Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

On sparse Fisher discriminant method for microarray data analysis

Eric S. Fung[1] and Michael K. Ng[1,*]

[1]Centre for Mathematical Imaging and Vision and Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong; Michael K. Ng* - E-mail: mng@math.hkbu.edu.hk; * Corresponding author

received December 14, 2007; accepted December 28, 2007; published online December 30, 2007

Abstract:
One of the applications of the discriminant analysis on microarray data is to classify patient and normal samples based on gene expression values. The analysis is especially important in medical trials and diagnosis of cancer subtypes. The main contribution of this paper is to propose a simple Fisher-type discriminant method on gene selection in microarray data. In the new algorithm, we calculate a weight for each gene and use the weight values as an indicator to identify the subsets of relevant genes that categorize patient and normal samples. A l_2 - l_1 norm minimization method is implemented to the discriminant process to automatically compute the weights of all genes in the samples. The experiments on two microarray data sets have shown that the new algorithm can generate classification results as good as other classification methods, and effectively determine relevant genes for classification purpose. In this study, we demonstrate the gene selection's ability and the computational effectiveness of the proposed algorithm. Experimental results are given to illustrate the usefulness of the proposed model.

Keywords: microarray; Fisher discriminant method; data; genes; algorithm

Background:
Microarray technologies for the analysis of biological samples provide information on a genomic scale. A major challenge in the context of microarray is the task of sample classification. One key problem in microarray data classification is that the number of features (gene expression levels) is extremely large compared to the number of observations (samples). Traditional pattern recognition methods may not handle this challenge properly. It is essential to identify which genes are relevant in the classification of disease so that better RNA-based diagnostic tests using laboratory techniques such as RT-PCR and better treatment can be developed.

Researchers **[3, 6]** have also developed methods to identify optimal sets of genes which together provide good discrimination of classes. These algorithms are generally very computationally intensive. Recently, various machine learning methods for gene selection have been developed, for instance, relevance vector machine **[11]**, Gaussian process models **[5]** and simple decision rules **[12]**. Fisher discriminant analysis and least squares support vector machines are used for sample classification **[9]**. Another approach is to use optimization algorithms in feature selection like sparse logistic regression **[14]** and modified Fisher optimization model **[7]**.

The main contribution of this paper is to propose a simple Fisher-type discriminant method on gene selection in microarray data. In the new algorithm, we calculate a weight for each gene and use the weight values as an indicator to identify the subsets of relevant genes that categorize patient and normal samples in two-class classification problems. This is achieved by including the weight sparsity term in the Fisher objective function that is minimized in the discriminant process as described in equation 1 (see supplementary material). Each entry of u represents a weight for each gene. An efficient l_2 - l_1 norm minimization method is implemented **[8]** to the above discriminant model to automatically compute the weights of all genes in the samples. The experiments on two microarray datasets have shown that the new algorithm can effectively determine a small set of genes for the purpose of classification, and can generate classification results that are as good as the other methods.

Results and discussion:
Datasets
In this paper, we apply the proposed method to two public microarray data sets, namely, colon cancer data set from **[1]** and the Leukaemia MIT AML/ALL data set from **[10]**.

Colon cancer data
In order to obtain more reliable results **[15]**, we performed ten-fold cross validation in the experiments. The k-nearest neighbor's method is used to determine a classifier that can be applied to predict the class of expression profiles of test samples. In the experiments, we tried several values of α. For each value of α, ten cross validation cases are generated

and therefore ten sets of weights of genes are obtained. Based on these ten sets of weights, the mean weights of genes can be calculated and thus genes are ranked according to the magnitude of their mean weights. We apply this ranking to the ten cross validation cases and evaluate how many numbers of important (relevant) genes to be selected such that the highest classification accuracy can be obtained.

In the tests, we found out that the highest classification accuracy is achieved when α = 1609 among all tested values of α. In Figure 1a, we show the classification accuracy curve for 10-fold cross validation based on the ranking of average weights of genes when α = 1609. We note that the classification accuracy is still 82.4% even when the number of genes selected is more than 30, i.e., even if we include more genes in the classifier, the classification accuracy cannot be improved. We see from the figure 1 that when the number of genes selected is three, we can obtain the highest classification accuracy (86.7%). Among the ten cross validation cases, 5 out of 10 cases are 100% correct. The type I and type II errors are 25.0% and 7.5% respectively when α = 1609.

In Table 1 (supplementary material), we list the mean weights, the mean values of cancer samples and the mean value of normal samples for the three selected genes. We observe that their sample mean discrepancies of two classes are quite large. This may also suggest why they are selected and why they are relevant to a normal/disease sample classification. In Figure 1b, we plot the value of equation (2) (see supplementary material) for each training sample *j*, where $[x_{.j}]$ is a vector containing those selected genes expression of the *j*-th sample and $\overline{u}_3$ represents a projection vector which is formed by using the average weights of the three selected genes.

Leukaemia MIT AML/ALL data

We also performed ten-fold cross validation for the Leukaemia data set. We found out that the highest average classification accuracy is achieved when α = 10 among all tested values of α. We show in Figure 1c that the classification accuracy curve for 10-fold cross validation based on the ranking of average weights of genes. We also note that the classification accuracy is still 91.5% even when the number of genes selected is more than 120. Obviously, we obtain the highest classification accuracy (95.8%) when the number of genes selected is 39. It is interesting to note that 7 out of 10 cases are 100% correct. The type I and type II errors are 0.0% and 11.7% respectively.

In Table 2 (supplementary material), we observe that their sample mean discrepancies of two classes are quite large. In Figure 1d, we plot the value of equation (3) (see supplementary material) for each training sample *j*, and it is clear from the figure 1 that the selected genes categorize patient and normal samples are well separated.

Comparison of methods

In this section, we compare the proposed method with other classification methods.

Modified Fisher discriminant method

In this subsection, we compare the performance of the proposed method with the modified Fisher discriminant method described in **[7]**. By using the colon cancer data set, we randomly selected half of the normal samples and patient samples as training samples and the rest of them as testing samples repeated 100 times. Here we fix α = 1069 as used in the previous subsection, and compare the results of the two methods. The classification accuracy for testing samples is 85.0 ± 13.8% and only one gene ("Hsa.8147") is selected. On the other hand, the classification accuracy for testing samples in **[7]** is 86.0 ± 5.7% and the number of genes selected is 29.9 ± 4.8%. We see that the proposed method is quite competitive with the modified Fisher discriminant method.

Secondly, we perform the same experiment by using the Leukaemia data set. We randomly selected half of the normal samples and patient samples as training samples and the rest of them as testing samples. Therefore, we have 36 training samples and 36 testing samples repeated 100 times. Here we fix α = 10 as used in the previous subsection. The classification accuracy for the test samples is 86.9 ± 14.7% and the number of genes selected is 58. No average result was given in **[7]** because large memory storage is required and the method is time- consuming. However, the proposed method can generate classification results efficiently.

Sparse logistic regression

In order to make a fair comparison with sparse logistic regression **[4]**, we also perform a leave-one-out validation procedure to test the performance of the proposed method. We calculate the mean weights of genes in the procedure and evaluate how many numbers of genes to be selected such that the highest classification accuracy can be obtained.

In the colon cancer data set, we find that when α is equal to 1, the classification accuracy, cross-entropy and number of selected genes of the proposed method are 83.9%, 0.31 and 9 respectively. It is better than those by the method (BLogReg) in **[4]**, which gives lower classification accuracy (82.3%), higher cross-entropy (0.51) and more number of selected genes (11). In the Leukaemia MIT AML/ALL data set, we find that when α is equal to 0.1, the classification accuracy, cross-entropy and number of selected genes of the proposed method are 95.8%, 0.087 and 8. It is better than those by the method (BLogReg) in **[4]**, which gives a lower classification accuracy (93.1%), a higher cross-entropy (0.259), and more selected genes (11). We remark that the lower cross-entropy is, the better the classification result is.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

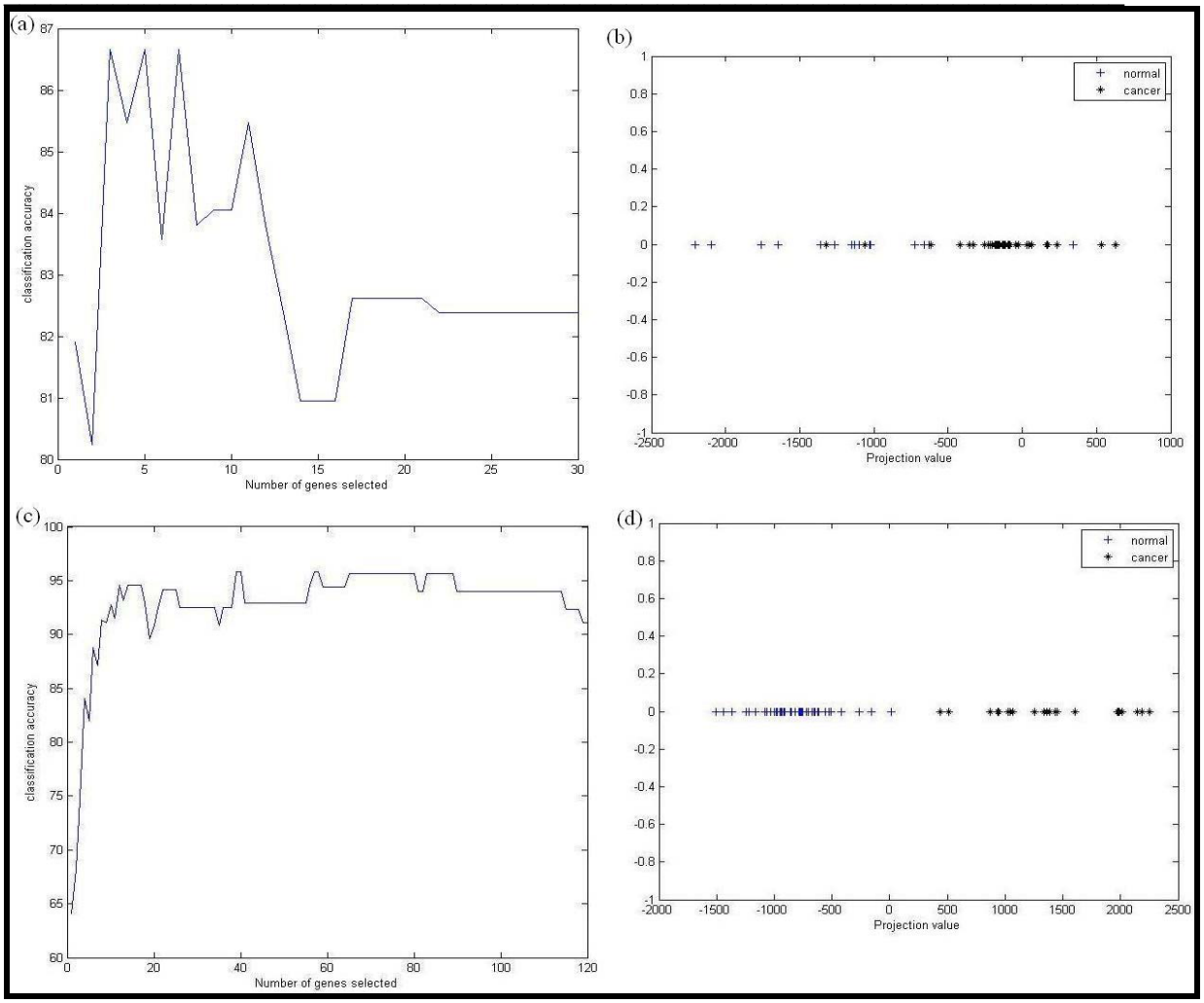

Figure 1: Classification accuracy and projection values. (a) classification accuracy (%) when α = 1609; (b) projection values when α = 1609; (c) classification accuracy (%) when α = 10; (d) projection values when α = 10

PAM

PAM is a tool for classifying normal/disease samples based on microarray data **[2]**. The idea behind nearest shrunken centroids **[13]** is to calculate each class centroid as a nearest centroid classifier. Each centroid is divided by the within-class standard deviation for each gene. This gives greater weight to genes whose expression is stable among samples in the same class. Soft thresholding is applied to the resulting normalized class centroids. If the normalized centroid is small, it is set to zero. This procedure is to reduce the number of genes that are used in the final classification model. The method is very efficient as it does not involve covariance matrix of genes, and the nearest shrunken centroids can be computed independently.

In **[2]**, it is mentioned that the discriminant weights in PAM are similar to those used in linear discriminant analysis. The main difference is that the calculation of distance between a given test observation and the class centroids where the pooled within-class variance/covariance matrix of the expression data is used. In PAM, it assumes that the covariance matrix is a diagonal matrix. In the proposed method, we use the covariance matrix in the formulation so that pairwise relations between any two genes are used in the calculation of discriminant weights. On the other hand, shrunken centroids are used in PAM. In the proposed method, we use a weight sparsity term $\| u \|_1$ in the objective function to control the discriminant weights. Similar to PAM, a cross-validation procedure is used to find out a good balance (α) between equation (4) (see supplementary material) and $\| u \|_1$. We remark that α is the regularization parameter to control the sparsity of *u*, i.e., very small values are set to zero. The corresponding gene does not contribute to the final classification.

Conclusion:
In this paper, we study a new Fisher discriminant method for gene selection in microarray data and propose a l_2 - l_1 norm minimization method for finding the projection vector in discriminant process. The experiments on two microarray data sets have shown that the new algorithm can generate classification results in a competitive manner compared with other classification methods, and can effectively determine relevant genes.

Acknowledgement:
Michael Ng is supported in part by Hong Kong RGC grant numbers 7035/04P, 7035/05P and HKBU FRGs.

References:
[01] U. Alon, *et al.*, *Proc. Nat. Acad. Sci.,* 96: 6745 (1999) [PMID: 10359783]
[02] E. Bair & R. Tibshirani, *SIGKDD Explorations,* 5: 48 (2003)
[03] T. Bo & I. Jonassen, *Genome Biology,* 3: 1 (2002) [PMID: 11983058]
[04] G. C. Cawley & N. L. Talbot, *Bioinformatics,* 22: 2348 (2006) [PMID: 16844704]
[05] W. Chu, *et al.*, *Bioinformatics,* 21: 3385 (2005) [PMID: 15937031]
[06] J. Deutsch, *Bioinformatics,* 19: 45 (2003) [PMID: 12499292]
[07] J. Feng, *et al.*, *Advances in data mining and modeling,* 15 (2002)
[08] H. Fu, *et al.*, *SIAM Journal on scientific computing,* 27: 937 (2006) [PMID: 16331583]
[09] T. S. Furey, *et al*., *Bioinformatics,* 16: 906 (2000) [PMID: 11120680]
[10] T. R. Golub, *et al.*, *Science,* 286: 53 (1999) [PMID: 10521349]
[11] Y. Li, *et al.*, *Bioinformatics,* 18: 1332 (2002) [PMID: 12376377]
[12] A. C. Tan, *et al.*, *Bioinformatics,* 21: 3896 (2005) [PMID: 16105897]
[13] R. Tibshirani, *et al.*, *Statist. Sci.* 18: 104 (2003)
[14] S. Shevade & S. Keerthi, *Bioinformatics,* 19: 2246 (2003) [PMID: 14630653]
[15] I. A. Wood, *et al.*, *Bioinformatics,* 23: 1363 (2007) [PMID: 17392326]

Edited by T. W. Tan & S. Ranganathan
Citation: Fung & Ng, Bioinformation 2(5): 230-234 (2007)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Supplementary material

Equations:

$$\| S_w u - z \|_2^2 + \alpha \| u \|_1 \quad \rightarrow \quad (1)$$

Here S_w is the within-class scatter matrix of the samples in a micro-array data and z comes from the between-class scatter matrix, u is the projection vector and α is the regularization parameter to control the sparsity of u.

$$\bar{u}_3^T [x_{.j}] \quad \rightarrow \quad (2)$$

$$\bar{u}_{39}^T [x_{.j}] \quad \rightarrow \quad (3)$$

$$\| S_w u - z \|_2^2 \quad \rightarrow \quad (4)$$

Tables:

Gene IDs	$\bar{u}_i$	$\bar{c}_i$	$\bar{n}_i$
Hsa:8147	0.329	597	2303
Hsa:1737	0.151	2578	3661
Hsa:140	0.105	3870	2090

Table 1: The information of the selected genes when α = 1609 (where, $\bar{u}_i$ = mean weights, $\bar{c}_i$ = mean values of cancer samples and $\bar{n}_i$ = mean value of normal samples).

Gene IDs	$\bar{u}_i \times 10^4$	$\bar{c}_i$	$\bar{n}_i$
J03779_at	0.039	1817	3200
J04164_at	0.036	1275	3369
Y00787_s_at	0.034	7790	766
M33680_at	0.031	2608	6351
Y00433_at	0.021	14155	8117
X69150_at	0.021	23913	22192
X95404_at	0.021	12260	11486
M27891_at	0.020	9120	185
M19507_at	0.020	8605	443
M84526_at	0.018	5125	-137
X68277_at	0.017	4947	5392
M13792_at	0.016	2062	5553
HG1872−HT1907_at	0.016	2540	1419
M20203_s_at	0.014	4510	110
X537774_at	0.014	11041	11731
X17093_at	0.013	1772	4177
J05614_at	0.013	1518	2892
U14970_at	0.012	17077	17501
X56997_rna1_at	0.012	12716	12046
M19045_f_at	0.012	6475	1736

Table 2: The information of the selected genes when α = 10.