The University of New Mexico

Undergraduate Honor Program

Statistical Modeling of Genomics Data to Improve Cancer Risk Classification and Outcome Prediction

Author:

Mo Li

Supervisor:

Dr. Huining Kang

May 1, 2015

Contents

1 Introduction 4

2 Literature Review 5

3 Method 9

3.1 Data set description . . . 9

3.2 Gene Filtering Approaches . . . 11

3.3 Modification of PAM . . . 12

3.4 Model evaluation . . . 15

3.4.1 Leave one out cross validation . . . 15

4 Results 17

4.1 Task 1: Induction Failure VS. Complete Remission . . . 17

4.1.1 Threshold Selection and Performance Evaluation by Nested LOOCV . . . 17

4.1.2 Performance Evaluation Using an Independent Data Set . . 19

4.2 Task 2: High-risk VS. Low-risk . . . 21

4.2.1 Threshold Selection and Performance Evaluation by Nested LOOCV . . . 21

4.2.2 Performance Evaluation Using an Independent Data Set . . 22

5 Discussion 23

Acknowledgment

First of all, I want to show my deepest gratitude to my supervisor, Dr. Huining Kang, a very respectable, responsible and resourceful scholar, who show valuable guidance in each stage of the research process. Without his enlightening instruction, I could not complete my honor thesis. Many thanks to the Gabrielles Angel Foundation, who funded this research. I would like to extend my thanks to Dr. Stuart S Winter and Dr. Ksenia Matlawska-Wasowska, University of New Mexico Department of Pediatric for providing the gene expression and patient data of Leukemia. Last but not least, I'd like to thank everyone, who give me a hand during the research process.

Abstract

The DNA microarray gene expression profiling technique has been widely applied in cancer research. The Prediction Analysis of Microarray (PAM) implemented in the statistical software R has been frequently used in classification with microarray gene expression data. However, when the classes of the data set are unbalanced, the software tends to make prediction towards to the majority class, leading to low sensitivity whereas in cancer screening test high sensitivity is often desired. We propose to select the optimal prediction model using the average error rate instead of using the over error rate as implemented in PAM. We also propose to treat the ratio of prior probabilities in the prediction model as an unknown parameter which will be determined by minimizing the average error rate in order to improve prediction performance. We applied the proposed approaches to risk classification and outcome prediction of T-Cell acute lymphoblastic leukemia using gene expression microarray data. The result indicates that our proposed approaches have effectively improved the prediction performance.

1 Introduction

Acute lymphoblastic leukemia (ALL) is the most common cancer diagnosed in children and represent approximately 25% of cancer diagnoses among children younger than 15 years [1]. The treatments for children diagnosed with ALL include induction, consolidation and maintenance therapy along with central nervous system (CNS) prophylaxis[1-5]. The intensity of treatment are usually determined based on the risk groups defined by both clinical and laboratory features. The patients with favorable clinical and biological features are classified into low risk groups. They are likely to have a very good outcome with modest therapy and can be spared more intensive and toxic treatment. The more aggressive, and potentially more toxic, therapeutic approach can be given to patients who have a high risk to fail the induction therapy or to relapse. Thanks to the development of the treatment including those based on risk classification, the cure rate for the childhood ALL has improved dramatically over the past four decades. Currently more than 80% childhood ALL have achieved long-term remissions [1].

However, significant challenges still remain. There are still 20% patients that have either failed the induction therapy or relapsed after therapy. The second common cause of cancer-related mortality in children in the US remains relapse ALL. To achieve high remission rate, up to one-third of children are likely over-treated. They may well be cured using less intensive therapy resulting in fewer acute toxicities and long term side effects. The key to improve the treatment outcome is to find a way to improve the risk classification. It has been hypothesized that the genomics technologies that measure global patterns of gene expression in leukemia patients will discover genes and gene profiles that may ultimately be useful for developing classification models based on gene expression profiles to improve the risk classification, thus risk-based treatment [7].

In this thesis we developed some models using gene expression microarray data that can be used to predict the early response and long term outcome of T-Cell childhood ALL. The early response is categorized as induction failure (IF) and complete remission (CR). The long term outcome refers to as the high risk and low risk groups where high risk group consists of patients who failed induction therapy (i.e. IF) and those how relapsed within four years of follow-up, and the low risk group consists of patients who have achieved continuous complete remission for at least four years.

Based on the structure of microarray gene expression data, there are some main challenges to overcome.

1. The high dimensional gene expression data set has the large number of genes (variables) and a relative small number of samples, which can complicate the search of gene combinations.

2. The model is prone to overfitting since there are super abundance genes (variables) but relative a small number of samples, which can result in the poor predictive performance.

Many statistical methods have been developed to deal with above challenges such as the least absolute shrinkage and selection operator (LASSO), ridge regression, random forest, prediction analysis of microarray (PAM). In the medical field, PAM has been widely used on the risk classification and outcome prediction tasks, which implement the nearest shrunken centroids (NSC) to improve the prediction performance on cancer risk classification [8].

2 Literature Review

PAM is a statistical classification technique using the NSC method to identify subsets of genes that best characterize each class and build classification rule with the selected subset of genes [9]. The classification rule of NSC is also developed from the rules in diagonal linear discriminant analysis (DLDA), which focus on the scaled distances between the expression values of samples and the class average expression value. The basic idea of NSC is to shrink the class centroid (average gene expression level) towards the overall centroid of each gene. The distance between these groups is small when the groups contain the non-differentiated expressed genes. Then these non-differential genes will be removed by the shrinkage value. The differential expressed genes will survive and combine together to contribute to the final classification model [8]. The method of ranking genes can be thought as the penalized t-statistics. The best shrinkage value will be determined based on the performance of each model in the cross-validation steps. Finally, the PAM use the selected gene combinations to predict the outcome and to do classification on the new samples.

The literature review will continue by introducing the PAM method in details. Assuming we have n samples, each with expression data for p genes. The expression value for the ith gene of the jth patient is denoted by x_{ij}, which can be measured by gene chips ($i = 1, \ldots, p$ and $j = 1, \ldots, n$). Each sample belongs to class k, where k is in the index C_k ($C_k = 1, \ldots, K$). Let

$$\bar{x}_{ij} = \frac{\sum_{j \in C_k} x_{ij}}{n_k},$$

denote the centroid of each gene in class k, n_k is the number of patients in class k. The overall centroid of each gene can be written by $\bar{x}_i = \frac{\sum_{j=1}^{n} x_{ij}}{n}$, where n is the number of patients for all classes. Let

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k(s_i + s_0)}.$$

The d-score is introduced from shrinking the class centroid $\bar{x}_i$ by the overall centroid $\bar{x}_i$. The standardization is done by the pooled within-class standardization for gene i, where $s_i = \frac{1}{n-k} \sum_{k=1}^{k} \sum_{j \in C_k} (x_{ij} - \bar{x}_{kj})^2$, and $m_k = \sqrt{\frac{1}{n_k} + \frac{1}{n}}$. s_0 is a positive constant, which is used to avoid the possibility of large d_{ik} values arising by chance, from genes at very low expression levels. It is usually set to the median value of s_is. Let

$$d_{ik}{}' = \text{sign}(d_{ik}(|_{ik}| - \Delta))_+,$$

where Δ is the soft shrunken threshold value.

The technique of "PAM" shrinks each d_{ik} to $d_{ik}{}'$. Then the shrunken centroid for each class k for each gene i is denoted by $\hat{x}_{ik} = \bar{x}_i + d_{ik}{}' m_k(s_i + s_0)$.

This method constructs new centroid for each class, which will be used in the discriminant score calculation. Another advantage is that some noisy signals can be shrunken to 0 by the optimal threshold value (Δ), which will decrease the number of variables. The model development is based on the soft threshold value selection. To predict the class of new sample x^*, PAM defines a discriminant score for class k:

$$\delta_k(x^*) = \textstyle\sum_{i=1}^{P} \frac{(x_i^* - \bar{x_{ik}}{}')^2}{(s_{ij} + s_0)^2} - 2 \log \pi_k, \text{ where } \sum_{k=1}^{K} \pi_k = 1.$$

π_k is the class prior probability and is estimated based on class sample size in PAM with the default setting. ($\pi_k = \frac{n_k}{n}$). It is used to penalize the effect that samples in the larger class always tend to have more error, which tends to give more weights for the larger classes to reduce more errors.

The classification rule [9] is

$$C(x^* = k^*), \text{ where } \delta_k(x^*) = \min_k \delta_k(x^*).$$

The estimated posterior probability can be constructed by the discriminant score,

$$\hat{p}_k(x^*) = \frac{e^{-\frac{1}{2}\delta_k(x^*)}}{\sum_{l=1}^{K} e^{-\frac{1}{2}\delta_l(x^*)}}.$$

The threshold value (Δ) selection is based on the cross-validation, as illustrated in the method section.

Although the NSC approach is powerful and easy to use for researchers and has been widely used in many classification and prediction situations, there is still room for improvement on the method application when we apply PAM on data sets with different features and different situations. For example, Wang and Zhu proposed that the adaptive L_∞-norm penalized NSC and the adaptive hierarchically penalized NSC has better performance on classification and more effective to do the feature selection [11].

Another issue of PAM application is caused by high dimensional data with unbalanced classes. Due to the factors of experimental design or some unchangeable reasons, the unbalanced data is produced very often. The original PAM analysis algorithm is not adaptable when applying in the unbalanced data [12]. Especially for classification problems, unbalanced data biases classification toward the majority class, the class with larger sample size [12]. Additionally, when classification is applied in the high dimensional data, the effect can increase further. Lusa reported that when high dimensional data is unbalanced, the NSC classifier is biased towards to the majority class. Although PAM can minimize the overall error rate, it always has the high error rate on the minority class since PAM is sensitive to the unbalanced data [10]. For the medical risk classification tasks, the high-risk class is often the minority class, which needs to have the lowest error rate on prediction. Generally, feature selection can decrease the error rate of the minority class by reducing the majority class bias, but it still has affects on the prediction [10].

In this article, three options of modification on model development and validation about PAM will be introduced, which are based on different combinations of finding the ratio of the prior probabilities and different evaluation criteria on selection of threshold value Δ.

3 Method

3.1 Data set description

The gene expression data were collected by measuring the fluorescent intensity of the prob sets located on the Affymetrix Microarray chips. The raw data file were saved in a .CEL file for each patient. The R packages “affy” and “frma” were implemented to read the raw data into the R work space and to perform data preprocessing for the gene expression values using different methods, such as Robust Multi-array Average (RMA) and frozen RMA algorithms.

The standard microarray data set contains three components. They are the matrix of patients annotation, the matrix of Gene expression values and the matrix of gene annotations.

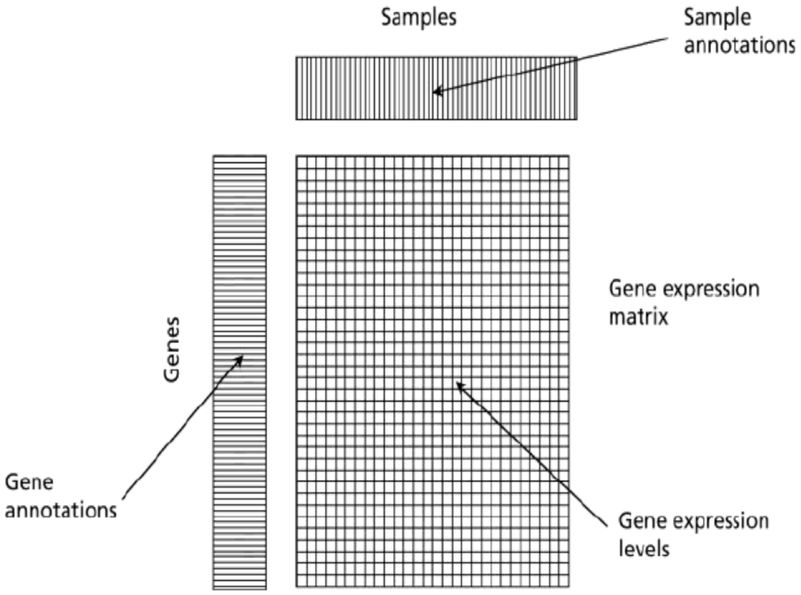


Figure 1: Data set format

The sample annotation matrix (9×213) contained the 213 patients' information, where one patient is in one column. The 4 variables, "patid","cohort","outcome" and "risk", from the of the sample annotation matrix were used in the risk classification model development and validation process. (The explanation of these variables can be seen in Table 1.)

Table 1: Patients Information from Sample Annotation Matrix

Variable Names	Number of Levels	Interpretation
patid	—	Patient ID
cohort	4	Cohort name of the data set
outcome	3	The recording leukemia patient outcome (CR: Complete Remission, RE: Relapse, ID: Induction Failure)
risk	2	High: ID+RE; Low Risk: CR

The gene expression matrix (54675×213) contained the 54675 gene expression values for each patients. The column names of this matrix were set up to the .CEL files' names for each patients. The row names were set to the probe set id for each gene. The gene annotation matrix (54675×7) contained the annotation information for the 54675 genes.

Totally there were 4 cohorts in these data matrix combination (9404, 0434_1, 0434_2 and 0434_3), which contained the patients in the same population but was collected in different groups. We applied the original PAM and our modified approaches on the 4th cohort, 0434_3, to develop our classification models and used the rest cohorts as an external independent data to get the unbiased performance evaluation of the selected models.

In this thesis, we were dealing with binary classification problems (IF VS. CR or High Risk VS. Low Risk). In order to investigate our approaches performance in the early response and in the long term outcome prediction situations. The patients were grouped into two kinds of different classes based on the definitions of early response and long term outcome separately. We can find the unbalanced structure in the data for both these situations according to the basic information of the two tasks in the Table 2 and 3.

The task 1 is to predict the early response with the unbalanced classes IF and CR. There are 17 patients in the IF class and 83 patients in the CR class in the model development data set, and there are 12 patients in IF class and 88 patients in CR class in the external independent validation data set. The total sample size is 100.

Table 2: Data Set Information for Predicting Early Response

Data Set for Model development				External independent Validation data set			
Sample Size	IF	CR	Number of Genes	Sample Size	IF	CR	Number of Genes
100	17	83	54675	100	12	88	54675

The task 2 is about the prediction in long term outcome with the unbalanced classes high risk and low risk. There are 25 patients in the high risk class and 75 patients in the low risk class in the model development data set, and there are 31 patients in the high risk class and 69 patients in the low risk class in the external independent validation data set. The total sample size is 100.

Table 3: Data Set Information for Predicting Long Term Outcome

Data Set for Model development				External independent Validation data set			
Sample Size	High	Low	Number of Genes	Sample Size	High	Low	Number of Genes
100	25	75	54675	100	31	69	54675

3.2 Gene Filtering Approaches

Unsupervised gene filtering is a step of removing genes with ignoring the labels of the samples. Gene filtering can reduce the dimensionality of the data and increase the statistical power in detecting genes differentially expressed by different experimental conditions. In addition, the computing efficiency is improved and the overfitting problems can be alleviated if we do the gene filtering to reduce the number of genes. According to the information in the gene annotation matrix, we removed the probsets that associated with the Affymetrix controls, the sex-related genes and the globins. We further removed the gene with expression value in the low inter quartile range (IQR). We filtered 50% genes based on IQR and 27252 genes were left for statistical prediction modeling.

3.3 Modification of PAM

In the binary classification problems, the number of classes K is equal to 2. The discriminant scores for class k $(k = 1, 2)$ are

$$\begin{aligned} \delta_1 &= \textstyle\sum_{i=1}^{P} \frac{(x_i^* - \bar{x}_{i1})^2}{\Delta_i} - 2\log\pi_1, \\ \delta_2 &= \textstyle\sum_{i=1}^{P} \frac{(x_i^* - \bar{x}_{i2})^2}{\Delta_i} - 2\log\pi_2, \end{aligned}$$

where π_1 and π_2 are the prior probabilities of the two classes. The prediction was based on the estimated probability of the sample belong to a certain group. In our case, the estimated probability can be written as

$$\begin{aligned} \hat{p}_1(x^*) &= \frac{e^{(-\frac{1}{2})\delta_k(x^*)}}{e^{-\frac{1}{2}\delta_1} + e^{(-\frac{1}{2}\delta_2))}} \\ &= \frac{1}{1 + e^{\frac{1}{2}\delta_1 - \delta_2}}; \\ \hat{p}_2(x^*) &= 1 - \hat{p}_1(x^*) \\ &= \frac{e^{\frac{1}{2}(\delta_1 - \delta_2)}}{1 + e^{\frac{1}{2}(\delta_1 - \delta_2)}} \\ &= \frac{1}{1 + e^{-\frac{1}{2}(\delta_1 - \delta_2)}}; \end{aligned}$$

In another hand, $\delta_1 - \delta 2$ can be denoted by

$\delta_1 - \delta 2 = \sum_{i=1}^{P} \frac{\alpha x_i * + \beta x_i^* + r}{\Delta_i} - 2\log\pi_1 + 2\log\pi_2 = \text{linear function} - 2\log\frac{\pi_1}{\pi_2}$

Then we can rewrite the probability of sample belong to high risk group formula as :

$\hat{p}_1 = \frac{1}{1+e^{a+log\frac{\pi_1}{\pi_2}}}$, where a is a linear function of x.
$\hat{p}_2 = 1 - \hat{p}_1$.

π_2/π_1 is the ratio of prior class probability, which can be affected by the unbalanced data set if we use the "PAM" with the default settings by setting the prior class probability by the class sample size, where $\pi_1 = n_1/n$ and $\pi_2 = n_2/n$.

The default decision rule in PAM can be rewrote as

$$\hat{p}_1 > 0.5, \text{new sample is classified to class 1;}$$
$$\hat{p}_2 < 0.5, \text{new sample is classified to class 2.}$$

The threshold selection corresponds to the classification model selection, where each threshold corresponds to one model. The optimal threshold is determined based on the prediction performance of the corresponding model. The prediction performance evaluation is based on the 4 possible prediction outcomes summarized in the two-way contingency table.

In our two different binary classification tasks, we formed a two way contingency table (Table 4) to address the concepts of sensitivity, specificity, overall error rate and average error rate by simply using the Bayes theorem. The true positive fraction (TPF) is the probability of the sample being predicted to class 1 given the real status is in class 1, which is called sensitivity. The false positive fraction (FPF) is the probability of the sample being predicted to class 1 given the real status is in class 2, which is 1 − Specificity. The overall error rate is the sum of the number of type-1 error and the number of type-2 error divided by total sample size. The average error rate is the sum of false positive rate and true negative rate divided by 2.

Table 4: Contingency Table for Binary Classification Problems

	Real Status	
Prediction	Class 1	Class 2
Class 1 ($p > 0.5$)	True Positive	False Positive (Type-1 Error)
Class 2 ($p < 0.5$)	False Negative (Type-2 Error)	True Negative

Since the PAM did not perform well on the unbalanced gene expression data, our modification on the feature selection was that we treated the prior probability ratio as a parameter, which will be determined by minimizing error rate from the cross-validation steps and be subject to the condition that sensitivity is greater than specificity. On another hand, since the overall error rate cannot help us correctly select the best threshold when applying on the unbalanced data, we derived using the average error criteria to evaluate and select the optimal threshold value, which can decrease the effect from the majority class and give more weights to the minority during error evaluation. The average error rate (E) can be calculated by the following equation,

$$E = W_1 E_1 + (1 - W_1) E_2,$$

where W_1 is the weights of the majority class, E_1 is the error rate of prediction on class 1 and E_2 is the error rate on class 2. Then the lowest average error ($\min E$) came with the optimal threshold shrunken value (Δ).

In the application on the two binary classification tasks, we combined the PAM default settings and our modification on feature selection with the performance evaluation (overall error rate and average error rate) Then we get 4 options, PAM default with overall error rate, PAM default with average error rate, our approach with overall error rate and our approach with average error rate.

3.4 Model evaluation

In the model evaluation part, we evaluated the model performance in terms of the the overall error rate, the average error rate, sensitivity, specificity and the Area Under the ROC Curve(AUC). The Receive Operating Characteristic (ROC) curve describes the performance of the predictions of a binary classification task by the calculation of the rates, which is based on the samples characteristics [13]. The curve is plotted by the true positive rate (sensitivity) against the false positive rate (1-specificity). Green and Swets (1966) states that the Area Under the ROC curve (AUC) is related to the probability of correctly prediction [14], which means that a bigger AUC value corresponding to better performance on the prediction of the binary classification system. In the model comparison part, the nested leave one out cross validation (LOOCV) [16] was processed to evaluate the model performance with little bias and the parallel computing were used to reach the effectively computation of nested LOOCV.

3.4.1 Leave one out cross validation

The nested LOOCV is one type of the exhaustive cross-validation technique for the model development and the model evaluation. It can be separated into two parts, the inner loop and the outer loop. The inner loop is used for the model selection and the outer loop is used for performance estimation with the decreasing bias [15]. In our case, the outer loop was used for evaluating the error rate for the selected model. The nested LOOCV procedure was following the steps showing in the Figure 2.

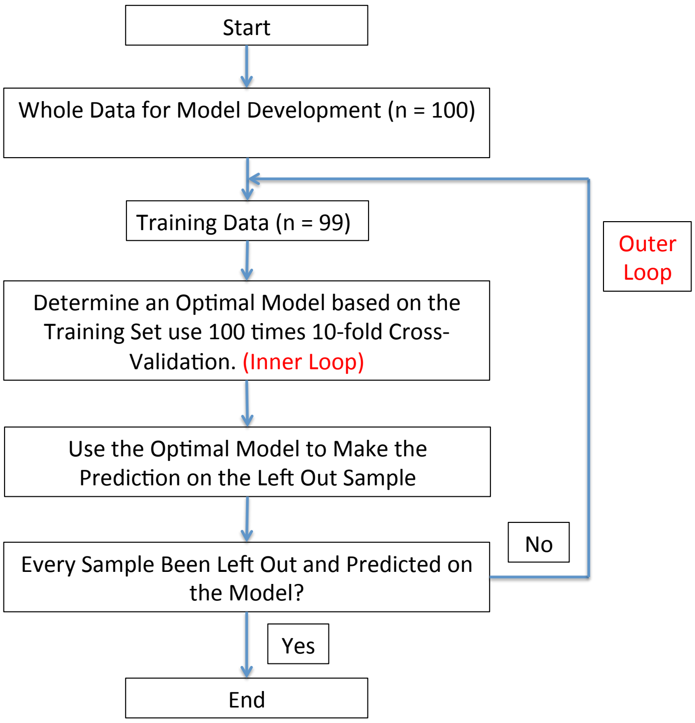

Start
Whole Data for Model Development (n = 100)
Training Data (n = 99)
Outer Loop
Determine an Optimal Model based on the Training Set use 100 times 10-fold Cross-Validation. (Inner Loop)
Use the Optimal Model to Make the Prediction on the Left Out Sample
No
Every Sample Been Left Out and Predicted on the Model?
Yes
End

Figure 2: LOOCV flow chart.

There are totally 100 samples in the cohort 0434_3 used to develop the classification model. One sample was left out as the test set to evaluate the model performance, and the left 99 samples were used to determine the model by the 100 times 10-fold cross-validation steps in the inner loop. After each sample using as the testing set to be predicted on, the loop finished and the mis-classification rate was calculated. Since the nested LOOCV calculation is time-consuming, the parallel computing was used to improve the computational efficiency by using the high performance computer in Center for Advanced Research Computing (CARC).

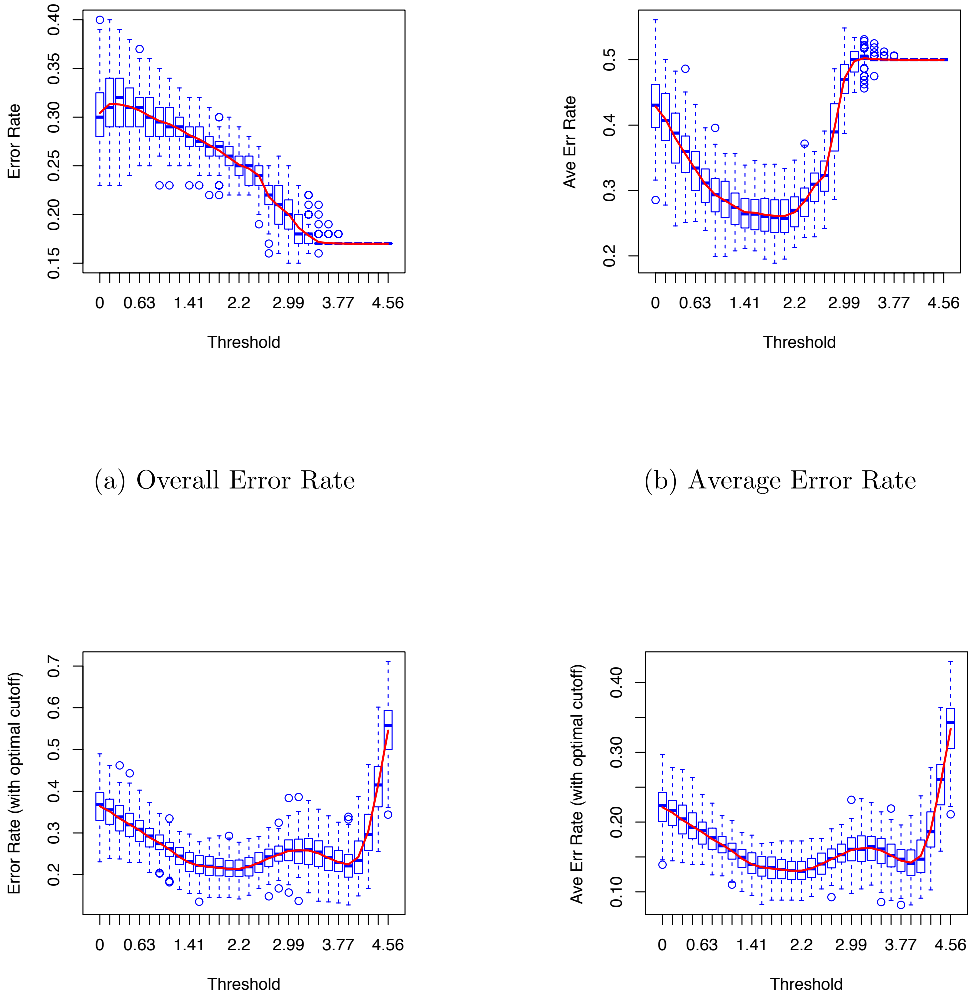
4 Results

In this section, I compared the results of the threshold selection by 4 different options and the performance evaluation of the selected models, which can address the improvement by applying our approaches on the unbalanced high dimensional data.

4.1 Task 1: Induction Failure VS. Complete Remission

4.1.1 Threshold Selection and Performance Evaluation by Nested LOOCV

Figure 3a shows the training, cross-validation, and test errors for different amount of the shrunken value Δ. The test errors were summarized from 100 times 10-fold cross-validation, which can decrease the random effects from the random partition. The box-plots in Figure 3a was ploted by 100 overall error rates from 100 times cross-validation for each threshold. The overall error rates keep decreasing from 0 shrinkage with 27252 genes (left) to complete shrinkage with 0 gene (right). The minimum overall error rate is 0.17 with the biggest threshold value, 4.56, which selects 0 genes in the classification model. Figure 3b shows the average error rate decrease first, then reach the minimum average error rate and finally increase to a certain level from 0 shrinkage value with 27252 genes (left) to 2.04 shrinkage value with 191 genes and end at 4.56 shrinkage with 0 genes. The minimum average error rate is 0.26 with the optimal threshold value 2.04. Figure 3c and 3d show that the average error rate with PAM default setting for feature selection and the average error rate and overall error rate with our approach selected the same shrinkage value, 2.04. The 191 genes and relative information show in the appendix.



(a) Overall Error Rate

(b) Average Error Rate

(c) Average Error Rate with Optimal Cut

(d) Overall Error Rate with Optimal Cut

Figure 3: Plots of Error Rates against by the Threshold Value

The nested LOOCV was used to evaluate the selected model's performance for the classification problems between the IF and CR groups through the parallel computing techniques. The PAM default methods with overall error rate criteria assigns all the IF patients to the CR groups in order to get the lowest overall error rate. However, the classification model with 0 genes has no power on classification and prediction. The sensitivity of the model is 0 and the p-value of the fisher exact test is equal to 1, which suggests the original PAM method predict new samples into the majority class under the unbalanced data.

The model determined from our three approaches is same and perform similar to each other in the nested LOOCV results. From the following table, we can see the three approaches successfully reduce the effect unbalance data, have the better performance in the LOOCV error rate evaluation and can effectively remove the non-informative genes by the 2.04 shrinkage.

Table 5: Prediction Results from LOOCV of IF VS. CR

Method	Overall.err	Ave.err	Sensitivity	Specificity	P-value
PAM.overall	0.17	0.50	0	1	1
PAM.average	0.26	0.2736	0.7059	0.7470	0.00109
Ours.overall	0.24	0.3317	0.5294	0.8072	0.01077
Ours.average	0.24	0.3317	0.5294	0.8072	0.01077

4.1.2 Performance Evaluation Using an Independent Data Set

It is easy to be overfitting when we use the data, which is also used for model development,to evaluate the model performance. In order to have the unbiased performance evaluation, we perform prediction on the test set, which contains the cohort 9404, 0434_1 and 0434_2. The average error rate with the PAM default settings predicts on the test set based on the original cut-off, 0.5, to classify the new samples by comparing with the posterior probability. If the sample's posterior probability is bigger than 0.5, it will be classified into the IF class. Otherwise, it will be classified into the CR group. Since we treated the ratio of class prior probabilities as the parameter determined by the cross-validation, we may consider the decision rule with different cut-offs by comparing with the posterior probability of each sample.

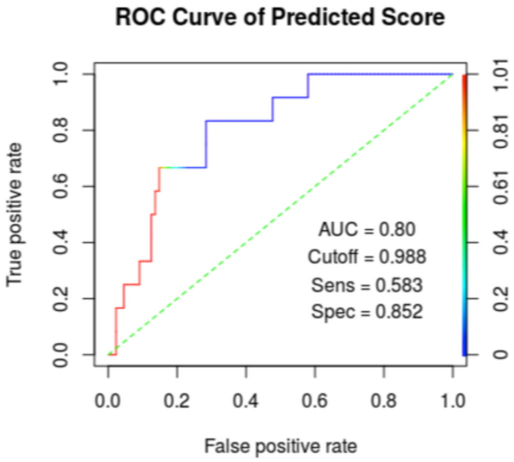


Figure 4: ROC Curve of Prediction on Total Test Set

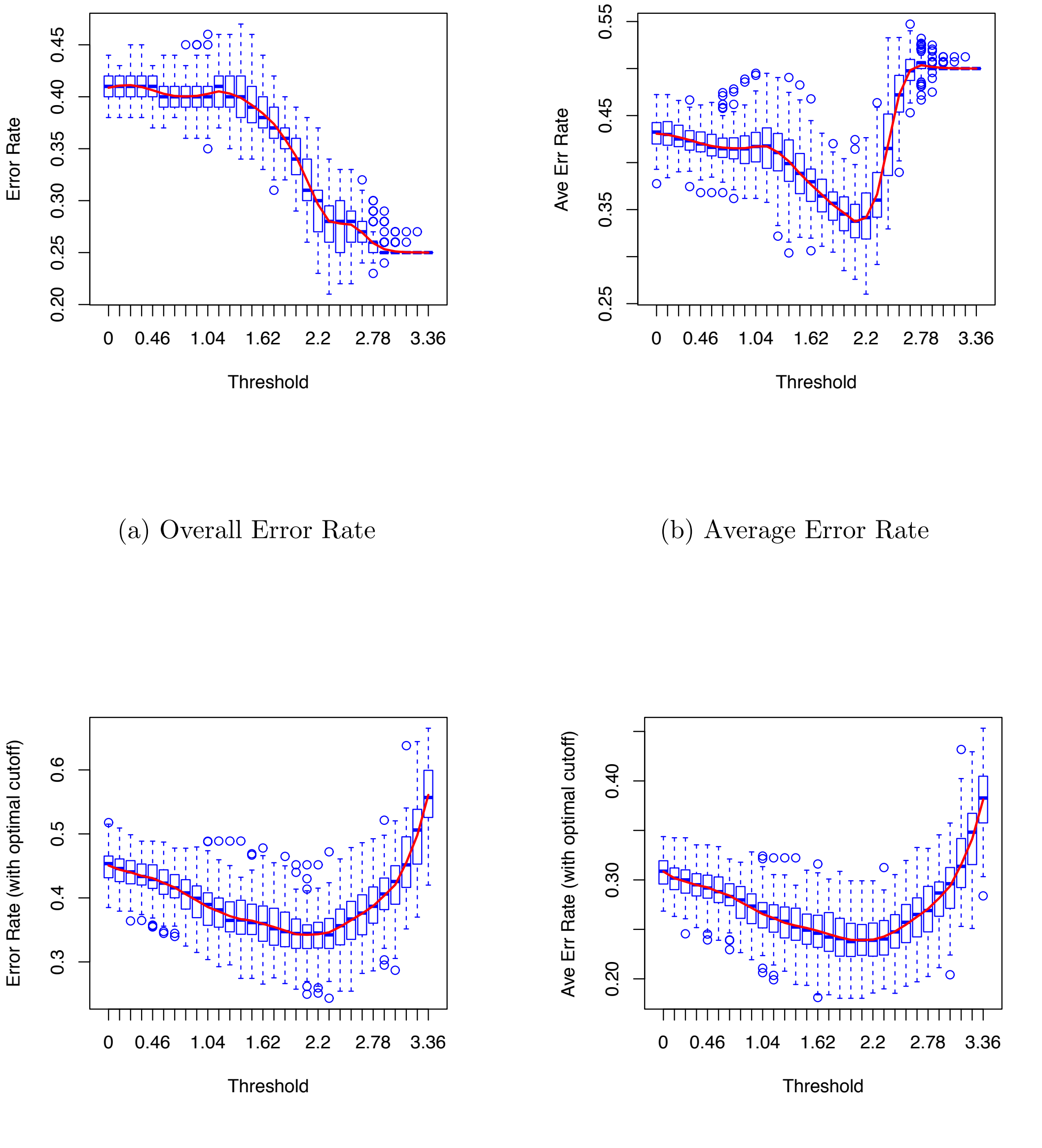
The ROC curve constructed from the clinical prediction shows the good performance in risk classification and outcome prediction on the external independent test set in Figure 4. The 0.80 AUC value suggests the high predictive accuracy of the classification model. The cut-off of the ROC curve is 0.988, which suggest under this cut-off, we have the highest sensitivity and highest specificity in the prediction results. The prediction with this cut-off has the overall error rate 0.18, the average error 0.2822 and the sensitivity is a little bit lower, which is 0.5833 and the sensitivity is 0.8523 (Table 6). Showing with statistical significance, the p-value of the fisher exact test is 0.00198, which suggest that the two classes' prediction results is not independent.

Table 6: Prediction Results from Independent data set of IF VS. CR

Predictions	Overall error	Average error	Sensitivity	Specificity	P-value
Original prediction	0.18	0.2462	0.6667	0.8409	0.00047
Prediction with cut point	0.18	0.2822	0.5833	0.8523	0.00198

4.2 Task 2: High-risk VS. Low-risk

4.2.1 Threshold Selection and Performance Evaluation by Nested LOOCV



(a) Overall Error Rate

(b) Average Error Rate

(c) Average Error Rate with Optimal Cut

(d) Overall Error Rate with Optimal Cut

Figure 5: Plots of Error Rates against by the Threshold Value

Figure 5 also shows the training, cross-validation, and test errors for different amount of the shrunken value Δ. From the nested LOOCV results, we can find the PAM default settings still select the biggest threshold value (3.36) in Figure 5a and contain 0 genes in its model. The sensitivity and the specificity is 0 and 1. Our three approaches select the 2.085 threshold value and contain 118 genes in the classification model with relative high sensitivity (Table 7).The p-values of fisher test from our approaches are statistical significant. These results can suggest that our modified approaches can improve the PAM in risk classification and outcome prediction under the unbalanced data.

Table 7: Prediction Results from LOOCV of High Risk vs. Low Risk

Method	Overall.err	Ave.err	Sensitivity	Specificity	P-value
PAM.overall	0.25	0.50	0	1	1
PAM.average	0.31	0.3267	0.64	0.7067	0.0037
Ours.overall	0.36	0.3467	0.68	0.6267	0.0105
Ours.average	0.36	0.3467	0.68	0.6267	0.0105

4.2.2 Performance Evaluation Using an Independent Data Set

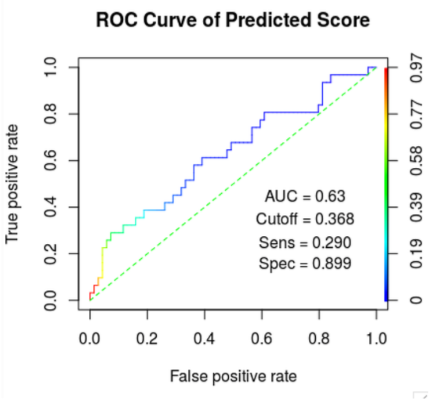


Figure 6: ROC Curve of Prediction on Total Test Set

Although the unbiased prediction result from external data set validation in the high risk VS. low risk task is not very good, our approaches still improve the performance from the PAM default settings. The sensitivity is 0.2258 and the specificity is 0.9420 for the original cut-off value (0.5). For the optimal cut-off, the average error rate is relative high in table 5, which is 0.71. It also came with the relative low sensitivity and high specificity. All the p-value shows statistical significance for the prediction results at the 95% confidence level.

Table 8: Prediction Results from External validation data set of High Risk vs. Low Risk

Predictions	Overall error	Average error	Sensitivity	Specificity	P-value
Original prediction	0.28	0.42	0.2258	0.9420	0.0322
Prediction with cut point	0.29	0.71	0.2903	0.8986	0.0354

5 Discussion

The three modified approaches were successful in improving the performance of the risk classification and outcome prediction by applying PAM on the unbalanced data set. Through the nested LOOCV evaluation, those approaches can effectively and similarly find a set of 191 genes for the early response task to classify the IF and CR groups with a low overall error rate around 0.24 and a set of 118 genes in the long term task to address the long term task to classify the high risk and the low risk groups with a relative low error around 0.35. The selection of the amount of shrinkage is very important in this method. The error rate changing as the shrinkage value change and the number of informative genes concluded in the classification is determined by the threshold value. One part of our modification on the shrinkage value selection is more intuitive and effective.

The evaluation results of original PAM method from the nested LOOCV show that the original PAM is not adaptable to the unbalanced data classification problems. It has the tendency to predict new sample into the majority class in order to minimize the overall error rate in both two tasks. Our modified PAM methods can somehow reduce the effects of the majority class in the unbalanced data set by simply giving the minority class more weights when doing the error evaluation and threshold value selection, which is more reasonable and meaningful.

According to the unbiased prediction performance evaluation on the external independent data set, the IF and CR groups in the early response task can be more correctly classified and the high risk and low risk groups in the long term task cannot be classified with a low error rate. It is easier to predict the induction failure and complete remission through the gene expression profile in the early response. Since some other factors can contribute to or affect the classification model by changing the gene expression value, it is still a problem for correctly prediction in the long term with a low error rate. In the future, we may continue to work on the improving the model performance of risk classification and outcome prediction by some other complex model.

References

[1] Howlader N, Noone AM, Krapcho M, et al, eds.: SEER Cancer Statistics Review, 1975-2010. Bethesda, MD: National Cancer Center Institute, 2013, Sections 28, 29.

[2] Larson, R. A., Dodge, R. K., Burns, C. P., Lee, E. J., Stone, R. M., Schulman, P., ... & Schiffer, C. A. (1995). A five-drug remission induction regimen with intensive consolidation for adults with acute lymphoblastic leukemia: cancer and leukemia group B study 8811. *Blood*, 85(8), 2025-2037.

[3] Rowe JM, Buck G, Burnett AK, et al. Induction therapy for adults with acute lymphoblastic leukemia: results of more than 1500 patients from the international ALL trial: MRC UKALL XII/ECOG E2993. *Blood.* Dec 1 2005;106(12):3760-7. [Medline].

[4] Thomas X, Boiron JM, Huguet F, et al. Outcome of treatment in adults with acute lymphoblastic leukemia: analysis of the LALA-94 trial. *J Clin Oncol.* Oct 15 2004;22(20):4075-86. [Medline].

[5] Cortes J, O'Brien SM, Pierce S, et al. The value of high-dose systemic chemotherapy and intrathecal therapy for central nervous system prophylaxis in different risk groups of adult acute lymphoblastic leukemia. *Blood.* Sep 15 1995;86(6):2091-7. [Medline].

[6] Kantarjian H, Thomas D, O'Brien S, et al. Long-term follow-up results of hyperfractionated cyclophosphamide, vincristine, doxorubicin, and dexamethasone (Hyper-CVAD), a dose-intensive regimen, in adult acute lymphocytic leukemia. *Cancer.* Dec 15, 2004;101(12):2788-801.

[7] Kang, H., Chen, I. M., Wilson, C. S., Bedrick, E. J., Harvey, R. C., Atlas, S. R., ... & Willman, C. L. (2010). Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric B-precursor acute lymphoblastic leukemia. *Blood*, 115(7), 1394-1405.

[8] Gohlmann, H., & Talloen, W. (2009). *Gene expression studies using Affymetrix microarrays.* CRC Press.

[9] Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10), 6567-6572.

[10] Lusa, L. (2013). Improved shrunken centroid classifiers for high-dimensional class-imbalanced data. *BMC bioinformatics*, 14(1), 64.

[11] Wang, S., & Zhu, J. (2007). Improved centroids estimation for the nearest shrunken centroid classifier. *Bioinformatics*, 23(8), 972-979.

[12] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. Knowledge and Data Engineering. *IEEE Transactions on*, 21(9), 1263-1284.

[13] Hanley, J. A., & McNeil, B. J. (1982). The averageing and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29-36.

[14] Green, D. M. (6). Swets, JA (1966). Signal detection theory and psychophysics. *Psychological Bulletin*, 75, 424-429.

[15] Cawley, G. C. (2006, July). Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. *In Neural Networks, 2006. IJCNN'06. International Joint Conference on* (pp. 1661-1668). IEEE.

[16] Simon, R. (2006). Development and evaluation of therapeutically relevant predictive classifiers using gene expression profiling. *Journal of the National Cancer Institute*, 98(17), 1169-1171.