A COMBINATORY ALGORITHM OF UNIVARIATE AND MULTIVARIATE GENE SELECTION

[1]H. Mahmoodian, [2]M. Hamiruce Marhaban, [3]R. A. Rahim, [4]R. Rosli, [5]M. Iqbal Saripan

[1]PhD student, Department of Electrical and Electronic, Eng. Faculty, UPM, Malaysia

[2]Assoc. Prof., Department of Electrical and Electronic, Eng. Faculty, UPM, Malaysia

[3]Prof, Department of Cell and Molecular Biology, Biotech. Faculty, UPM, Malaysia

[4]Assoc. Prof., Medicine and Health Science Faculty, UPM, Malaysia

[5]Lecturer, Department of Computer and Communication System, Eng Faculty, UPM, Malaysia

E-mail: gs19182@mutiara.upm.edu.my, hamiruc@eng.upm.edu.my, raha@bioteh.upm.edu.my, rozita@medic.upm.edu.my, eqbal@eng.upm.edu.my

ABSTRACT

Microarray technology has provided the means to monitor the expression levels of a large number of genes simultaneously. Constructing a classifier based on microarray data has emerged as an important problem for diseases such as cancer. Difficulty arises from the fact that the number of samples are usually less than the number of genes which may interact with one another. Selection of a small number of significant genes is fundamental to correctly analyze the samples. Gene selection is usually based on univariate or multivariate methods. Univariate methods for gene selection cannot address interactions among multiple genes, a situation which demands the multivariate methods [1], [2]. In this paper, we considered new parameters which come up from singular value decomposition and present a combination algorithm for gene selection to integrate the univariate and multivariate approaches and compare it with gene selection based on correlation coefficient with binary output classes to analyze the effect of new parameters. Repeatability of selected genes is evaluated by external 10-fold cross validation whereas SVM and PLR classifiers are used to classify two well known datasets for cancers. We calculated the misclassification error in training samples and independent samples of two datasets (breast cancer and Leukemia). The results show that the mean of misclassification error of training samples in 100 iteration are almost equal in two algorithms but our algorithm have the better ability to classify independent samples.

Keywords -*Singular value decomposition, penalized logistic regression, gene selection*

1. INTRODUCTION

DNA microarrays constitute one of the most powerful high-throughput technologies in molecular biology today. It has emerged in recent years as a powerful tool that provides a glimpse into the complexities of cellular behavior through the window of transcriptional activity. Gene expression array experiments present us with vast amounts of data containing substantial biological information. In order to obtain the most from these data, a considerable variety of approaches for statistical and algorithmic analyses have been developed. One of the most important problems in analyzing microarray data is gene selection which is usually based on two major methods called univariate and multivariate gene selection.

In univariate approach to gene selection the behavior of each gene is considered alone and does not address interactions among multiple genes, a situation which demands the multivariate approach. Some univariate approach are such t-score-base statistic which sorts genes based on the t-test values [3]-[6], maximum likelihood ratio approach to rank genes in the order of most discriminating to least discriminating between two classes [7], X^2statistics or a correlation-based feature selection [8].Interactions between genes is considered in multivariate approaches which have been developed in different methods such as using

PCA (Principal Component Analysis) or SVD (Singular Value Decomposition) that explicitly use the high dimensional nature of the gene expression space. Some previous researches used SVD to reduce the dimension of gene expression profiles with at least losing information [9],[10],[11],[12],[13], other works on multivariate gene selection have been presented in [14] and [15]. In this paper, it is tried to analyze an algorithm based on combination of univariate gene ranking such as correlation coefficient algorithm and multivariate gene selection based on SVD.

The remainder of this paper is organized as follows. Section II presents the mathematical framework and the gene selection algorithm. Then, in section III we illustrate the results of the classification of two well known microarray datasets which are classified by Support Vector Machine (SVM) and Penalized Logistic Regression (PLR) [16]. In section IV, we discuss about the results and compare the ability of two methods of gene selection in section V.

2. MATHEMATICAL FRAMEWORK AND GENE SELECTION ALGORITHM

2.1 Singular Value Decomposition

The SVD of a linear transformation $G: R^n \rightarrow R^m$ is $G = USV^T$ where $U: R^m \rightarrow R^m$ and $V^T: R^n \rightarrow R^n$ are orthogonal matrixes and $S: R^n \rightarrow R^m$ is a nonnegative diagonal matrix whose diagonal elements are named singular values and usually are sorted descending ($\sigma_0 > \sigma_1 > \ldots > \sigma_{r-1}$). It can be shown that matrix G is equal to:

$$G = \sum_{i=1}^{r} \sigma_i U_i V_i^T \qquad (1)$$

where, r is the rank of matrix G, σ_i is ith singular value, U_i and V_i^T are the first r columns of U and V respectively. Choosing 0< p <r-1 there is a new matrix which is defined as:

$$G_p = \sum_{i=1}^{p} \sigma_i U_i V_i^T \qquad (2)$$

where $G_p$is the optimum approximation of G in view two norm ($\|.\|_2$)and Frobenius norm ($\|.\|_F$) [17].In gene expression dataset, it is supposed that there is $G^{n \times m}$ matrix (usually n>> m) consisting n row of gene expression in m samples. SVD of matrix G produces two orthogonal bases $U$and V^T (left and right singular vectors respectively) which the former defines new space of samples and the latter defines a new space of gene in G. Column vectors of U and row vectors of V^T are usually called eigenarray and eigengene respectively [10].

By considering (1), it can be shown that i^{th} row of matrix$G$, is

$$i^{th} row\ of\ G = \sum_{k=1}^{m} u_{ik} \sigma_k V_k \qquad (3)$$

which V_k is k^{th} row of matrix V. So, each gene can be represented by a new space which is spanned based on the orthogonal vectors (eigengenes).

2.2 Pearson Correlation Coefficient

Pearson Correlation Coefficient (PCC) (5) has been usually used to show linearity dependence of two vectors X and Y with N dimension. This parameter is defined such as:

$$PCC(X,Y) = \frac{(\sum XY - (\frac{\sum X \sum Y}{N}))}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} \qquad (4)$$

The correlation is 1 in the case of an increasing linear relationship, −1 in the case of a decreasing linear relationship, and some value in between all other cases, indicating the degree of linear dependence between the variables. The closer the coefficient is to either −1 or 1, the stronger the correlation between the variables.

2.3 Gene Selection

Suppose that g_i is the i^{th} row of matrix G which presents the expression of i^{th} gene in the samples. For i^{th} gene, we define Square Root of Sum of Squares ($SRSS$) such as (5).

$$SRSS_i = \sqrt{(C_{iV1}^2 + C_{iV2}^2 + C_{icv}^2)} \qquad (5)$$

In (5) C_{iV1}=PCC(g_i,V_1) and C_{iV2}=PCC(g_i,V_2) are the PCC between g_i and two first rows of V^Twhich are two eigengenes with respect to two largest singular value of gene expression profile matrix. High correlation between genes and these two eigengenes shows high linearity relation between genes and matrix$G$. Since more information of G is in two largest singular value [10], $C_{iV1}$and C_{iV2} increase the rank of genes that handle more information of G which includes interaction between genes. C_{icv} is the PCC of genes with the binary classes of samples.

For gene selection, we sorted the genes descending respect to the SRSS and classified the training samples with the first gene and then added the remain genes one by one. For each set of

genes, we used external 10-fold cross validation to calculate the misclassification errors in 100 iteration. Adding genes one by one, till

there was not significant change in mean of misclassification errors. This occurred when the number of genes reached to around 150 for both breast cancer and Leukemia dataset.

In all steps of gene selection and classification (training and independent samples), the rank of genes are measured with both equation (5) (we call it CC+SVD method) and absolute of C_{icv} (we call it CC method) separately to analyze the effect of C_{iV1} and C_{iV2} on selected genes.

3. RESULTS

To analyze the effect of consideration of C_{iV1} and C_{iV2}, we implemented this algorithm on breast cancer dataset and leukemia datasets and classified the training and independent samples by two different classifiers (SVM and PLR).

The first dataset is classification of relapse time of lymph node negative breast cancer samples, reported by van't Veer [18] and the second is molecular class discovery and class prediction of cancer between Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL) samples reported by Golub [19]. We analyzed the breast cancer dataset in two categories with BRCA1 and BRCA2 samples and without them.

In the breast cancer dataset, there are 97 training samples (20 samples have BRCA1 or BRCA2 mutation) including 46 tumors with relapse time greater than 5 years (2 of 46 are mutated with BRCA1) and 51 tumors with the relapse time less than 5 years (18 of 52 are mutated with BRCA1 or BRCA2). In the dataset, there are about 25000 human genes which after preselecting the genes with at least 2-fold changes in ratio with p-value less than 0.01, the number of genes reduced to about 5100 genes. For evaluation the gene selection method and classification, there are 19 extra samples including 7 and 12 samples with relapse time greater and less than 5 years respectively. Because we analyzed the data set with mutated genes (BRCA1 and BRCA2) and without them, we trained the classifier with 77 and 97 samples separately and 19 samples for validation.

Table 1. Misclassification errors in training 77 samples of breast cancer

Breast Cancer dataset (77 samples)	Percentage of misclassification error (PLR classifier)		Percentage of misclassification error (SVM classifier)	
Number of selected genes	CC+SVD	CC	CC+SVD	CC
10	33.66	33.21	41.11	41.80
20	32.99	32.75	42.8	42.72
30	31.10	32.37	42.85	42.85
40	32.40	31.78	42.85	42.85
50	31.63	32.7	42.85	42.85
60	32.03	31.98	42.85	42.85
70	31.51	31.85	42.85	42.85
80	31.88	31.94	42.85	42.85
90	32.92	32.50	42.85	42.85
100	33.54	33.05	42.85	42.85

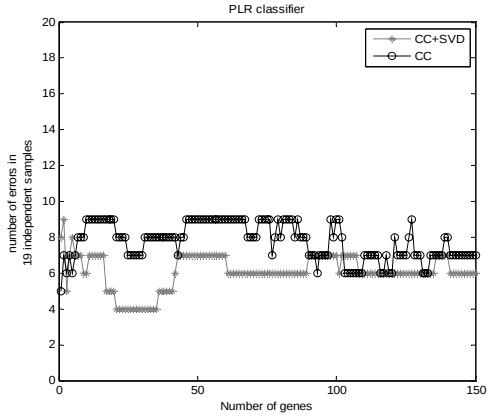

Figure 1. The number of errors in 19 independent samples (PLR classifier)

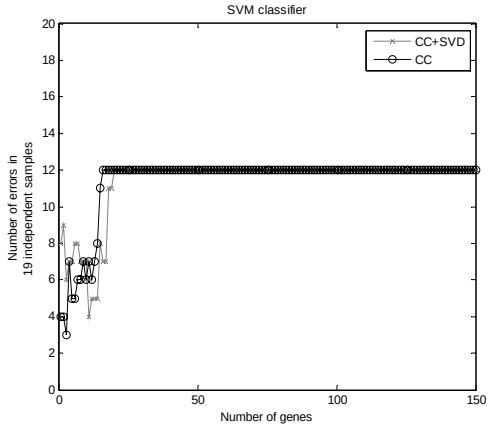

Figure 2. The number of errors in 19 independent samples(SVM classifier)

The leukemia dataset includes 38 training samples (27 ALL and 11 AML) with 7129 genes and 34 independent samples (20 ALL and 14 AML).

3.1 Breast Cancer (77 Training Samples)

PLR and SVM classifiers were used to classify the training samples based on external 10-fold cross validation. The folded training samples for both CC+SVD method and CC method were the same. The procedure of calculating the misclassification error was repeated 100 times and the mean of misclassification error in 100 iteration was measured for each subset of selected genes. For each subset of selected genes, the misclassification error for independent samples was also measured. For PLR classifier, we used cross validation method to select the best value for penalty factor (λ). Table 1 show the percentage of misclassified errors and the number of selected genes for training and independent samples which are analyzed based on CC and CC+SVD methods.

To validate the procedure of gene selection and classification methods, we classified the 19 extra independent samples by SVM and PLR based on two methods of gene selection. Figure 1 and 2 shows the results for 77 breast cancer samples.

3.2 Breast Cancer (97 Training Samples)

We repeated the same procedure for 97 samples which include BRCA1 and BRCA2 mutated samples. Table 2 compares the results between two methods.

To validate the subset of selected genes, the 19 independent samples were classified by PLR and SVM-RBF classifiers which the results are shown in figure 3 and 4.

3.3 Leukemia dataset

The same procedure is used to select the significant genes of 7129 genes in the Golub dataset. Table 3 presents the mean of misclassified errors in many subsets of selected genes with linear SVM and PLR classifiers

For validation the procedure, the 34 extra independent samples were used. Figures 5 and 6 show the number of errors out of 34 samples.

Table 2. Misclassification errors in training 97 samples of breast cancer

Breast Cancer dataset (97 samples)	Percentage of misclassification error (PLR classifier)		Percentage of misclassification error (SVM classifier)	
Number of selected genes	CC+SVD	CC	CC+SVD	CC
10	50.10	50.92	40.82	38.67
20	50.05	48.07	36.80	37.36
30	42.87	43.26	36.98	38.91
40	41.91	42.25	38.70	38.18
50	39.15	39.37	40.17	39.95
60	38.18	39.19	40.18	41.25
70	35.71	37.2	37.20	39.33
80	36.84	38.77	40.23	41.33
90	37.93	37.01	41.38	41.22
100	37.52	36.87	41.38	41.33

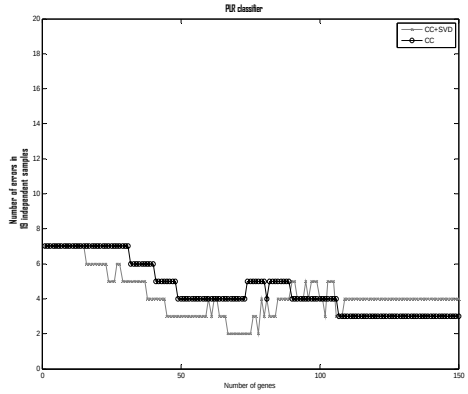

Figure 3. The number of errors in 19 independent samples (PLR classifier)

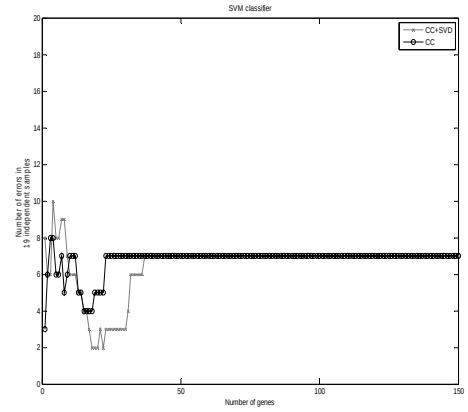

Figure 4. The number of errors in 19 independent samples (SVM classifier)

4. DISCUSSION AND CONCLUSION

Table 1,2 and 3 show that there are no significant differences in the rate of the misclassification error in training samples between two methods. But figures 1, 2,3,4,5 and 6 present that using CC+SVd method reduces the rate of misclassification error in independent samples.

We measured the mean of misclassification error for independent samples in all 150 subset of selected genes to compare the ability of classification in both methods in independent samples. The results are in table 4 which illustrates that error rate in CC+SVD method is less than CC method in all datasets and two method of classifiers (only for ALL/AML dataset and with SVM classifier, the error rate of CC method is a little bit less than CC+SVD method).

In addition, considering the figures 1,3 and 5, shows that CC+SVD method have the ability to classify the independent samples with a suitable accuracy. For instance, using PLR classifier, with 22,67 and 27 of high ranked genes, the number of errors were 4 out of 19, 2 out of 19 and 1 out of 34 in independent samples of 77 breast cancer dataset, 97 breast cancer dataset and ALL/AML dataset respectively.

It shows that CC+SVD method which considers the genes that are high correlated with outcome and eigengenes have more ability to classify the independent samples.

Considering weighting parameters in relation (5) may promote the quality of selected subsets of genes which need more effort and research in future.

Table 4. mean of misclassification error of independent samples in all subsets of genes

Mean of errors in independent samples		CC+SVD	CC
77 breast cancer	PLR	6.1	7.7
	SVM	11.39	11.41
97 breast cancer	PLR	4.24	4.6
	SVM	6.5	6.7
ALL/AML	PLR	2.9	4.4
	SVM	13.61	12.98

Table 3. Misclassification errors in training 38 samples of ALL/AML

ALL/AML dataset	Percentage of misclassification error (PLR classifier)		Percentage of misclassification error (SVM classifier)	
Number of selected genes	CC+SVD	CC	CC+SVD	CC
10	13.15	9.20	18.15	8.45
20	7.1	6.5	17.12	15.1
30	9.7	18	14.12	15.0
40	10	10.70	12.12	13.1
50	11.80	10.58	12.12	13.1
60	10.12	10.15	12.12	13.1
70	9.56	10.34	12.12	13.1
80	9.2	7.5	12.12	13.1
90	10.1	9.4	12.12	13.1
100	8.89	7.89	12.12	13.1

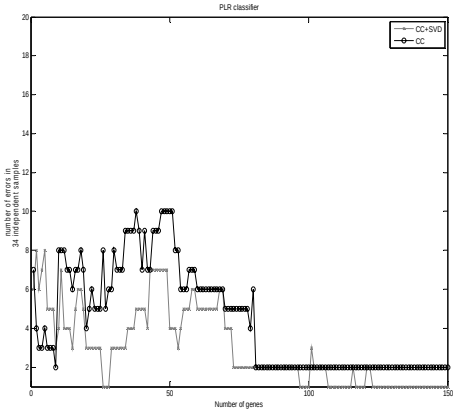

Figure 5. The number of errors in 34 independent samples (PLR classifier)

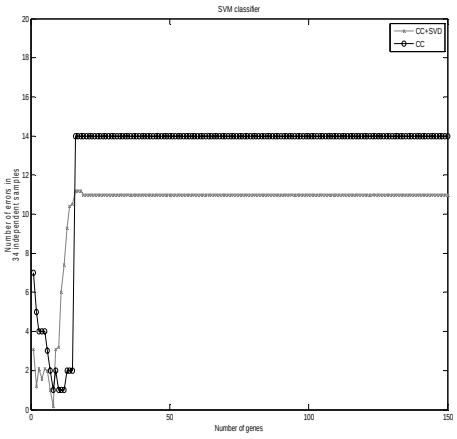

Figure 6. The number of errors in 34 independent samples (SVM classifier)

REFERENCES

[1] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, 2002, pp. 389–422.

[2] T. Bo, and I. Jonassen, "New feature subset selection procedures for classification of expression profiles," *Genome Biol.*, vol. 3, no.4, pp.0017.1-0017.11, 2002, PMCID:MC115205.

[3] W. Pan, "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 18**,** no.4, 2002, pp.546–554.

[4] O.G. Troyanskaya, M.E. Garber, P.O. Brown, D. Botstein, and R.B. Altman," Nonparametric methods for identifying differentially expressed genes in microarray data," *Bioinformatics,* vol.18, no.11, 2002, pp. 1454–1461.

[5] W. Pan, J. Lin, and C.T. Le," How many replicates of arrays are required to detect gene expression changes in microarray experiments? a mixture model approach," *Genome Biol* , 3, no.5, pp.0022.1-0022.10, 2002, PMCID: PMC115224.

[6] J. Li, H Liu, J.R Downing, A.E. Yeoh, and L. Wong," Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients," *Bioinformatics***,** vol. 19, no. 1, 2003, pp.71–78.

[7] W. Li, and Y. Yang, "Zipf's law in importance of genes for cancer classification using microarray data," *Theor Biol* , 219, no. 4, 2002, pp.539–551.

[8] H. Liu, J. Li, and L. Wong," A comparative study on feature selection and classification methods using gene expression profiles and proteomics patterns," *Genome Informatics*, vol. 13, 2002, pp.51–60.

[9] N.S. Holter, M. Mitra, A. Maritan, M. Cieplak, J.R. Banavar, and N.V. Fedoroff," Fundamental patterns underlying gene expression profiles: simplicity from complexity," *Proc Natl Acad Sci USA,* vol. 97, no. 15, 2000, pp.8409–8414.

[10] O. Alter, P.O. Brown, and D. Botstein," Singular value decomposition for genome-wide expression data processing and modeling," *Proc Natl Acad Sci USA,* vol. 97, no.18, 2000, pp.10101–10106.

[11] T.O. Nielsen, R.B. West, S.C. Linn, O. Alter, M.A. Knowling, J.X. O'Connell, S. Zhu, M. Fero, G. Sherlock, J.R. Pollack, P.O. Brown, D. Botstein, and M. van de Rijn," Molecular characterisation of soft tissue tumours: a gene expression study," *Lancet*, vol. 359, pp.1301–1307, 2002, doi:10.1016/S0140-6736(02)08270-3.

[12] Y. Kluger, R. Basri, J.T. Chang, and M. Gerstein," Spectral biclustering of microarray data: coclustering genes and conditions," *Genome Res*, vol. 13, 2003, pp.703–716.

[13] S. Bicciato, A. Luchini, and C. D. Bello," PCA disjoint models for multiclass cancer analysis using gene expression data," *Bioinformatics*, vol.19, no. 5, 2003, pp.571–578.

[14] M. Dettling, and P. Buhlmann," Supervised clustering of genes," *Genome Biol*, vol. 3, no. 12, 2002, pp.0069.1-0069.15.

[15] J.M. Deutsch," Evolutionary algorithms for finding optimal gene sets in microarray prediction," *Bioinformatics*, vol. 19, no.1, 2003, pp.45–52.

[16] M.Y. Park, and T. Hastie,"Penalized logistic regression for detecting gene interaction," *Biostatistics*, vol. 9, no. 1, pp.30-50, 2008,doi:101093/biostatistics/kxm010.

[17] S. Theodoridis and K. Koutroumbas, Pattern recognition, second edition, academic press, 2003, pp. 215-216.

[18] L. J. Van't Veer, H. Dai, M. J. Van de Vijver, Y. D. He, et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Letters to Nature* , 415, 2002, pp. 530-536.

[19] T. Golub, D.K. Slonim, P. Tamayo, C. Huard, M Gaasenbeek, J.P. Mesirov, H. Coller, ML. Loh, J.R. Downing, M.A. Caliqiuri, C.D. Bloomfield, and E.S. Lander," Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science* , vol. 286, 1999,pp. 531–537.