Improvement of Survival Prediction from Gene Expression Profiles by Mining of Prior Knowledge

Siyuan Ren and Zoran Obradovic*

Center for Information Science and Technology, Temple University, Philadelphia, PA 19122, USA

** Corresponding author: Zoran Obradovic E-mail: zoran@ist.temple.edu*

Abstract

Identification of a small set of discriminative genes is a crucial step for effective prediction of disease or patient survival based on microarray gene expression data. Previous approaches to this problem are mainly based on analyzing differential gene expression data. In this work, an additional step is introduced to take advantage of prior knowledge about the relation of genes and a disease. In the proposed approach, keyword scanning of human proteins at the Swissprot database is performed to select genes related to the disease of interest followed by analysis of differential gene expressions. In results obtained on lung cancer data where a differential expression-based selection of genes is fairly inaccurate, our prior knowledge mining based approach offered a large improvement of prediction accuracy (0.74 vs. 0.58 ROC curve when using 20 genes). Furthermore, experimental results on a breast cancer dataset, where prediction based on differential gene expression alone was quite accurate can be further improved by integrating with our new approach.

Keywords

Feature selection; classification; gene expression analysis

1. Introduction

Compared to traditional methods that study a single or a few genes at a time, microarray technology measures expression of thousands of genes at a time. Assuming appropriate data analysis and validation, this allows more accurate disease profiling, diagnosis and treatment. One of the key objectives in this process is selecting a small subset of genes expected to be closely related to the disease whose expression levels are able to effectively diagnose diseases [1, 2] or test disease treatments [3].

A wide variety of feature selection methods have been proposed for microarray data. The most widely used method is by ranking the genes according to their significance in differential expression (DE) between diseased and normal samples using a statistical test (e.g. t-test) and selecting the best ones. Other methods are based on machine learning and other statistical methods such as SVM-recursive feature elimination [4], genetic algorithms [5], Nearest shrunken centroid [6] and Significance analysis of microarrays [7]. However, most of these feature selection methods are largely confined to analysis of expression data from microarray or from enriched functional annotation [8].

In this work, we extracted the disease related information through a prior knowledge mining technique to aid the prediction of patient survival and compared the results with conventional approaches within two cancer related microarray datasets. We show that our prior knowledge mining based approach (PKM) can offer significantly better prediction accuracy in cases where the differential expression based method (DE) fails. Furthermore, in applications where DE is fairly accurate, combining genes selected from both DE and PKM can further increase the predictive accuracy.

2. Material and Methods

2.1 Data

The methods described in Section 2 will be evaluated on the problem of predicting cancer survival based on gene expression data. To better characterize the proposed method, it will be tested on two types of cancer (lung and breast) with very different properties. The lung cancer microarray data used in this work is from [9], which contains 86 sample assays where 24 patients died and 62 survived. The breast cancer data is from [10], which contain 78 sample assays where 34 patients died and 44 survived. The expression level of genes on each chip, representing one patient sample, was normalized (divided by the mean value of that chip). The mean and standard deviation of expression levels of each gene in the training dataset were used to normalize both the training and testing dataset.

2.2 The Differential Expression (DE) Based Selection of the Most Discriminative Genes

For each gene from the training dataset, the p-value is calculated as the difference in expression between the survival and deceased group based on the t-test and the expression difference ratio calculated as the fold change between the two groups. Genes with a low p-value and high fold change were selected based on thresholds as the most informative genes.

2.3 Prior Knowledge Mining (PKM) for Selection of the Most Discriminative Genes

For each gene in the Swissprot database (Nov. 2006 version downloaded from ftp://ftp.ncbi.nih.gov) key words highly associated with the disease location and type were scanned. Only those genes that contain both location and type keywords associated with the disease were selected.

In particular, in our experiments a gene is considered to be associated with lung cancer if its description contains both cancer related keywords (five keywords were used: “oncogene”, “cancer”, “carcinoma”, “sarcoma” and “tumor”) and keywords for the location of lung cancer (“lung” and “vascular”). The same cancer related keywords are used in the breast cancer dataset, but “breast” is used as the keyword for the disease location.

Selected disease associated genes are further analyzed based on their gene expressions. A subset of low p–value genes with high fold change is selected as described for the DE method in Section 2.2.

2.4 A Hybrid Method (HY) for Selection of the Most Discriminative Genes

While genes selected by DE and PKM methods have very small overlap, as we will demonstrate in the Results section, it might be beneficial to combine the two methods into a hybrid method (HY). In this approach the genes selected based on the differential gene expression are combined with *k* top ranked genes based on the prior knowledge mining based selection. Experiments reported in Section 3 were performed using *k*=10.

2.5 Survival Prediction

Expression values of genes selected by DE, PKM, or HY process were used as features for training neural network classification models for survival prediction. This choice was made based on the demonstrated effectiveness of neural network in applications related to biomedical prediction from noisy and correlated variables. We also considered other machine learning methods (SVM, simple logistic regression and random forest) but these results were omitted, as the findings were very similar.

In our experiments the number of hidden neurons was set to 5. In the 5-cross validation process, data were randomly partitioned into five disjoint subsets. In each of the 5-cross validation experiments, since the training data were changed, different genes were identified and neural networks were trained based on the information from the training dataset alone and then tested on the test dataset. To address the non-determinism in neural networks optimization, at each round of the 5-cross validation, 30 neural networks were developed and tested (the average and standard deviation of these 30 trials were reported in the results section). In each trial, a different 20% of data were reserved for validation, while the remaining 80% were used for training of a predictor.

3. Results

3.1 Comparison on the Lung Cancer Dataset

In the lung cancer dataset [9] gene markers selected by the DE and PKM method have very small overlap. The fraction of genes shared between the two approaches averaged among the 5-cross validation was below 10% indicating that these two approaches are quite independent.

We then compared the area below ROC curves using neural network algorithm with genes selected through different methods to compare accuracy of the new methods PKM and HY to DE. As shown in Table 1 and Figure 1, using different feature selection methods, we selected between 10 and 200 genes to predict disease survival. For each method, different p-value and ratio thresholds were chosen so that the number of genes selected was about the same.

No. of Genes	10	20	30	50	100	200
DE	0.62	0.58	0.60	0.62	0.65	0.66
PKM	0.68	0.74	0.75	0.77	0.74	0.74
HY	0.68	0.70	0.65	0.67	0.69	0.68

Table 1. Area under ROC curves with different number of selected genes using DE, PKM and HY for lung cancer prediction.

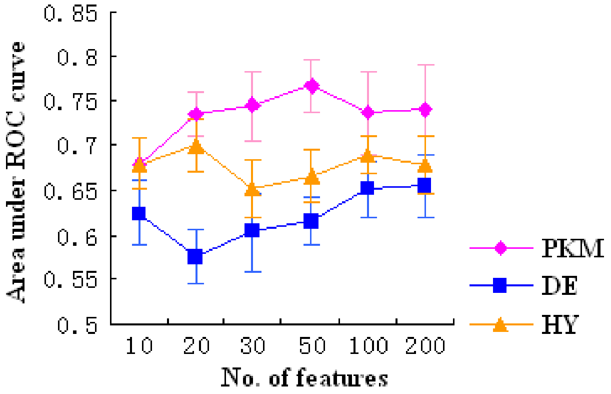

Figure 1. Comparison of area under ROC curves using different feature selection methods in the lung cancer dataset. The y-axis indicates area under ROC curves of neural network models built on features selected by the DE, PKM and HY method with different number of genes and the x-axis indicates number of genes selected. Error bars indicate the standard deviation among 30 trials.

The results obtained by selecting different numbers of genes suggest that the proposed prior knowledge mining method greatly facilitates the prediction of patient survival. The ROC curve for the differential expression based feature selection was close to the diagonal, which means that DE method was just slightly more accurate than a trivial model. The PKM showed a significant improvement when compared to DE. However, for some applications DE method alone is quite accurate. In the next section we report the results of experiments aimed at determining if PKM is beneficial in such situations.

3.2 Comparison on the Breast Cancer Dataset

We further tested our disease prior knowledge mining approach to select biomarkers on the breast cancer dataset [10]. Genes selected by the DE and PKM are again very different from each other. The fraction of selected gene markers shared between the two approaches was below 2% with 10, 20 up to 100 genes.

No. of Genes	10	20	30	50	100
DE	0.92	0.94	0.95	0.97	0.98
PKM	0.95	0.90	0.91	0.93	0.92
HY	0.95	0.97	0.97	0.98	0.99

Table 2. Area under ROC curves with different number of selected genes using DE, PKM and HY for breast cancer prediction.

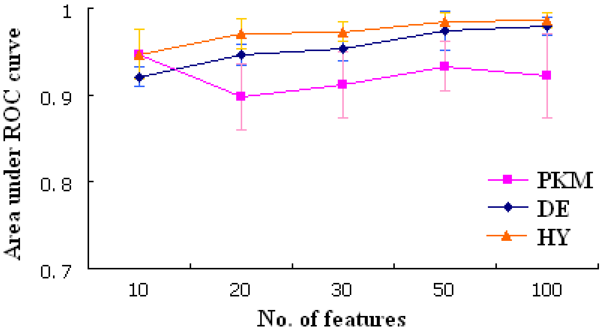

Figure 2. Comparison of area under ROC curves using different feature selection methods in the breast cancer dataset. The y-axis indicates area under ROC curves of neural network models built on genes selected by the DE, PKM and HY method with different number of genes and the x-axis indicates the number of genes selected. Error bars indicate the standard deviation among 30 trials.

As shown in Table 2 and Figure 2, using different feature selection methods, we selected between 10 and 100 genes to predict disease survival. For each method, different p-value and ratio thresholds were chosen so that the numbers of genes selected were about the same.

This result suggests that even in a dataset where DE works quite well, HY which combined PKM and DE can effectively enhance the performance of prediction. However, the previous lung cancer example shows that the hybrid method is not necessarily better than the two individual methods in all cases. It is possible that the DE method in the lung cancer example was performing poorly such that combining the DE method with the PKM method did worse than the PKM method alone. Therefore, it could be necessary to first test on validation data whether to use the prior knowledge based method or the hybrid method.

4. Conclusion

Feature selection is an important step in the prediction of diseases from gene expression patterns. While previous feature selection methods are mainly confined to information from the micro-array or gene functional annotations, we proposed a novel approach that introduces prior knowledge of the disease to achieve better predictive power. Our results obtained on lung cancer data suggest that disease prior knowledge mining based feature selection can offer improved survival prediction when differential expression based selection is inadequate. In the breast cancer dataset, where the differential expression based selection works quite well, including genes selected based on the disease prior knowledge mining was still beneficial.

The contribution of the proposed approach is that through combing disease prior knowledge mining and differential gene expression based feature selection methods, we show that integration of information from low throughput studies of diseases and high throughput micro-arrays can provide more accurate guidance for future discoveries. Nevertheless, there are limitations to our approach. Our method may be less effective in cases where the disease is not well studied and less prior knowledge is available. Furthermore, we currently have only retrieved disease information from Swissprot database. In the future, it would be useful to incorporate information from multiple databases, which is expected to capture additional relevant information and thus result in more accurate prediction.

Acknowledgements:

This project is funded in part under a grant with the Pennsylvania Department of Health. The Department specifically disclaims responsibility for any analyses, interpretations, or conclusions. We thank Joe Jupin, Uros Midic and Slobodan Vucetic for reading preliminary drafts of this manuscript and suggesting valuable comments.

References

1. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM: **Delineation of prognostic biomarkers in prostate cancer**. *Nature* 2001, **412**(6849):822-826.
2. Ressom HW, Varghese RS, Zhang Z, Xuan J, Clarke R: **Classification algorithms for phenotype prediction in genomics and proteomics**. *Front Biosci* 2008, **13**:691-708.
3. Wang S, Cheng Q: **Microarray analysis in drug discovery and clinical applications**. *Methods Mol Biol* 2006, **316**:49-65.
4. Huang TM, Kecman V: **Gene extraction for cancer diagnosis by support vector machines--an improvement**. *Artif Intell Med* 2005, **35**(1-2):185-194.
5. Yang JY, Li GZ, Meng HH, Yang MQ, Deng Y: **Improving prediction accuracy of tumor classification by reusing genes discarded during gene selection**. *BMC Genomics* 2008, **9 Suppl 1**:S3.
6. Wang S, Zhu J: **Improved centroids estimation for the nearest shrunken centroid classifier**. *Bioinformatics* 2007, **23**(8):972-979.
7. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *Proc Natl Acad Sci U S A* 2001, **98**(9):5116-5121.
8. Lottaz C, Spang R: **Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data**. *Bioinformatics* 2005, **21**(9):1971-1978.
9. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG *et al*: **Gene-expression profiles predict survival of patients with lung adenocarcinoma**. *Nat Med* 2002, **8**(8):816-824.
10. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT *et al*: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415**(6871):530-536.