University of California

Los Angeles

Prostate Cancer Classification Based on Gene Expression and Splicing Profiles

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Meng Meng

2016

Abstract of the Thesis

Prostate Cancer Classification Based on Gene Expression and Splicing Profiles

by

Meng Meng

Master of Science in Statistics

University of California, Los Angeles, 2016

Professor Yingnian Wu, Chair

The purpose of this study was to propose a method for classifying prostate cells into specific diagnostic categories based on their gene expression and exon inclusion level and compare their performance in classification. In order to build a concise statistical model with meaningful biological information, we combining univariate analysis with multivariate analysis with LASSO regularization for variable selection. Missing data is an important problem for exon inclusion level in our data. We apply two imputation methods and compare their results. Our questions in concern were answered by error rates of 100 iterations of cross-validation in testing after training. We found: (1) Exon inclusion level has a much stronger prediction ability than gene expression on our data by making lower error rates (p-value=1.29e-11 for exon inclusion level imputed by median and 2.20e-16 for exon inclusion level imputed by KNN); (2) The model built on exon inclusion level is more concise with less variables than that built on gene expression (p-value=8.15e-6); (3) Imputation methods on exon inclusion level does not affect classification results (p-value=5.37e-1).

The thesis of Meng Meng is approved.

Hongquan Xu

Yi Xing

Yingnian Wu, Committee Chair

University of California, Los Angeles
2016

To my mother . . .

who gives me unconditional love at all times

Table of Contents

List of Figures

Acknowledgments

Thanks for the support from all my family members, friends, faculty members, and staff at UCLA, I had the most memorable three years in the U.S. Without these help, I cannot grow and be the person I am now.

I would like to show my greatest gratitude to my advisor, Prof. Ying Nian Wu, who helped me continue my path in academia and taught me patience, the most precious quality in doing research and teaching. I would like to thank my co-adviser Prof. Yi Xing, who gave me this inspiring and interesting topic to finish this degree. I would like to thank Dr. Shihao Shen, who guided me all the way through this project and gave me profound ideas and open the door for me to the new area of bioinfomatics. I would like to thank Prof. Hongquan Xu, whose wisdom and insight gave me the positive energy to face all the obstacles.

Finally I would like to thank Dr. Akram Almohalwas and Dr. Juana Sanchez for giving me the opportunity to work as teaching assistant and showed me how much responsibility and passion it took to be a good instructor. I would like to thank Safaa Dabagh and Linda Zanontian, who cared me like sister. I would like to thank Glenda Jones who treated me as daughter in this new home in Statistics. And many thanks to all the staff members and Ph.D students in the Statistics department.

CHAPTER 1

Introduction

Prostate cancer remains the most frequently diagnosed cancer world wide and is the second leading cancer of men in the United States. According to recent estimates, it accounts for 33% of cancer diagnosed and 6% of cancer death of men worldwide[1][2]. Besides the high death rates, prostate cancer is also an indolent disease and hidden for years for patients. Current prognostic indicators including clinical stage, bioppsy Gleason grade (a measure of tumor differentiation), and serum PSA levels are not accurate in clinical prediction. The clinical heterogeneity of prostate cancer indicates the molecular heterogeneity among tumors. Therefore, it is urgent to find good biomarkers for early clinical diagnosis and treatment.

RNA-Seq has recently become an attractive method of choice in the studies of transcriptomes. It uses deep-sequencing technologies and provides more precise measurement of levels of transcripts and their isoforms than other methods, such as microarray. These methods allow the simultaneous monitoring of expression levels and splicing profiles of thousands of genes[5][10], and therefore propels the computational analysis using machine learning techniques. These analysis extract patterns and build classification models from gene expression and alternative splicing data and aid the prediction[8][7][3] and prognosis[12][4] of cancer.

Gene expression reveal the overall picture of change, but does not address the complex events that occur within individual genes in a given sample. As many as 80% genes undergo a process called alternative splicing (AS) which generates multiple mRNA isoforms and contributes the proteomic diversity in higher eukaryotes[13][9]. One of the big advantage of RNA-Seq over microarray is to accurately measure AS. The occurrence of many diseases, including cancer, comes along with increases of AS that contribute to their pathogenesis. Mutations can also affect AS by changing gene sequences that control the binding of slicing factors or in splice enhancer or inhibitor sequences. It has been estimated that 15% of point mutations that cause human genetic disease affect splicing[6]. We therefore incorporate features from AS by including exon inclusion levels, which is the percentage of gene isoforms that include the exon in our classification. Our results show exon inclusion level outperform gene expression in accuracy of classification and generate a more concise model with less variables than gene expression with higher performance in classification.

This paper is arranged as following: Chapter 2 introduces statistical methods for variable selection including univariate logistic regression and multivariate selection using LASSO regularization, which will be used for variable selection and classification for cancer data. In chapter 3, we introduce our data set, which includes gene expression and exon inclusion level, and methods for missing data inputation. Then a two-round feature selection method is presented for variable selection and followed by cross-validation. Finally we present our model for prostate cancer classification with variables selected from gene expression and exon inclusion level. Chapter 4 provides discussion and conclusion.

CHAPTER 2

Statistical Methods

2.1 Logistic Regression

Logistic regression is widely used to model independent binary response data in medical and epidemiologic studies. The crucial limitation of linear regression is that it cannot deal with dependent variables that are binary. Logistic regression uses link function to transform the continuous output from the linear predictor to fall between 0 and 1. The model for logistic regression is

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i + ... + \beta_p x_p \tag{2.1}$$

where

$$\text{logit}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i} \tag{2.2}$$

The right hand side of equation 2.2 is called log-odds of π_i. Odds is the ratio of the probability to its complement. As π_i goes down to zero the odds approach zero and the logit approaches $-\infty$. At the other extreme, as π_i approaches one the odds approach $+\infty$ and so does the logit. Thus, logits map probabilities from (0,1) to the entire real line.

The main idea behind bivariate logistic regression is the distribution of binomial variable Y_i with parameter π_i and n_i, which can be written as

$$Y_i \sim B(n_i, \pi_i) \tag{2.3}$$

where y_i is the realization of Y_i that take the value 0,1,...,n_i. If the n_i observations in group i are independent and have the same probability π_i of having the attribute of interest, then the probability distribution function of Y_i is given by

$$Pr\{Y_i = y_i\} = \binom{n_i}{y_i} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{2.4}$$

for $y_i = 0, 1, ..., n_i$. Here $\pi_i^{y_i}(1-\pi_i)^{1-y_i}$ is the probability of obtaining y_i successes and $n_i - y_i$ failures after n_i trials of Bernoulli experiment with probability π_i in some specific order, and the combinatorial coefficient is the number of ways of obtaining y_i successes in n_i trials.

For example, in cancer diagnosis, the number of patients getting cancer can be treated as a binomial variable, π_i will be the probability that patient has cancer, and y_i will be the dichotomous outcome with 1 as cancer, and 0 as no cancer. Features used for diagnosis, such as gene expression, will be treated as variables in the regression model.

2.2 LASSO

The name LASSO is an acronym for Least Absolute Selection and Shrinkage Operator. Given a linear regression with predictors x_i and response value y_i for $i = 1, 2, ..., N$, the LASSO solves the L1 penalized regression problem by finding $\beta = \{\beta_j\}$, $j = 1, 2, 3, ..., p$ that minimize the sum of a loss and a penalty

$$\hat{\beta}^{LASSO} = \min_{\beta \in R^p} \|y - X\beta\|^2 + \lambda\|\beta\|_1 \tag{2.5}$$

where λ is the complexity parameter which controls the amount of shrinkage. There is no penalty if λ equals to zero, and big penalty if λ goes larger.

The idea behind LASSO is to minimize the sum of squares with a constraint $\sum \beta_j < s$. Another popular shrinkage method is called ridge regression, which uses L2 penalty instead of L1 penalty for coefficient shrinkage.

$$\hat{\beta}^{Ridge} = \min_{\beta \in R^p} \|y - X\beta\|^2 + \lambda\|\beta\|_2^2 \tag{2.6}$$

The problem looks similar, but their solution behave very differently. Compared with usual regression, which solves the unconstrained least squares problem, Lasso and ridge regression estimate constrain the coefficient vector to lie in some geometric shape centered around the origin. However, when $\lambda = 0$, $\hat{\beta}^{LASSO} = 0$, LASSO is able to perform variable selection by shrinking coefficients towards exactly 0. This makes LASSO substantially different from ridge regression.

CHAPTER 3

Classification

3.1 Introduction to Data

This data consists of 30 observations of gene expression and exon inclusion levels from normal cells in normal prostate, and 15 observations of gene expression and exon inclusion levels from normal cells adjacent to prostate tumors. Each observation contains measures of 57150 gene expressions, and 56612 exon inclusion levels.

Gene expression is measure by RNA-Seq which provides more accurate measurement of levels of transcripts and their isoforms. In a standard workflow of an RNA-Seq experiment, RNAs of interest in the sample are firstly fragmented and reverse-transcribed into complementary DNAs (cDNAs). Then the obtained cDNAs are then amplified and subjected to Next Generation Sequencing (NGS). Finally the reads generated are mapped to a reference genome and the number of reads aligned gives the measure of gene expression levels. The value of each gene expression is a non-negative number.

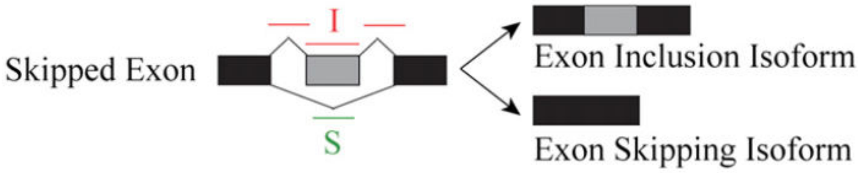

Figure 3.1: Diagram for alternative splicing.

Inclusion level of an exon is estimated as

$$\hat{\Psi} = \frac{I/l_I}{I/l_I + S/l_S} \tag{3.1}$$

where I is the counts of reads of inclusion isoforms that connecting the upstream splice junction, the alternative exon itself and the downstream splice junction, and S is the counts of reads of skipping isoforms that directly connects the upstream exon to the downstream exon. Refer to 3.1 for illustration. l_I and l_S are the effective lengths of the inclusion isoform and skipping isoform. The counts of reads of inclusion isoforms I can be assumed following a binomial distribution[11], written as

$$I|\Psi \sim Binomial(n = I + S, p = f(\Psi) = \frac{l_I\Psi}{l_I\Psi + l_S(1-\Psi)}) \tag{3.2}$$

Under the binomial assumption, the inclusion level Ψ is affected by the number of reads n and the proportion of reads of the exon inclusion isoforms from total reads after length normalization.

3.2 Univariate Analysis

3.2.1 Univariate variable selection

Univariable analysis tests the association of one explanatory variable with the dependent variable at a time without worrying about other variables. This is essential in order to shortlist variables for multivariable analysis in case of a large number of explanatory variables. It helps to exclude the variables that do not show any significant association with the outcome from further analysis.

We feed each gene and exon for all observations into logistic regression to perform univariate analysis. In total, 57150 (number of genes)+56612 (number of exons) times of univariate logistic regression are executed. For each run of logistic regression, we record the p-value as the measure of the significance of the variable (gene/exon) in predicting the response. P-value is an important measurement in hypothesis testing. It is the probability of obtaining the value of the test statistic at least as extreme as obtained, given the null hypothesis is true. Here our null hypothesis is the coefficient of the variable (gene/exon) in logistic regression is zero. A low p-value means that there is little chance that the null hypothesis is true, which mean there is a statistically significant association between the tested variable and the response.

3.2.2 Normalization for gene expression

As we mentioned earlier, the data value for gene expression is a non-negative number, while for exon, the inclusion level falls in [0,1]. In order to compare gene and exon on the same level and more importantly combine them together for future multivariate analysis, we need to normalize gene expression by scaling its values in between 0 and 1. We use min-max normalization to perform linear transformation on the original data. Suppose x_{min} and x_{max} are the minimum and maximum of variable x, the normalized value for x_i is calculated by

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Min-max normalization preserves the relationship among the original values. We normalize the top 100 variables selected by univariate analysis in this way.

3.2.3 Missing data imputation for exon inclusion level

Missing data sometimes is unavoidable in data collection. About 38% of the 56612 exons have at least one of 45 observations missing in our exon inclusion data. The missingness is due to miss-detection of certain exon in the inclusion isoform and the skipping form. After univariate analysis, the missing rate among the selected 100 exons decreases to 10%, which indicates most of the exons with missing values are not statistically significant in predicting cancer. We therefore focus on missing data imputation on the selected exons.

Before we jump into methods for missing data imputation, an important task is to know the reason for missing. There are generally three data missing mechanism:

- Missing Completely At Random (MCAR). Here missing completely at random means the missing of a variable is independent of itself and other variables. This is an ideal situation and rare in real data. A common approach is removing the cases with missing values which won't bias the inference.

- Missing at random (MAR). Most missingness is not completely at random. The missing of the variable is dependent on other variables but itself.

- Not Missing At Random (NMAR). The missing value depends on the variable that is missing. Data that's NMAR can be difficult to analyze properly. Currently approaches to analyzing NMAR data include the use of selection models and pattern mixtures.

The missing type of exon inclusion level is Missing At Random (MAR), because the missingness is due to the low abundance of the gene transcript, which is not related to the exon inclusion level. There are 90% of exons with complete data for all observations. Among the remaining exons, 90% of them have missing values less than 50%. We apply two imputation methods: imputation with median, and imputation with KNN on exon data. The impact of the methods are evaluated in later classification.

The reason we choose median is the distributions of exons are skewed, therefore, median would be a good indication for missing values. In this imputation method, missing value is replaced with median of the variable, where median is the middle point of the variable.

The idea for KNN missing data imputation is quite similar with KNN classification, in which the class label is determined by k nearest neighbors. The algorithm for KNN imputation is:

- Compute the Euclidean distance between variable with missingness and all other complete variables. Find the k-closest variables.
- Replace missingness with the average of the corresponding entries of k-closest variables.

This method is quite simple in principle but is effective and often preferred over some of the more sophisticated methods. The disadvantage is that it does not model the variation of the imputed value and therefore the uncertainty of the imputed value is underestimated.

3.3 Multivariate Selection

We choose top 100, 300, and 500 variables from both gene and exon by sorting their p-value from small to large in univariate analysis. For the top 100 variables, the cut-off p-value is 1.20e-3 for gene, and 5.50e-3 for exon. For the top 300 variables, it is 3.29e-3 for gene, and 1.43e-2 for exon. For the top 500 variables, it is 4.36e-3 for gene, and 2.36e-2 for exon. We further run 100 iteration of cross-validation using top 100, 300, and 500 variables for both gene expression and exon inclusion level, and find similar error rates among each of the three groups (For gene, mean=0.261 for top 100 variables, mean=0.304 for top 300 variables, and mean=0.22 for top 500 variables; for exon, mean=0.087 for top 100, 200 and 500 variables). Therefore, these models lead to similar outcome. For simplicity of the model, we select the top 100 variables as the input for the multivariate analysis.

We refer to the 100 variables of exons selected by univariate analysis with median imputation and KNN imputation as $Exon_{median}$ and $Exon_{KNN}$, respectively, and the 100 variables of genes selected by univariate analysis with normalization as $Gene_{norm}$. There are 45 observations for each of $Exon_{median}$, $Exon_{KNN}$, and

$Gene_{norm}$. The response variable is 1 or 0 depending on the whether the patient has cancer or not. Among the 45 observations, 30 of them are from normal prostate cells and are labeled 0 in response, and the remaining 15 observations are from cells adjacent to prostate tumor and are labeled 1 in response.

We perform 100 iterations of training and testing on $Exon_{median}$, $Exon_{KNN}$ and $Gene_{norm}$, as well as the combination of $Gene_{norm}$ and each of imputed exon data. In each iteration, half of the data are randomly chosen as training set and the remaining as testing set. Then we use the "glmnet" package in R to implement the training process through cross-validation. During training, "glmnet" fits the regularization path on 100 different values of λ. The decision of which λ is the best is a tradeoff between low error rates and amount of shrinkage and up to the user. Here we chose λ that minimize the cross-validation error. We calculate the error rates for normal prostate cells , cells adjacent to prostate tumors, and both cells. We use group1, group2 and group1+group2 as notations in the graph respectively. Error rate is measured by the number of miss-classification observations divided by the total number of observations being classified.

Fig.3.2 is the boxplot summary of 100 iterations of classification error rates for $Gene_{norm}$ and exon imputed by median $Exon_{median}$, as well as combination of $Gene_{norm}$ and $Exon_{median}$. The three columns show error rates among three cell groups using exon, gene, and exon+gene as classification variable, respectively. We run t-test to test the statistical significance of the difference among the three columns. We get p-value=1.29e-11 in test between the first column and the second column, and p-value=8.16e-3 between the second and the third column. Both p-values are smaller than 0.05, which indicates exon is much better than gene in predicting prostate cancer in terms of classification error.

Fig.3.3 show the summary of 100 iterations of classification error rates for $Gene_{norm}$, exon imputed by KNN $Exon_{KNN}$, and combination of $Gene_{norm}$ and $Exon_{KNN}$. The configuration is the same with Fig.3.2 except we use $Exon_{KNN}$

instead of $Exon_{median}$ in classification. In order to compare the prediction ability of gene and exon imputed by KNN, we perform the similar two t-tests as before, and get p-value= 2.20e-16 for the test between exon and gene and p-value=3.00e-16 between exon+gene and gene for all the observations, which again verifies variables selected from exon is better predictor than those selected from gene. The bench effect in gene expression may affect the prediction accuracy. Because exon inclusion levels are calculated from the isoform counts of the same individuals, exon inlucsion level calculation can remove bench effect in the transcript expression.

The above two experiments both indicate variables selected from exon inclusion level have much better prediction than those selected from gene expression. We are further curious about whether the method for imputation affects the prediction ability of exon inclusion level. The advantage of KNN imputation is that it takes into consideration the correlation structure of the data and easily treat instance with multiple missing values without creating a predictive model for each attribute with missing data. While median imputation is robust to outliers due to the nature of median, especially in our case the variables have skewed distribution. We perform a two sample t-test between the error rates of $Exon_{median}$ and $Exon_{KNN}$ over all observations, and get p-value= 5.37e-1, which shows no statistical significance of the difference. Another t-test is performed between the combination of each imputed exon inclusion level with gene expression, namely $Exon_{median}$+$Gene_{norm}$ vs. $Exon_{KNN}$+$Gene_{norm}$, we get p-value=8.86e-4, which shows a significance of the difference. It is interesting that when combined with gene expression, exon inclusion level with KNN imputation shows a stronger prediction ability than with median imputation. It is possible that variables in gene expression and exon inclusion level imputed by KNN are more correlated with each other and LASSO punishes correlation and shrinks the correlated ones. In the future analysis and our prediction model, we will use exon inclusion level imputed by KNN as exon data.

We further compare the number of non-zero coefficients selected by LASSO for exon, gene, and the combination of exon and gene in the 100 iterations of training. Fig.3.4 shows the distribution for the number of variables selected from gene, exon, and combination of gene and exon after 100 iterations. ANOVA test gives p-value=3.14e-5, which shows the difference of the number of coefficients among the three groups is significant. We further compare the significance of difference between each of the two groups by performing two sample t-tests and get p-value=8.15e-06 for exon (6.2) vs. gene (7.8), p-value=7.01e-5 for gene (7.8) vs. exon (4.2) + gene(2.8), and p-value=9.24e-1 for exon (6.2) vs. exon (4.2) + gene (2.8). The number in the parenthesis is the average of number of variables selected. These p-values show that variables selected from exon are more effective in building concise model than those from genes.

3.4 Prediction Model

In this section, we build three classification models using all the observations from the data.

The top 100 variables selected from gene expression by univariate analysis is used to build the first model. Different from the training and testing, we use multivariate analysis with LASSO regularization to train a model using all the 45 observations of the 100 variables in one shot. 12 genes are selected using λ minimizing the classification error. Fig.3.5 visualizes the 12 genes with non-zero coefficients. Each row represents one gene, and each cell is colorized based on the level of gene expression, ranging from red for the lowest value through green for the highest value. Data was normalized with mean zero for better visualization. We can observe a dissimilarity of color between the first 20 columns which represents cells from normal prostate and the remaining 15 columns which represents cells

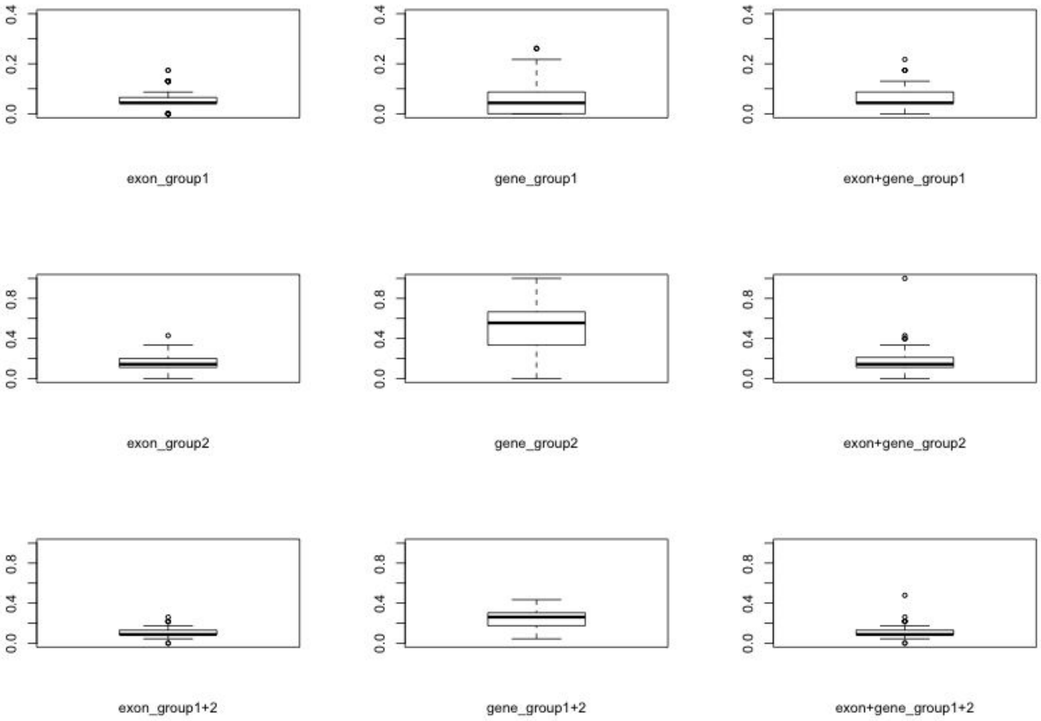

Figure 3.2: Cross validation and error rates of prostate cancer classification where missing values of exon inclusion level are imputed by median.

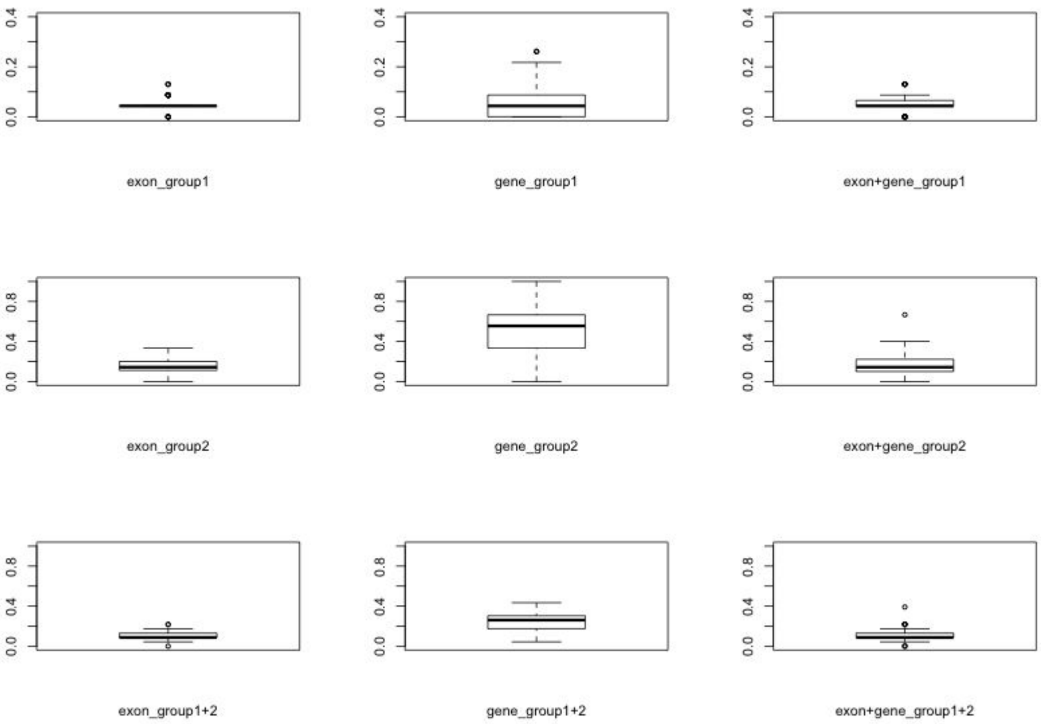

Figure 3.3: Error rates of prostate cancer classification where missing value of exon inclusion level is imputed by KNN.

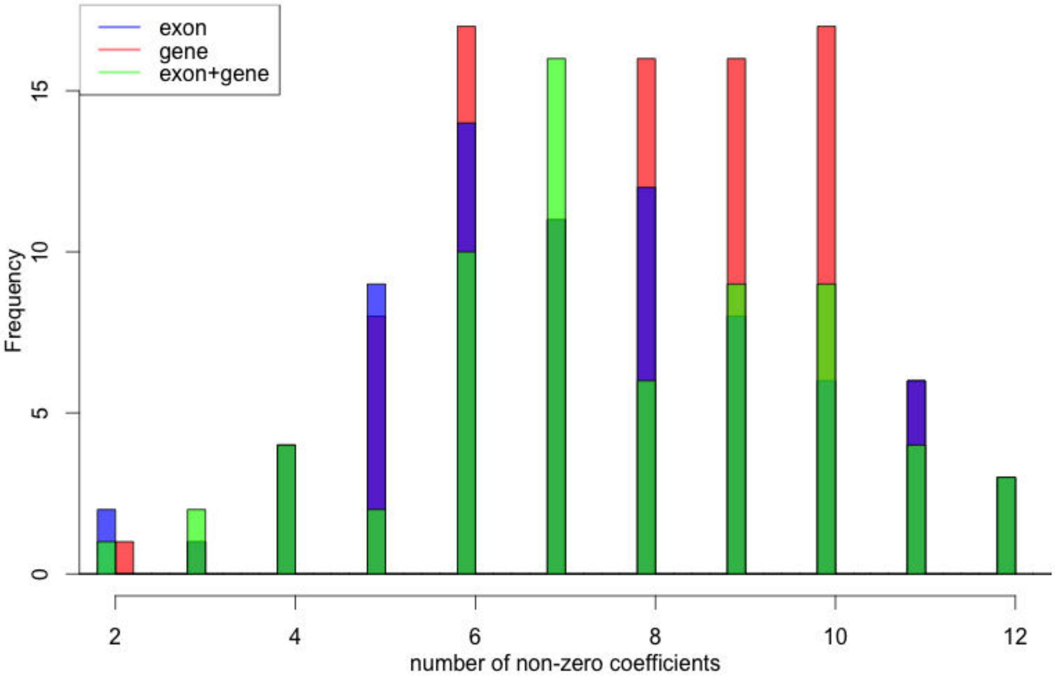

Figure 3.4: The distribution of the number of non-zero coefficients selected by multivariate analysis using LASSO regularization for exon, gene, and the combination of exon and gene in 100 iterations of training.

adjacent to prostate tumors.

The second model is based on exon inclusion levels, we use the top 100 variables selected from exon inclusion level by univariate analysis and imputed by KNN imputation. 11 out of 100 exons are selected in the prediction model. Fig.3.6 is the heatmap for the selected exons. Each row represents one exon, the first 30 columns are observations of cells from normal prostate, and the remaining 15 columns are observations of cells adjacent to prostate tumors. The color represents the level of exon inclusion level. The difference of color is significant between the normal group and the tumor group for each exon, which verifies that these variables are distinctive features for classification.

Finally we combine the data from the first two models to construct the third model. 100 variables of gene expression and 100 variables of exon inclusion level with 45 observations are used to train the model. 7 out of 100 variables are selected from gene and 10 out of 100 variables are selected from exon. Fig.3.7 shows the visualization of these variables. The first 10 rows represent exons, and the remaining 7 rows represent genes. The dissimilarity of color shows these features have distinctive values among the two classification groups, and are important for classification.

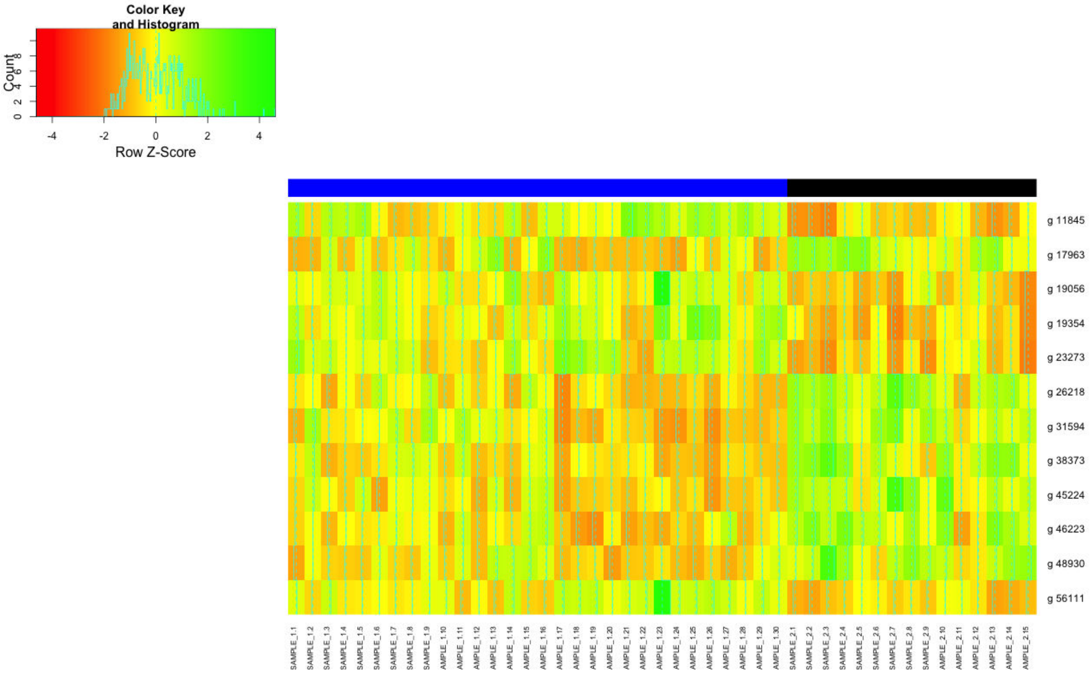

Figure 3.5: Heatmap of genes selected by model trained on 30 gene samples from normal cells and 15 gene samples from cells adjacent to prostate cancer. Each row represents one gene, and each cell is colorized based on the level of expression of that gene in the sample.

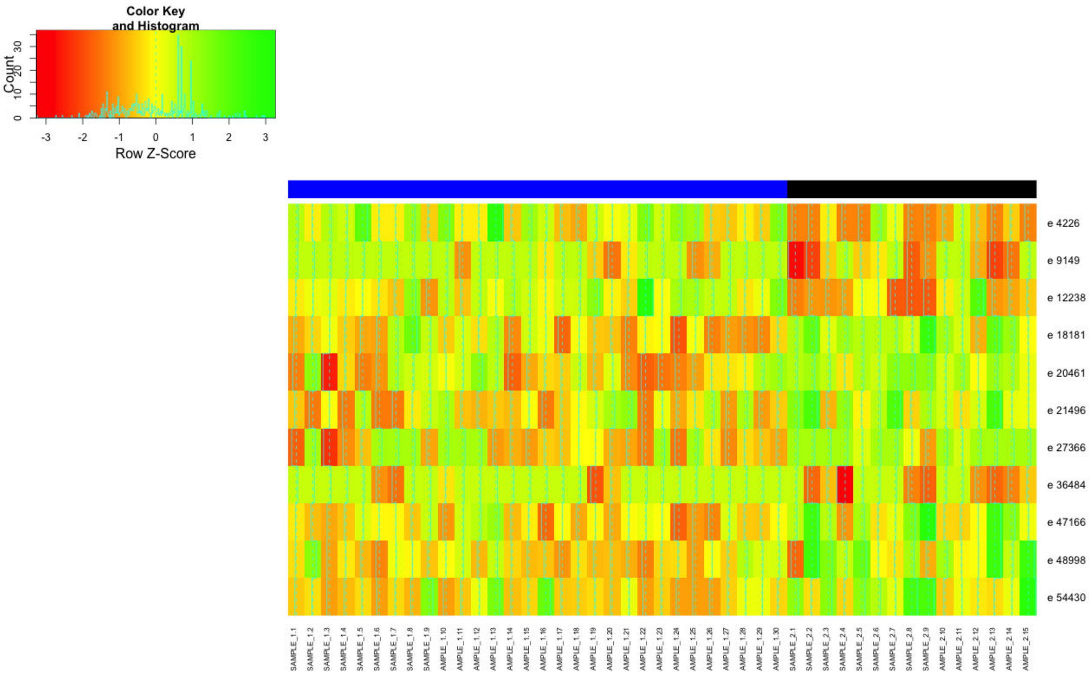

Figure 3.6: Heatmap of exons selected by model trained on 30 exon samples from normal cells and 15 exon samples from cells adjacent to prostate cancer. Each row represnets one exon, and each cell is colorized based on the level of inclusion level of that exon in the sample.

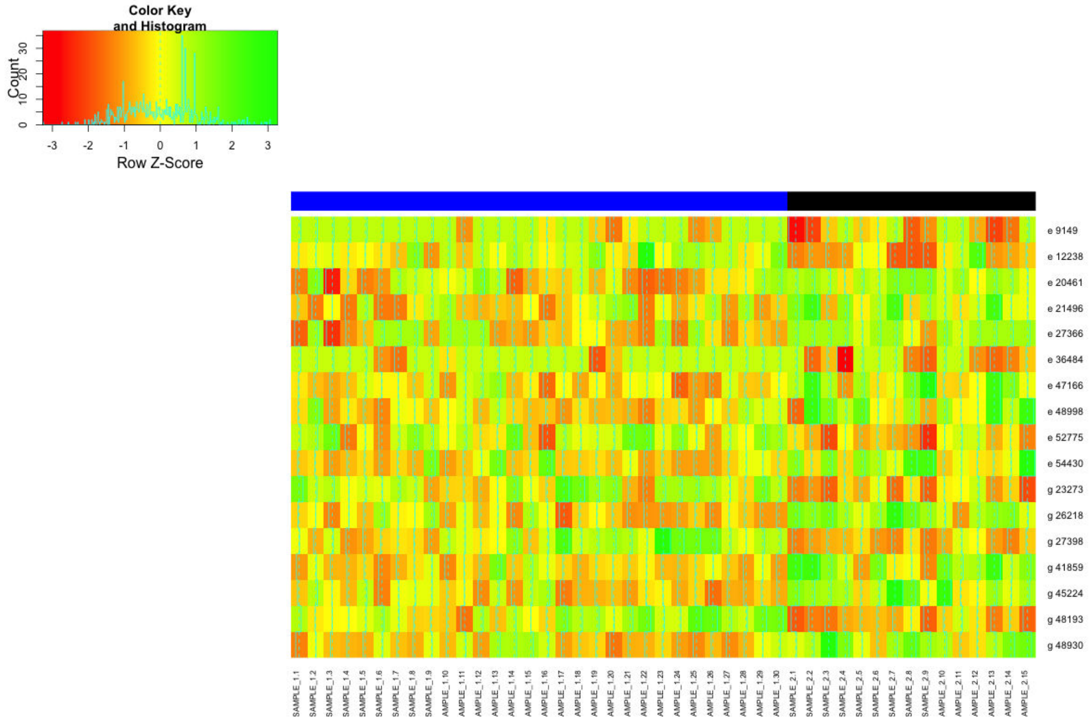

Figure 3.7: Heatmap of exons and genes selected by model trained on 30 exon and gene samples from normal cells and 15 exon and gene samples from cells adjacent to prostate cancer. Each row represents one exon or gene, and each cell is colorized based on the level of inclusion level of that exon or gene in the sample.

CHAPTER 4

Discussion and Conclusion

In this paper, a method for classifying prostate cells into specific diagnostic categories are proposed. We use gene expression and exon inclusion level as two types of features in the classification. We construct the classification model by running two rounds of variable selections. In the first round, univariate analysis using logistic regression is applied to select variables that are significant in prediction the categories. Also we found this step largely contributes to missing data imputation for exon inclusion level through dropping variables with most missing values. Top 100 variables from both gene and exon are selected as input for the second round of multivariate analysis with LASSO regularization. In this step, we apply two imputation methods to solve missing data problem in exon inclusion level and compare their performance. The results show no statistical significant difference between error rates of model built on exon inclusion level imputed by the two methods (p-value=5.37e-1), but KNN imputation shows a stronger prediction ability than median in the combined data by generating lower error rates. (p-value=8.86e-4). We choose KNN as imputation method for further analysis. We then compare the performance of variables selected from gene expression, exon inclusion level imputed by KNN, and the combination of gene with exon inclusion level imputed by KNN. ANOVA test on error rates after 100 iterations of cross-validation shows significant difference among the three models (p-value=3.14e-5). The average error rate is 18% for gene, 9% for exon, and 10% for combination of gene and exon. Further two group t-tests also show a significant difference of error rates for gene vs. exon (p-value=8.149e-06), gene vs. gene+exon (p-value=7.01e-5), but no significant difference for exon vs. gene+exon (p-value=9.24e-1). This indicates exon inclusion levels are better predictors than gene expression for classification of prostate tumor patients in our data. Also the model constructed using exon inclusion level is more concise with less variables than that built on gene expression.

For future work, first, we would like to apply our method on a larger data set, since our current one contains only 30 non-patients and 15 patients, also the imbalance number of the two groups may impair the generalization ability of our model. Second, it is interesting we find the two imputation methods we used in this paper do not affect the performance of exon inclusion level in classification, but more imputation methods need to be tested to draw conclusion about the influence of imputation. Third, we would like to add histone modification, DNA methylation or somatic mutation in the survival prediction model. Finally, when read counts are small, the exon inclusion level estimation is not as reliable as estimation from large read counts, therefore, modeling estimation uncertainty of exon inclusion level in the prediction model will be a good way to solve this problem.

References

[1] World cancer report 2014. World Health Organization. Chapter 5.11.

[2] Prostate cancer. National Cancer Institute, 2014.

[3] Sandrine Dudoit, Jane Fridlyand, and Terence P Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87, 2002.

[4] Bard Erin and Hu Wei. Identification of a 12-gene signature for lung cancer prognosis through machine learning. *Journal of Cancer Therapy*, 2011, 2011.

[5] Christina A Harrington, Carsten Rosenow, and Jacques Retief. Monitoring gene expression using dna microarrays. *Current opinion in Microbiology*, 3(3):285–291, 2000.

[6] Jason M Johnson, John Castle, Philip Garrett-Engele, Zhengyan Kan, Patrick M Loerch, Christopher D Armour, Ralph Santos, Eric E Schadt, Roland Stoughton, and Daniel D Shoemaker. Genome-wide survey of human alternative pre-mrna splicing with exon junction microarrays. *Science*, 302(5653):2141–2144, 2003.

[7] JingJing Liu, WenSheng Cai, and XueGuang Shao. Cancer classification based on microarray gene expression data using a principal component accumulation method. *Science China Chemistry*, 54(5):802–811, 2011.

[8] Qingzhong Liu, Andrew H Sung, Zhongxue Chen, Jianzhong Liu, Lei Chen, Mengyu Qiao, Zhaohui Wang, Xudong Huang, and Youping Deng. Gene selection and classification for cancer microarray data based on machine learning and similarity measures. *BMC genomics*, 12(Suppl 5):S1, 2011.

[9] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12):1413–1415, 2008.

[10] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[11] Shihao Shen, Juw Won Park, Zhi-xiang Lu, Lan Lin, Michael D Henry, Ying Nian Wu, Qing Zhou, and Yi Xing. rmats: Robust and flexible detection of differential alternative splicing from replicate rna-seq data. *Proceedings of the National Academy of Sciences*, 111(51):E5593–E5601, 2014.

[12] Leonardo Vanneschi, Antonella Farinaccio, Giancarlo Mauri, Mauro Antoniotti, Paolo Provero, and Mario Giacobini. A comparison of machine learning techniques for survival prediction in breast cancer. *BioData mining*, 4(1):12, 2011.

[13] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.