Feature Ranking and Feature Redundancy Reduction for Prognostic Microarray Study of Tumor Clinical Outcomes

Qihua Tan, Mads Thomassen, Kaare Christensen, Torben A. Kruse

***Abstract*— Different from significant gene expression analysis which looks for all genes that are differentially regulated, feature selection in prognostic gene expression analysis aims at finding a subset of informative marker genes that are discriminative for prediction. Unfortunately feature selection in the literature of microarray study is predominated by the simple heuristic univariate gene filter paradigm that selects differentially expressed genes according to their statistical significance. Since the univariate approach does not take into account the correlated or interactive structure among the genes, classifiers built on genes so selected can be less accurate. More advanced approaches based on multivariate models have to be considered. Here, we introduce a feature ranking method through forward orthogonal search to assist prognostic gene selection. Application to published gene-lists selected by univariate models shows that the feature space can be largely reduced while achieving improved testing performances. Our results indicate that "significant" features selected using the gene-wised approaches can contain irrelevant genes that only serve to complicate model building. Multivariate feature ranking can help to reduce feature redundancy and to select highly informative prognostic marker genes.**

***Index Terms*— feature selection; tumor; clinical outcome prediction; microarray gene expression data**

I. Introduction

Similar to significant gene expression analysis, one demanding challenge in prognostic microarray experiments in cancer studies is the development of a powerful prognostic profile based on informative genes or features selected from a large pool of candidate genes measured on a small number of arrays or samples [1]. Among the thousands of genes measured in an experiment, it is anticipated that only a limited number of genes are informative for prognostic purposes while a large number of genes are redundant or irrelative and thus can be ignored. Inclusion of uninformative genes for tumor outcome prediction only introduces unnecessary noise and will inevitably complicate model building and introduces computational difficulties. Obtaining a smaller subset of representative genes while retaining the prognostic characteristics of the original data should lead to a more accurate and efficient learning system with improved classification performance [2]. Furthermore, for prognostic purpose, predictive expression profiles built upon limited number of genes are more useful in practice because their expression levels can be easily measured using economic techniques, for example, the quantitative real-time PCR.

Different from significant gene expression analysis which looks for all genes that are differentially regulated, feature selection in prognostic gene expression analysis aims at finding a subset of informative marker genes that are discriminative for prediction, ideally without redundancy. Ein-Dor *et al*. [3] reported that the set of outcome predictive genes is not unique due to the existence of multiple genes that are correlated with the clinical outcomes and some of them may have only small differences in their correlations. Such a context represents the hitting-set problem in finding the smallest set of features (hitting set) that encompass or characterizes all the classes [4]. The difficulty in this context is the exponential search space created by all the possible genes or markers to be considered.

In the literature of prognostic microarray study, feature selection is predominated by the simple heuristic univariate gene filtering paradigm [1]. Since the univariate approach does not take into account the correlated or interactive structure among the genes, classifiers built on genes so selected can be less accurate. More advanced approaches based on multivariate models have been considered, among them the variance-based dimension reduction [5,6]. In this paper, we introduce a feature ranking method through forward orthogonal search and apply it to published gene-lists selected by univariate models in prognostic microarray analysis. Example application of the method shows that the predictive feature space can be largely reduced while achieving improved testing performances.

II. Methods

In a microarray experiment with the expression levels of n measured genes for N samples, we use $x_{i,j}$ ($i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, n$) to represent each measurement point in the data space. Suppose the genes have been filtered to remove irrelevant genes using a gene filtering method (for example, a simple univariate statistic) and obtain a list of m potential genes (the feature space) for prediction purpose. Our objective here is to find a subset of informative marker genes or features of size d ($d \le m$) for predicting the outcomes of a testing sample. As mentioned above, the selected subset of

Manuscript received November 7, 2008. This work was partially supported by the US National Institute on Aging research grant NIA-P01-AG08761.

Q. Tan, M. Thomassen, and T.A. Kruse are with the Dept of Biochemistry, Pharmacology and Genetics (BFG), Odense University Hospital, Sdr. Boulevard 29, DK-5000, Odense C, Denmark (phone: 0045 6550 3536; e-mail: qtan@ health.sdu.dk).

Q. Tan and K. Christensen are with Epidemiology, Institute of Public Health, University of Southern Denmark, J.B. Winsløws Vej 9B, Odense C, Denmark (phone: 0045 6541 2822; fax: 0045 6541 1911; e-mail: qihua.tan@ouh.regionsyddanmark.dk).

genes should characterize the major features of the overall feature space. For that purpose, we start with calculating the squared-correlation coefficient for two vectors x_s and x_t, $s,\ t \in \{1, 2, \ldots\ldots, m\}$, each representing one feature in the feature space,

$$r^2(x_s, x_t) = \frac{\left(x_s^T x_t\right)^2}{\left(x_s^T x_s\right)\left(x_t^T x_t\right)} \quad (1)$$

We calculate the squared-correlation coefficients for all combinations of s and t. For each gene (for example *j*), we obtain the mean of the squared-correlation with all the genes as $r^2_{mean}(j) = \frac{1}{n}\sum_{s=1}^{n} r^2(x_s, x_j)$. The gene with the highest mean is then selected as the first most representative gene.

To select the second gene, each of the unselected genes indicated as *j* is orthogonalized to the selected gene using the Gram-Schmidt algorithm with the orthogonalization of the first gene z_1 equaling to x_1.

$$z_j^{(2)} = x_j - \frac{x_j^T z_1}{z_1^T z_1} z.$$

Now we repeat the procedure for selecting the first gene by calculating the squared-correlation coefficient between each of the unselected genes *j* but using its orthogonalization and each of the *n* original genes, $r^2(x_s, z_j^{(2)})$ and then obtain its mean as $r^2_{mean}(j) = \frac{1}{n}\sum_{s=1}^{n} r^2(x_s, z_j^{(2)})$.

We select the gene with the highest mean as the second gene.

Likewise, in order to select the *kth* gene, each of the unselected genes *j* is orthogonalized to the *k-1* selected genes as

$$z_j^{(k)} = x_j - \frac{x_j^T z_1}{z_1^T z_1} z_1 - \frac{x_j^T z_2}{z_2^T z_2} - \cdots - \frac{x_j^T z_{k-1}}{z_{k-1}^T z_{k-1}} z_{k-1}. \quad (2)$$

We calculate the squared-correlation coefficient between the unselected gene *j* and each of the *n* original genes as $r^2(x_s, z_j^{(k)})$ and the mean of its correlation, $r^2_{mean}(j) = \frac{1}{n}\sum_{s=1}^{n} r^2(x_s, z_j^{(k)})$.

The *kth* gene is selected as the gene with the highest mean. The process is repeated until all the genes are selected and meanwhile ranked.

With the above procedure, the most representative genes that account for the variation of the overall features with the highest percentage can be selected. The data vector for each gene or feature can be approximated by a linear combination of the selected subset of features of size *d* ($d \le m$). Following Korenberg et al. [7], we can calculate the error reduction ratio (*ERR*) as a measurement for accounting for the variation in gene *j* by the *kth* gene ($k = 1, 2, \ldots, d$) in the selected feature subset,

$$ERR(j,k) = \frac{\left(x_j^T z_k\right)^2}{\left(x_j^T x_j\right)\left(z_k^T z_k\right)} \times 100\%. \quad (3)$$

The mean percentage of variation in the overall features or genes that are accounted for by gene *k* can be calculated as

$$\overline{ERR}(k) = \frac{1}{m}\sum_{j=1}^{n} ERR(j,k). \quad (4)$$

Finally, the accumulated percentage of variation in the overall features or genes that are accounted for by the subset of *d* selected genes can be calculated as

$$\overline{SERR}(d) = \sum_{k=1}^{d} \overline{ERR}(k). \quad (5)$$

$\overline{SERR}$ serves as a measurement for the performance of the selected subset of genes and for setting up a threshold for defining the subset of genes to be selected to sufficiently represent the overall features.

All calculations are done under the free R programming environment for statistical computing.

III. Results

A. Ovarian cancer survival data

The method is first applied to a microarray study on cancer survival from Spentzos *et al.* [8] who reported prognostic significance of gene expression profiling in survival of epithelial ovarian cancer in a sample of 68 patients using Affymetrix U95A2 array containing approximately 12,000 genes. Their study identified a 115-gene signature that predicted patients with unfavorable and favorable survival outcomes at a significance level of p=0.004.

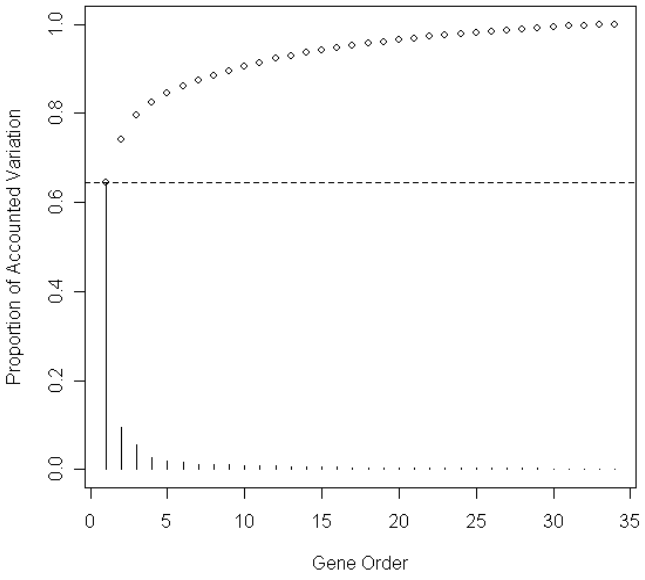

Fig. 1. The ranked (*solid bar*) and the accumulated (*empty dot*) $\overline{ERR}$ for the top genes from the 115-gene signature reported by Spentzos et al. [8].

Since the development of the 115-gene signature was based on a gene-wised testing approach, we think that the selected feature set could contain highly correlated or redundant genes that can be removed by our proposed method. After applying our method, we obtain Figure 1 showing the ranked and the accumulated $\overline{ERR}$ for the top rank genes.

Surprisingly the number one rank gene is already responsible for 65% of the total variation in the overall gene

set. The $\overline{SERR}$ for the top 3 genes can even explain about 80% of the total variation in the 115 genes. In order to examine the performance of the top rank genes, we follow the same way of dividing the samples for training and testing, i.e. 34 samples for training and 34 for testing using exactly the same samples in each group as did in the original study. We also adopt the step-wised strategy by Spentzos et al. [8] for training the model. That is we first train a classifier based on the extreme samples (shortest survivors without censoring and longest survivors, 9 for each) to classify the remaining training samples in the middle into favorable and unfavorable groups. Then the whole training set together with their group membership is used to train the final model. For convenience, we build our classifier with the support vector machines (SVM) using the free R package *e1071*. In Figure 2, we show the SVM probability for favorable survival for each sample predicted by our classifier using the top 3 genes (2a) and the Kaplan-Meier survival curves for the predicted favorable and unfavorable groups using a cut-off for SVM probability of 0.5 (2b). We can see that the long survivors (most of them censored; indicated by empty circles) are plotted on top and short survivors (most of them dead; indicated by solid circles) to the bottom of Figure 2a. As a result, we observe a remarkable difference in the survival distribution of the two groups in Figure 2b. Statistical test on differential survival between the two groups gives a χ^2 of 10.65 with 1 degree of freedom which amounts to a p value of 0.001 which is in contrast to 0.004 in the original study.

B. Breast cancer metastasis data

A 70-gene signature was reported by van't Veer et al. [9] for predicting breast cancer metastasis within 5 years with high accuracy using a 25K chip with 60-mer oligonucleotides from Rosetta. The same data was re-analyzed by Thomassen et al. [10] using similar training (61 samples: 31 metastasis and 30 non-metastasis) and testing (180 samples: 42 metastasis and 138 non-metastasis) sets as in original study but using SVM as classifier obtaining a sensitivity of 83% and a specificity of 60%. The 70-gene signature was developed using gene-wised correlation between single gene expression and metastasis outcomes. Similar to example 1, feature redundancy reduction can be conducted and improvement in prediction anticipated. We display the ranked and the accumulated $\overline{ERR}$ for the top rank genes from the 70-gene signature in Figure 3. Different from Figure 1, there is no gene with extreme contribution to the total variation. However, with the top 15 genes included in the feature subset (accounting for about 72% of the total variation), we achieve a sensitivity of 71% and a specificity of 74% when the trained SVM classifier using 61 training samples is applied to the testing set of 180 samples (Figure 4). Note that the above testing accuracy is based on setting the cut-off for SVM probability of metastasis to 0.5. We see in Figure 4 that one can easily push down the threshold to achieve a higher sensitivity while still maintain an acceptable specificity. When the cut-off is moved down to 0.45, a higher sensitivity of 86% can be reached without sacrificing so much for specificity (a lower specificity of 64%) which is an obvious improvement in both sensitivity and specificity in comparison with the results obtained using the full set of 70 genes [10].

(a)

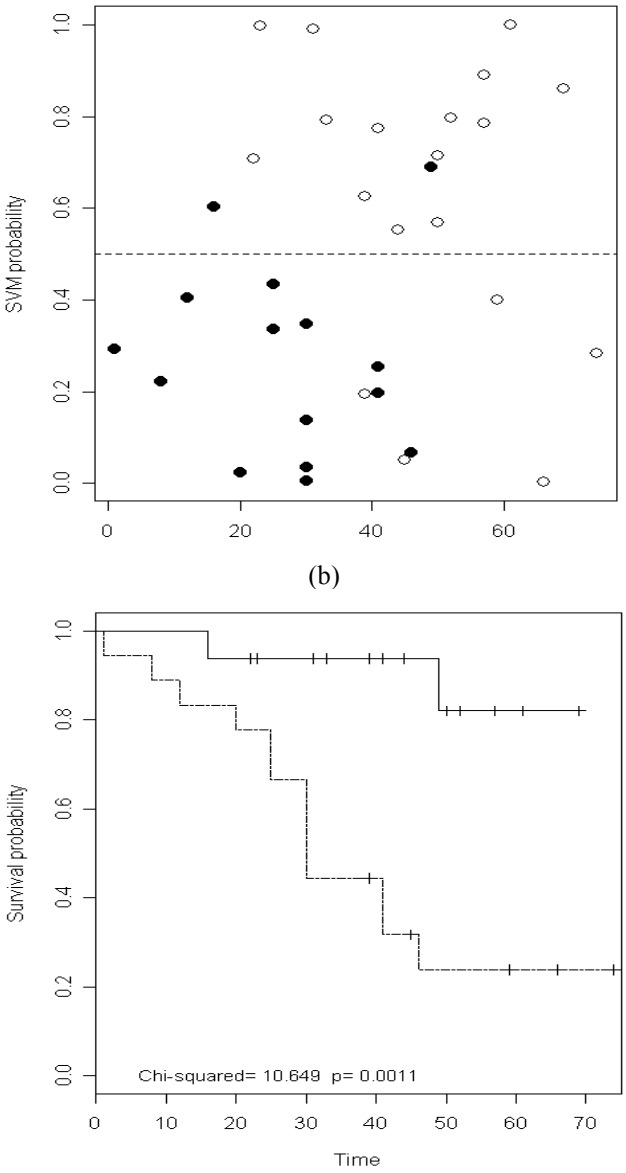

Fig. 2. Performance of the top 3 genes shown by the SVM probability for favorable survival for each sample (2a, *solid circle* for uncensored and *empty circle* for censored) and the Kaplan-Meier survival curves for the predicted favorable (*solid line*) and unfavorable (*dashed line*) groups using a cut-off for SVM probability of 0.5 (2b).

IV. DISCUSSIONS

We have shown that our unsupervised learning algorithm can be applied for feature ranking and feature redundancy reduction in prognostic studies of tumor clinical outcomes using the array-based technology. The method can be used to remove correlated genes that are with low impact on classification so that, as shown by the two examples, improved performance on an independent testing set can be expected. Our results indicate that "significant" features selected using the gene-wised approaches can contain irrelative or redundant genes that serve only to complicate model building for a classifier. Here we emphasis the difference between significant and prognostic gene expression analyses because the former looks for all genes significantly regulated (including correlated genes

co-expressed in a biological pathway) while the latter, on the other hand, tries to extract only informative and highly representative gene markers to characterize the outcomes.

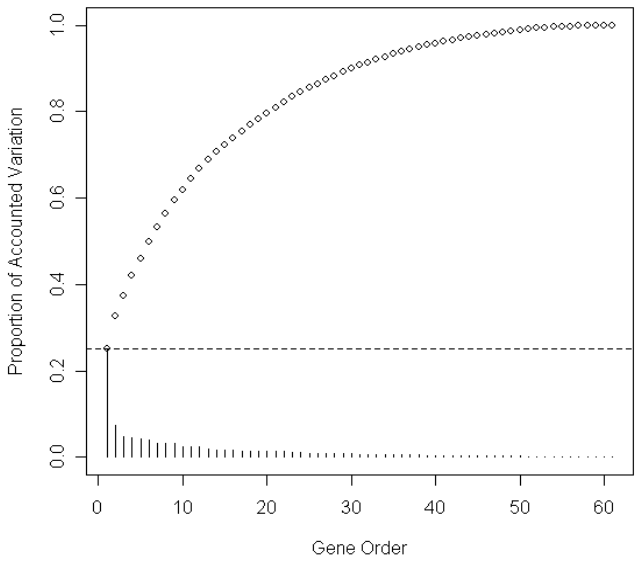

Fig. 3. The ranked (*solid bar*) and the accumulated (*empty dot*) $\overline{ERR}$ for the top genes from the 70-gene signature reported by van't Veer et al. [9].

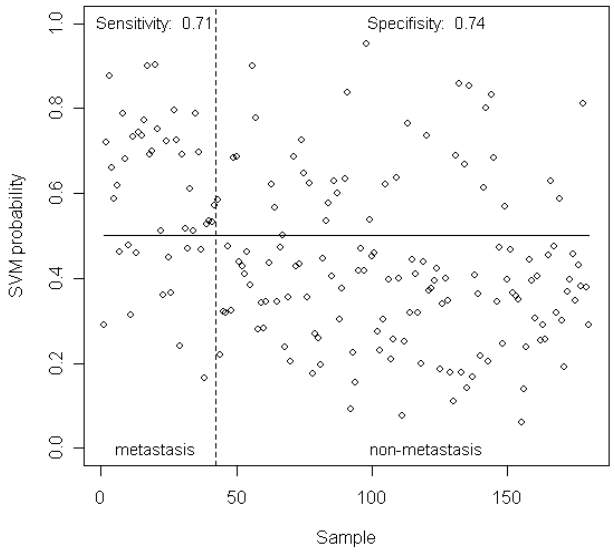

Fig. 4. SVM probability for metastasis predicted for each of the 180 samples in the testing set using the 15 selected top rank genes.

Our proposed feature ranking and feature reduction method is an unsupervised approach by nature. By unsupervised feature ranking, all genes in the feature space are ordered according to their ability in representing the original nature or explaining the total variation in the overall data. This means that one is not expected to apply this method for feature selection from the large number of genes measured in a microarray experiment because the major variation in the whole data may not be predominated by the outcome status and its related expression profiles. For practical application, we suggest first find all genes that are significantly correlated with tumor outcome status including both dependently and independently regulated genes and then use the proposed method to remove significant genes but with low impact.

Feature redundancy reduction not only helps to improve performance and generalization of the classifier, it is also advantageous for clinical applications. As mentioned in the beginning, clinical use of the confirmed subset of highly representative and prognostic genes can be made possible by engaging economic methods such as quantitative real time PCR (qrt-PCR) technology which can be used as a routine bioinstrumentation for gene expression level measurement [11]. Implementation of such methods will certainly help to develop more efficient and individualized treatment strategy to improve survival of cancer patients.

ACKNOWLEDGMENTS

We thank Dr. Dimitrios Spentzos at Beth Israel Deaconess Medical Center in Boston for help in accessing their data and for providing the list of their signature genes.

REFERENCES

[1] Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics. 23, 2507--2517 (2007)
[2] Gasca, E., Sanchez, J. S., Alonso, R.: Eliminating redundancy and irrelevance using a new MLP-based feature selection method. Pattern Recognition. 39, 313--315 (2006)
[3] Ein-Dor, L., Kela, I, Getz, G., Givol, D., Domany, E.: Outcome signature genes in breast cancer: is there a unique set? Bioinformatics. 21, 171--178 (2005)
[4] Selman B. : A hard statistical view. Nature. 451, 639--640 (2008)
[5] Guo, Q., Wu, W., Massart, D. L., Boucon, C., de Jong, S.: Feature selection in principal component analysis of analytical data. Chemometrics and Intelligent Laboratory System. 61, 123--132 (2002)
[6] Gasca, E., Sanchez, J. S., Alonso, R.: Eliminating redundancy and irrelevance using a new MLP-based feature selection method. Pattern Recognition. 39, 313--315 (2006)
[7] Korenberg, M., Billings, S. A., Liu, Y. P., Mcllroy, P. J.: Orthogonal parameter estimation algorithm for non-linear stochastic systems. Int'l J. Control. 48, 193--210 (1988)
[8] Spentzos, D., Levine, D. A., Ramoni, M. F., Joseph, M., Gu, X., Boyd, J., Libermann, T. A., Cannistra, S. A.: Gene expression signature with independent prognostic significance in epithelial ovarian cancer. J. Clin. Oncol., 22, 4700--4710 (2004)
[9] van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., et al.: Gene expression profiling predicts clinical outcome of breast cancer. Nature. 415, 530--536 (2002)
[10] Thomassen, M., Tan, Q., Eiriksdottir, F., Bak, M., Cold, S., and Kruse, T.A.: Prediction of metastasis from low-malignant breast cancer by gene expression profiling. International Journal of Cancer. 120, 1070--1075 (2006)
[11] Livak, K. J., Schmittgen, T. D.: Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. Methods. 25, 402--408 (2001)