A Novel Single Genes Ensemble Decision Tree Classifier on Microarray Cancer Prediction

Hui Zhang[a], Ziyi Yang[a], Yanqiong Ren[a], Yong Liang[a*] and Junjie Taoa[a]
[a]Building C407,Macau University of Science and Technology Avenida Wai Long, Taipa, Macau
zhanghui.nwu@foxmail.com, yangziyi091100@163.com, yanqiongren@gmail.com,
yliang@must.edu.mo*,932967652@qq.com

Abstract

Cancer classification problem based on microarray data analysis is becoming more and more popular. In biology, many different genotypes can produce the same phenotype. Therefore, it is necessary to find out the pathogenic gene for cancer patients with same genotypes. In this paper, we propose a novel classification method for gene selection and cancer classification based on microarray data. We first select the significant genes through four popular filter methods, T-test, entropy, Chernoff bound and Wilcoxon test. And then we build ensemble classifiers based on the decision trees induced by single genes. We apply our approach to three publicly available lung cancer gene expression datasets and compare our single genes ensemble classifier with six standard methods including single gene classifier with the t-test, single gene classifier with the WMW, Support Vector Machine, sparse logistic regression with L_1 penalty, K-Nearest Neighbour and Random Forest. The single genes ensemble classifier provides classification accuracy comparable to or better than those obtained by existing methods in most cases. In comparison with other methods, our method can select the most significant pathogenic gene for cancer patients with same genotype, which is helpful for cancer patients to get better targeted therapy.

1 The Framework of Our Method

The framework of our method is shown in Figure 1. Our method combines statistical feature selection measure method and ensemble decision tree classifiers for cancer detection based on microarray data. We select four statistical feature selection measure methods, T-test, entropy, Chernoff bound and Wilcoxon test. These methods are more easily accepted by biologists and clinicians because they are easy to understand. Each statistical method can select significant genes which are related to a person whether have cancer. We merge all the significant genes from the four statistical methods and generate a weighting candidate gene set. Based on the weighting candidate gene set, each single decision tree classifier is constructed based on one candidate gene. Therefore, we can get a single gene ensemble classifier, and abbreviate it as SGEC.

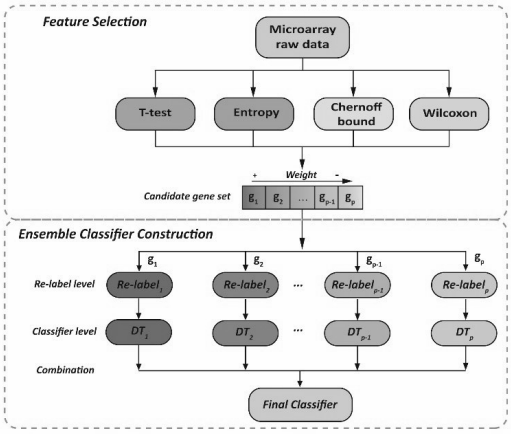

Figure 1: The framework of single gene ensemble classifier.

2 Results

We apply SGEC to three publicly available lung cancer gene expression datasets, and compare SGEC with other six standard methods including SGC-t, SGC-W, SVM, sparse logistic regression with L_1 penalty, K-NN and Random Forest. The experiment result shows in Table1.

The experiment result shows that simple classifier with a small number of features appear to work as well as some more complex multivariate classifiers. In microarray- based cancer prediction, we focus on the interpretability of the classier, so the simple classifier with a small number of features can perform well. SGC-t and SGC-W are the simplest classier which only select one significant gene to classify cancer samples, but might be influenced by noise gene, the classification performances are not ideal. For example, in GSE19188 dataset, SGC-W achieves the worst classification accuracy. Furthermore, there are different genotypes behind the same phenotype of human cancer disease in the view of biology. In other words, different cancer patients may be caused by different pathogenic genes. Our SGEC selects more than one

significant gene to classify cancer samples, and each selected significant gene has significant difference in the expression between two classes. SFTPC [Golub, 1999] and AGER [Pan, 2017] have been shown that are associated with lung cancer. FTPC and AGER are selected by SGEC but not discovered by SGC-W and logistic with L_1 penalty. Moreover, SGEC can identify pathogenic gene for each sample, it is meaningful because of the advantage for interpretability and applicability for biological study and medical use. Furthermore, SGEC provide classification accuracy comparable to or better than some more complex multivariate classifiers. For example, in GSE19804 dataset, our SGEC achieves the best classification accuracy, in other two datasets, SGEC provide classification accuracy comparable to the other six methods.

Table 1. The classification performance of seven different methods for three lung cancer datasets.

Accuracy (%)							
Dataset	**SGEC**	**SGC-t**	**SGC-W**	**SVM**	**Logistic + L_1**	**K-NN**	**RF**
GSE10072	97.60	93.75	78.13	**100.00**	**100.00**	93.75	98.85
GSE19188	95.43	86.96	52.17	**95.65**	87.17	84.78	87.68
GSE19804	**96.02**	88.89	88.89	94.44	94.44	66.67	92.78
Sensitivity (%)							
Dataset	**SGEC**	**SGC-t**	**SGC-W**	**SVM**	**Logistic + L_1**	**K-NN**	**RF**
GSE10072	96.12	94.12	**100.00**	**100.00**	**100.00**	**100.00**	97.96
GSE19188	**100.00**	95.65	**100.00**	**100.00**	86.52	**100.00**	84.95
GSE19804	92.85	93.75	**100.00**	94.44	94.44	87.50	91.46
Specificity (%)							
Dataset	**SGEC**	**SGC-t**	**SGC-W**	**SVM**	**Logistic + L_1**	**K-NN**	**RF**
GSE10072	**100.00**	93.33	68.18	**100.00**	**100.00**	88.24	**100.00**
GSE19188	90.14	78.26	46.34	90.48	88.30	73.08	**93.27**
GSE19804	**100.00**	85.00	81.82	94.44	94.44	60.71	94.25

3 Conclusions

Microarray technology has much facilitated the detection of cancerous molecular markers, abundant explorations have been conducted to carry out cancer diagnosis, prognosis or prediction based on DNA microarray data. Using gene expression data to conduct classification or prediction of cancer is often faced with the dilemma: genes far outnumber samples. Hence, the first difficulty in this topic is how to effectively identify the genes relating to the pathogenesis of specific cancers from the extremely high-dimensionality gene expression data. The second issue is the interpretability of predictive classifiers that biologists and clinicians care about.

To overcome the first difficulty, we employ four statistical feature selection methods to identify the significant genes relating to the pathogenesis of specific cancers. The four statistical feature selection methods are T-test, entropy, Chernoff bound and Wilcoxon test. These methods are more easily accepted by biologists and clinicians because they are easy to understand. Each statistical method can identify significant genes which are related to the pathogenesis of specific cancers. We merge all the significant genes from the four statistical methods and generate a weighting candidate gene set.

To handle the second trouble, we use the weighting candidate gene set to build ensemble classifier SGEC, SGEC is based on the decision trees induced by single gene. SGEC can identify the candidate significant genes for every cancer sample.

To verify the validity of SGEC, we apply SGEC to three publicly available lung cancer gene expression datasets, and compare the performance of SGEC with six standard methods including SGC-t, SGC-W, SVM, sparse logistic regression with L_1 penalty, K-NN and Random Forest. These six methods are widely used on microarray cancer detection. SGEC provides classification accuracy comparable to or better than those obtained by existing methods in most cases. In comparison with other methods, SGEC can find out the most significant pathogenic gene for each cancer patient, it is meaningful because of the advantage for interpretability and applicability for biological study and medical use.

Acknowledgments

This work was supported by FDCT Grant $No.003/2016/AFJ$ from the Macau Special Administrative Region of the Peoples Republic of China, the National Grand Fundamental Research 973 Program of China under Grant $No.2013CB329404$ and the China NSFC projects under Contracts $61373114, 61661166011, 11690011, 61721002$.

References

[Golub, 1999] Golub, T. e. a. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531537.

[Pan, 2017] Pan, Z. e. a. (2017). Long non-coding rna ager-1 functionally upregulates the innate immunity gene ager and approximates its anti-tumor effect in lung cancer. *Molecular carcinogenesis*.