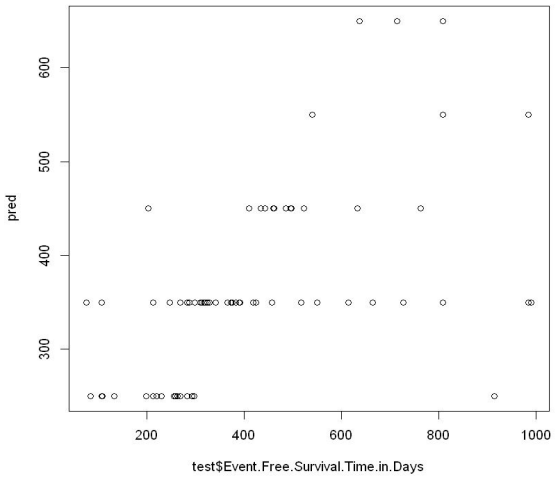

Figure 1: Predicted and true EFST in days

4. RESULTS

4.1 Clustering of AML Phenotypes

Attached three figures show the comparison of these three clustering methods with 4 clusters.

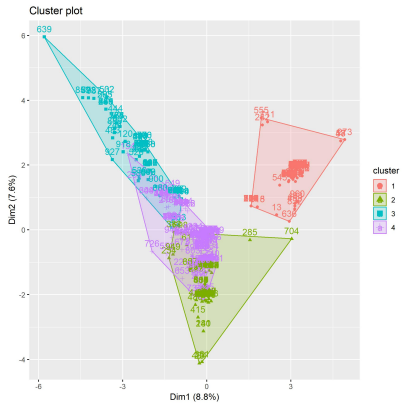

Figure 2: K-means Clustering with 4 Clusters

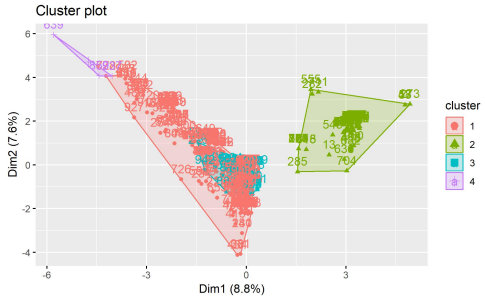

Figure 3: Complete Linkage Hierarchy Clustering with 4 Clusters

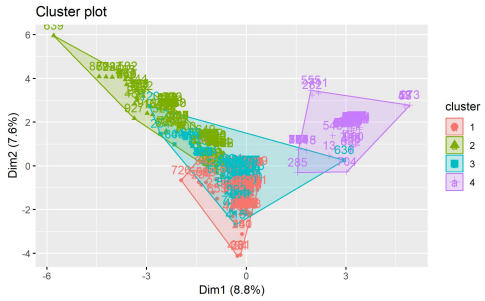

Figure 4: Ward Hierarchy Clustering with 4 Clusters

In order to evaluate the performances of these three models, we need to evaluate the inner cohesion and external separation of these clusters. We use the Between Sum of Squares and total sum of squares ratio to evaluate the performance. For 4 clusters with k-means, this ratio equals to 37.2 % which is quite low. As the number of cluster increases, this ratio will go up and when the number of cluster reaches 10, this ratio can go up to 58.4 %.

4.2 Multi-Classification on Event Free Survival Time of AML Patients

In my experiment, I compare couple groups of variable and they are showed below in TABLE I and TABLE II.

Table 2:

One vs. One			
Accuracy	p.value=0	p.value=0.05	p.value=0.1
k=3	0.28125	0.21875	0.25
k=4	0.234375	0.296875	0.25
k=5	0.34375	0.34375	0.359375

Table 3:

One vs. Rest			
Accuracy	p.value=0	p.value=0.05	p.value=0.1
k=3	0.1818182	0.2424242	0.1515152
k=4	0.1969697	0.1969697	0.2727273
k=5	0.2272727	0.1969697	0.3181818

We can see from the tables, the best accuracy is given by One vs. One method. It is easy to explain that we use six classifiers to give results among four classifications. Under the same case, One vs. Rest only use four classifiers. It is obvious that the previous one may be more precise. And the Pearson correlation coefficient can only measure linear correlation. But we have no evidence to prove that it is linear correlation between Event Free Survival Time and RNA-seq gene expression. So we can see in the table that stricter condition doesn't bring better accuracy. In all, my conclusion is that Pearson correlation coefficient feature selection method seems not to perform very good in this case. Also kNN is not a pure binary classifier, other better binary classifiers may perform better in this case. Actually I am not very satisfied with this result. If given more time, I can do more comparison and analysis.

4.3 Predictive Model on AML Patient Survival Time

As a result of the experiments, we build predictive models to predict EFST of AML patients within an acceptable error. Furthermore, we affirm the assumption that to train models with EFST level of 100 days will improve the performance. We also propose the novel model of one versus rest multi classifier to predict EFST level of 100 days and reaches best performance.

As fig.1 shown, most predicted samples are located near the main diagonal, and the predicted values are discrete into levels of 100.

5. DISCUSSION

There are many possible future improvements on our work. Exploring the width of the level could be an important extension. Another possible direction may be to build heterogeneous models for different element of the input feature vector, since the role of them are not the same, especially for the gene expressions and mutations. Furthermore, sociological factors may significantly impact EFST of a disease. The TARGET data contains many such factors that can hardly be quantized, of which further works can take advantage. A combination work with the research on phenotypes of AML could also be interesting and make improvement on the prediction of EFST.

6. REFERENCES

[1] J. Tyner, C. Tognon, D. Bottomly, B. Wilmot, S. E. Kurtz, S. L. Savage, N. Long, A. Reister Schultz, E. Traer, M. Abel, A. Agarwal, A. Blucher, U. Borate, J. Bryant, R. Burke, A. Carlos, R. Carpenter, J. Carroll, B. Chang, and B. J. Druker, "Functional genomic landscape of acute myeloid leukaemia," *Nature*, vol. 562, 10 2018.

[2] G. Yi, A. T. Wierenga, F. Petraglia, P. Narang, E. M. Janssen-Megens, A. Mandoli, A. Merkel, K. Berentsen, B. Kim, F. Matarese, A. A. Singh, E. Habibi, K. H. Prange, A. B. Mulder, J. H. Jansen, L. Clarke, S. Heath, B. A. van der Reijden, P. Flicek, M.-L. Yaspo, I. Gut, C. Bock, J. J. Schuringa, L. Altucci, E. Vellenga, H. G. Stunnenberg, and J. H. Martens, "Chromatin-based classification of genetically heterogeneous amls into two distinct subtypes with diverse stemness phenotypes," *Cell Reports*, vol. 26, no. 4, pp. 1059 – 1069.e6, 2019.

[3] D. Lee, S. Maden, V. Peddu, R. Shean, J. Smith, and R. Buie, "Rnaseq cancer biomarkers: Aml rna-seq biomarker and feature selection with ml," 2019.

[4] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc B.*, vol. 58, pp. 267–288, 01 1996.

[5] B. Baltagi, *Multiple Regression Analysis*, pp. 73–93. 04 2011.

[6] H. Akaike, "Information theory and an extension of the maximum likelihood principle," *2Nd International Symposium on Information Theory*, vol. 73, pp. 1033–1055, 01 1973.

[7] C. Cortes and V. Vapnik, "Support-vector networks," in *Machine Learning*, pp. 273–297, 1995.

[8] H. Drucker, C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Linear support vector regression machines," vol. 9, pp. 155–161, 01 1996.

[9] S. Kim, T. Park, and M. Kon, "Cancer survival classification using integrated data sets and intermediate information," *Artificial Intelligence in Medicine*, vol. 62, 09 2014.

[10] J. Thomas and L. Sael, "Multi-kernel ls-svm based bio-clinical data integration: Applications to ovarian cancer," *International Journal of Data Mining and Bioinformatics*, vol. 19, 04 2017.

[11] J. Xu, "An extended one-versus-rest support vector machine for multi-label classification," *Neurocomputing*, vol. 74, no. 17, pp. 3114 – 3124, 2011.