Identifying The Most Significant Genes From Gene Expression Profiles For Sample Classification

Hisham Al-Mubaid and Noushin Ghaffari, *Member, IEEE*

Abstract-- **The gene expression data generated by the Microarray technology for thousands of genes simultaneously provide huge amounts of biomedical data in forms of gene expression profiles. This generated gene data include complex variations of expression levels of thousands of gene in the classes of samples. The gene level variations allow for classifying and clustering the samples based on only a small subset of genes. In this work, we want to identify the *most significant* genes that demonstrate the highest capabilities of discrimination between the classes of samples. We present a new gene selection technique for extracting the most significant genes from the huge gene/feature space in a given gene expression dataset. Our method is based on computing the *discriminating capability* of each gene, and classifying the data according to only those *most significant* genes that have highest discriminating capabilities. We also adapted from text categorization and information retrieval five feature selection techniques into the gene selection task to compare with our method. We evaluated the method using four well-known gene expression datasets. The experimental results showed that our method produces impressive and competitive results in terms of classification performance with few selected genes compared with the existing techniques.**

Keywords: Bioinformatics, Gene Selection, Gene Classification.

I. INTRODUCTION

The DNA Microarray technology was first introduced in 1995 for measuring the expression levels of thousands of genes simultaneously [1, 2, 7]. These genes are amplified by Polymerase Chain Reaction (PCR) technique. A robot spots the PCR resulted genes onto an ordinary glass microscope slide. The next process denatures and links the spotted DNA to the glass slide [7]. Each microscope slide contains a grid-like pattern like an array with thousands of spots of amplified copies of each gene. Immobilized DNA on the microarray will be hybridized with a probe, which is a known labeled DNA sequence. In order to make the probe, mRNA is isolated from control or diseased samples, and converted into cDNA. The nucleotides used to produce cDNA include a green dye (Cy3) or a red dye (Cy5) [5, 7]. Since each microarray contains thousands of DNA spots, the output numeric data is too much to be processed manually. So, there is a great demand for efficient methods for analysis and manipulation of gene expression data. These data include complex variations among expression levels of each gene in the normal vs disease tissue samples, which allows for classifying and clustering the samples into normal vs disease based on only a small subset of the genes. The goal of this work is to extract those genes that demonstrate high discriminating capabilities between the classes of samples. We propose a new method for gene classification and extraction using various feature selection techniques. Our method is based on computing thresholds and discriminating capabilities of each gene and classifying the data according to only those genes that have highest capabilities to discriminate between the two classes (viz. *normal, disease)* of samples. The method extracts very small subsets of useful salient genes that can improve the classification accuracy of tissue samples. We applied the method on four well-known gene expression datasets. We also applied five other feature selection techniques from the text classification literature to compare with our method. The method produces encouraging and competitive results in terms of classification performance compared with recent similar techniques.

A number of methods have been proposed and applied into gene expression profiles in the last few decades. Paul and Iba [1] modified the Probabilistic Model Building Genetic Algorithm (PMBGA) into a Random PMBGA (RPMBGA) for gene selection. They tried to reduce the size of gene subsets while keeping accuracy of classification in the high level. For the same task, Liu et al. [2] used the neural network for gene expression profiles. They used 100 iterations of resampled data as an input to their architecture, which consists of three neural network feature selection methods. They used Kent Ridge datasets [11] and found 100% accuracy for ALLAML and Lung Cancer data, and 97.06% for the prostate cancer dataset [2].

II. THE PROPOSED METHODS

DNA Microarray technology produces 2D representation of gene expression levels containing 2 or more classes of (*tissue*) samples. The variation in the expression levels of each gene between *class-1* vs. *class-2* determines how much that gene is related to one of the two classes. The gene that demonstrates high differences in its expression levels between *class-1* and *class-2* is a good "*significant*" gene that is typically highly related with the disease of *class-2* samples (assuming *normal* vs *disease* tissue samples). Our method for gene selection is adapted from the feature selection techniques in the text categorization (TC) and information retrieval (IR) literatures

H. Al-Mubaid and N. Ghaffari are with the Computer Science department, University of Houston-Clear Lake, Houston, TX 77058 USA (e-mail: hisham@uhcl.edu, tel: +1 (282) 283-3802)

[12, 13]. These feature selection techniques, like *Mutual Information* and *Chi-square*, are based on selecting the salient features from a huge feature space based on the feature values in the underlying classes. Our method is based on computing a discriminating value *V* for each gene in the dataset. A gene with the highest *V* value is the one that have the highest differences in its expression levels between the two classes of samples. Then we sort the genes based on their computed *V* values, select the top *n* genes, and delete the remaining (unselected) genes for the data. Before we delve into the details we explain how we compute the thresholds which are needed for our feature selection techniques.

A. Selecting Thresholds

We want to find such a threshold *t* that separates the gene expression levels in *class-1* for its levels in *class-2* with the least *noise*. For example, if the expression levels of gene *g* in *class-1* are all positive and in *class-2* are all negative, then in this case a threshold value of zero (*t=0*) is the best value that gives the least noise (zero noise). For that, we examine, for each gene, all the values from the minimum gene expression value in all samples to the maximum one and select the value that gives lowest noise as a threshold.

B. Computing V Values

Suppose we are given a gene expression matrix with two classes of samples: *classe-1* and *class-2*. Assume further that we have a threshold value *t*. For each gene we define four values *a*, *b*, *c*, and *d* as follows:

a = # of gene expression values of gene *g* in *class-1* $\geq$ t
b = # of gene expression values of gene *g* in *class-1* $<$ t
c = # of gene expression values of gene *g* in *class-2* $\geq$t
d = # of gene expression values of gene *g* in *class-2* $< t$

For example, if the threshold $t = 0$, then we compute for a given gene how many one of its expression values in *class-1* are above or equal to 0 (*a* value), or below 0 (*b* value); and how many one of its expression values in *class-2* are above or equal 0 (*c* value), or below 0 (*d* value). Furthermore, for a given threshold *t*, the most useful gene is the one that has the highest *a* and *d* values and lowest *b* and *c* values. Then, the measure [(a + d) - (b + c)] is a good indicator of how much a gene differentiates between the two classes. Thus, we compute for each gene a *V* score using our method as follows:

$$V = (a+d)-(b+c) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{..(1)}$$

This method (*Eq.1*) selects the genes that demonstrate the highest separation in their expression values from *class-1* to *class-2*. Then, to evaluate our method and compare it with the similar feature selection techniques, we borrowed and adapted from the *IR* and *TC* research four other feature selection techniques: *Mutual Information* (*MI*), *Chi-square* (*X^2*), *GSS-Coefficient*, and *Odd Ratio* (*OR*) [12, 13] and are defined as follows:

$$\mathrm{MI} = \frac{N*a}{(a+c)*(a+b)} \qquad \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

$$\mathrm{X}^2 = \frac{N*(a.d - c.b)}{(a+c)*(b+d)*(a+b)*(c+d)} \qquad \ldots\ldots\ldots\ldots(3)$$

$$\text{GSS-Coeff} = \frac{a.d - a.c}{N^2} \qquad \ldots\ldots\ldots\ldots\ldots\ldots(4)$$

$$\mathrm{OR} = \frac{(a + 0.5)*(d + 0.5)}{(c + 0.5)*(b + 0.5)} \qquad \ldots\ldots\ldots\ldots\ldots\ldots(5)$$

where N is the total number of samples in both classes. We further adapted from *MI* (*Eq.2*) a new feature selection technique by giving more weight to the *a* value of each gene and to the difference (*a* - *b*). So we multiplied *MI* by *a* and also by (*a* – *b*), and the resulting formula are shown in *Eq.6:*

$$\text{MI-2} = a*(a-b) * \frac{N*a}{(a+c)*(a+b)} \qquad \ldots\ldots\ldots\ldots\ldots(6)$$

C. Learning and classification

We evaluate the selected gene subset using machine learning with a *two-class* classification based on only the *n* selected genes. We use support vector machines (SVM) [9, 10] for learning and classification. Numerous theoretical justifications exist in the literature to support SVM [10]. We take two classes at a time and apply SVM to train on them and produce a classifier (*model*). The classifier will then be used in the classification phase to classify the testing samples. We use the SVM-*light* implementation with the default parameters (svmlight.joachims.org*)*.

III. EXPERIMENTS AND RESULTS

A. Datasets

We used four microarray datasets to evaluate our method: ALL-AML Leukemia [4], Lung cancer [3], Prostate cancer [2], and the Diffuse Large B-cell Lymphoma (DLBCL)[5]. Table 1 contains the details of these datasets. These datasets (downloaded from [11]) are used commonly for sample classification and gene clustering research.

B. Results and Discussion

We ran our method to select subsets of most significant genes for a number of subset sizes. Subsets with size of 1, 2,… ,10, 50 and 100 genes were selected. We evaluated our method (*Eq.1*) using the four datasets described in Table 1. Table 2 summarizes the results; each dataset was tested with subsets of 1, 2,…,10 selected genes.

TABLE 1
The four gene expression datasets used in our experiments

Dataset	Number of Genes	Number of Classes	Training Set	Testing Set
ALL-AML Leukemia [4]	7129	2 AML vs. ALL	38 samples: 27 ALL & 11 AML	34 samples: 20 ALL and 14 AML
Lung cancer [3]	12533	2 MPM vs. ADCA	32 samples: 16 MPM and 16 ADCA	149 samples: 134 ADCA and 15 MPM
Prostate Cancer [2]	12600	2 tumor vs. normal	102 samples: 52 tumor and 50 normal	34 samples: 25 tumor and 9 normal
DLBCL [5]	4026	2 germinal vs. activated	47 samples: 24 germinal and 23 activated	NA

TABLE 2
The classification accuracy results of our proposed method with subsets of sizes 1 to 10 genes on the four gene expression datasets.

Number of Genes	**Datasets**			
	AMLALL	**Prostate Cancer**	**Lung Cancer**	**DLBCL**
10	97.06%	100.00%	98.66%	100.00%
9	97.06%	100.00%	98.66%	100.00%
8	97.06%	100.00%	98.66%	100.00%
7	97.06%	100.00%	98.66%	100.00%
6	**97.06%**	100.00%	98.66%	100.00%
5	91.18%	100.00%	98.66%	100.00%
4	94.12%	100.00%	**99.33%**	90.00%
3	94.12%	100.00%	96.64%	**100.00%**
2	88.24%	**100.00%**	98.66%	80.00%
1	94.12%	97.06%	98.66%	70.00%
Average	**94.71%**	**99.71%**	**98.53%**	**94.00%**

These results (Table 2) demonstrate that our feature selection technique produces the highly significant features as the classification accuracy is very impressive. For example, with 6 genes only, the method produces accuracy of 97.06% correct classifications in the AML-ALL dataset, and 100% accuracy with only two genes in the prostate cancer dataset. To further evaluate the method, we conducted experiments on the four methods *MI, X^2, GSS-Coeff,* and *OR* (*Eqs. 2, 3, 4, 5*) using the same experimental settings of Table 2, and the results are in Tables 3, 4, 5, 6. From these experimental results we can see that our method is superior in selecting the most significant genes. The classification performance with only ten genes or less showed that our technique can produce accuracy on average 94.00% to 99.71% (Table 2), whereas the other four methods are lagging behind, from which, X^2 comes next with average accuracy of one to ten genes on the four datasets ranges from 93.00% to 97.06%. Moreover, we examined our second technique (*MI-2)* that we adapted from *MI* on the same setting, and the results are in Table 7. These results, in Table 7, can be easily compared with the *MI* results in Table 3 to realize that our technique (*MI-2*) outperforms *MI* significantly. For example, on the *Lung Cancer* dataset, our method gave average accuracy of 98.93% (Table 7) while *MI* produced for the same dataset 67.32% (Table 3). Furthermore, if we compare the performance of our *MI-2* technique to X^2, *GSS*, and *OR* (Tables 4, 5, 6) we again notice that it is competitive and effective in selecting the significant genes. In another set of experiments, we used 50 and 100 genes selected by our method and by the other methods, and the results are in Table 8. Again, our method outperforms the other techniques on three out of the four datasets (Table 8). We also notice that the *GSS-Coefficient* technique works very well in the case of 50 and 100 genes. Finally, Table 9 summarizes, for each dataset, the average accuracy of one to ten genes selected using our method and the other four feature selection techniques. As we can see in Table 9 that our method on average (of 1 to 10 genes) produced the best accuracy results on three datasets *AML-ALL, Prostate cancer,* and *DLBCL.* And in the fourth dataset (*Lung cancer*) our method produced the second best accuracy (98.53%) and very close to the best accuracy of 98.93% produced by the *GSS* method.

TABLE 3
The classification accuracy results using *MI* for feature selection with subsets of sizes 1 to 10 genes on the four datasets.

Number of genes	**AMLALL**	**Prostate Cancer**	**Lung Cancer**	**DLBCL**
	Accuracy	**Accuracy**	**Accuracy**	**Accuracy**
10 genes	64.71%	73.53%	75.84%	90.00%
9 genes	67.76%	73.53%	74.50%	90.00%
8 genes	64.71%	73.53%	73.83%	90.00%
7 genes	64.71%	70.59%	65.77%	90.00%
6 genes	64.71%	70.59%	55.03%	70.00%
5 genes	50.00%	73.53%	47.65%	70.00%
4 genes	64.71%	73.53%	47.65%	70.00%
3 genes	64.71%	73.53%	52.35%	80.00%
2 genes	58.82%	73.53%	90.60%	80.00%
1 gene	58.82%	73.53%	89.93%	60.00%
Average	**62.37%**	**72.94%**	**67.32%**	**79.00%**

TABLE 4
The classification accuracy results using X^2 for feature selection with subsets of sizes 1 to 10 genes on the four datasets.

Number of genes	**AMLALL**	**Prostate Cancer**	**Lung Cancer**	**DLBCL**
	Accuracy	**Accuracy**	**Accuracy**	**Accuracy**
10 genes	97.06%	97.06%	96.64%	100.00%
9 genes	97.06%	97.06%	96.64%	100.00%
8 genes	97.06%	97.06%	96.64%	100.00%
7 genes	97.06%	97.06%	96.64%	100.00%
6 genes	97.06%	97.06%	96.64%	100.00%
5 genes	97.06%	97.06%	96.64%	100.00%
4 genes	97.06%	97.06%	97.32%	90.00%
3 genes	94.12%	97.06%	97.32%	90.00%
2 genes	85.29%	97.06%	97.32%	80.00%
1 gene	85.29%	97.06%	98.66%	70.00%
Average	**94.41%**	**97.06%**	**97.05%**	**93.00%**

TABLE 5
The classification accuracy results using *GSS-Coefficient* for feature selection with subsets of sizes 1 to 10 genes on the four datasets.

Number of genes	**AMLALL**	**Prostate Cancer**	**Lung Cancer**	**DLBCL**
	Accuracy	**Accuracy**	**Accuracy**	**Accuracy**
10 genes	61.76%	97.06%	99.33%	100.00%
9 genes	73.53%	97.06%	99.33%	100.00%
8 genes	73.53%	97.06%	99.33%	100.00%
7 genes	67.65%	97.06%	99.33%	100.00%
6 genes	64.71%	97.06%	99.33%	100.00%
5 genes	94.12%	97.06%	97.99%	100.00%
4 genes	88.24%	97.06%	98.66%	90.00%
3 genes	88.24%	97.06%	98.66%	100.00%
2 genes	88.24%	97.06%	98.66%	80.00%
1 gene	94.12%	97.06%	98.66%	70.00%
Average	**79.41%**	**97.06%**	**98.93%**	**94.00%**

TABLE 6
The classification accuracy results using *OR* technique for feature selection with subsets of sizes 1 to 10 genes on the four datasets.

Number of genes	AMLALL	Prostate Cancer	Lung Cancer	DLBCL
	Accuracy	Accuracy	Accuracy	Accuracy
10 genes	94.12%	100.00%	97.99%	80.00%
9 genes	94.12%	100.00%	97.99%	80.00%
8 genes	97.06%	100.00%	97.32%	90.00%
7 genes	58.82%	100.00%	97.32%	70.00%
6 genes	58.82%	100.00%	96.64%	100.00%
5 genes	58.82%	100.00%	96.64%	90.00%
4 genes	91.18%	97.06%	97.32%	90.00%
3 genes	82.29%	97.06%	97.32%	90.00%
2 genes	76.47%	97.06%	97.32%	80.00%
1 gene	73.53%	94.12%	98.66%	70.00%
Average	**78.52%**	**98.53%**	**97.45%**	**84.00%**

TABLE 7
The classification accuracy results using *MI-2* technique for feature selection with subsets of sizes 1 to 10 genes on the four datasets.

Number of genes	AMLALL	Prostate Cancer	Lung Cancer	DLBCL
	Accuracy	Accuracy	Accuracy	Accuracy
10 genes	97.06%	38.24%	98.66%	100.00%
9 genes	94.12%	38.24%	99.33%	100.00%
8 genes	88.24%	41.18%	99.33%	100.00%
7 genes	91.18%	35.29%	98.66%	70.00%
6 genes	91.18%	88.24%	98.66%	90.00%
5 genes	94.12%	94.12%	98.66%	90.00%
4 genes	88.24%	91.18%	99.33%	90.00%
3 genes	94.12%	97.06%	99.33%	90.00%
2 genes	94.12%	97.06%	98.66%	80.00%
1 gene	88.24%	97.06%	98.66%	70.00%
Average	**92.06%**	**71.77%**	**98.93%**	**88.00%**

IV. CONCLUSION

This paper explores feature selection techniques within the context of gene expression data for sample classification. We presented two new gene selection techniques and compared them with several features selection techniques adapted from the information retrieval literature. The methods extract small gene subsets that allow for sample classification with high accuracy. Since the gene expression profiles are usually produced from disease and normal tissue samples, the extracted genes are considered as related with that underlying disease. Furthermore, identifying a small group of genes that can classify the gene expression data with very high accuracy leads to the discovery that these selected genes are *associated* with that disease studied in the gene expression dataset. In the experimental results, the proposed techniques demonstrated superior or very competitive performance in terms of accuracy of sample classification. In the future work, we plan to investigate how we can support our findings by exploiting the huge amounts of biomedical literature (*e.g. Medline*) using text-mining techniques.

TABLE 8

TABLE 8: The classification accuracies for all datasets with 50 and 100 genes selected using our method as well as the other feature selection techniques

Dataset		Our method	X^2	MI	GSS Coeff	OR
AMLAL L	100 genes	85.29%	85.29%	67.65%	97.06%	76.47%
	50 genes	97.06%	85.29%	58.82%	97.06%	82.35%
Prostate Cancer	100 genes	100.00%	32.35%	61.76%	97.06%	26.47%
	50 genes	100.00%	26.47%	50.00%	100.00%	29.41%
Lung Cancer	100 genes	93.29%	90.60%	91.28%	100.00%	93.96%
	50 genes	95.97%	93.29%	91.28%	100.00%	95.97%
DLBCL	100 genes	100.00%	100.00%	60.00%	100.00%	100.00%
	50 genes	100.00%	100.00%	60.00%	100.00%	100.00%

TABLE 9
Summary of the average accuracy of one to ten genes on the four datasets for our method and the other four feature selection techniques.

	Our method	X^2	MI	GSS Coeff	OR
AMLALL	94.71%	94.41%	62.37%	79.41%	78.52%
Prostate Cancer	99.71%	97.06%	72.94%	97.06%	98.53%
Lung Cancer	98.53%	97.05%	67.32%	98.93%	97.45%
DLBCL	94.00%	93.00%	79.00%	94.00%	84.00%

REFERENCES

[1] T. K. Paul and H. Iba. Extraction of Informative Genes from Microarray Data. *GECCO'05,* June 25–29, 2005, Washington, DC, 2005 USA.

[2] B. Liu, Q. Cui, T. Jiang and S. Ma. A combinational feature selection and ensemble neural network method for classification of gene expression data. *BCM Bioinformatics* 5:136, 2004,.

[3] G. J. Gordon, R. V. Jensen, L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker and R Bueno. Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma. *Cancer Research* 62, 4963-4967, September 2002.

[4] T.R. Golub, Slonim, Tamayo, Huard, Gaasenbeek, Mesirov, Coller, Loh, Downing, Caligiuri, LoomÞeld, Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *SCIENCE* Vol. 286, October 1999.

[5] A.A. Alizadeh et. al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *NATURE,* vol. 403, Feb. 2000.

[6] D. V. Nguyen and D. M. Rocke. Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics* 18(9), Pages 1216-1226, Oxford University Press 2002.

[7] J. Newton. Analysis of Microarray Gene Expression Data Using Machine Learning Techniques .CMPUT 606: Computational Molecular Biology and Bioinformatics.

[8] N. Ghaffari and H.Al-Mubaid. A New Gene Selection Technique Using Feature Selection Methodology, *To Appear*, *Proc. CATA-06*, Seattle, 2006.

[9] V. Vapnik. The nature of statistical learning theory. Springer, New York, USA, 1995.

[10] B. E. Boser, I. Guyon, V. Vapnik. A training algorithm for optimal margin classifiers. In COLT, pages 144-152, 1992.

[11] Kent Ridge biomedical datasets:
http://sdmc.lit.org.sg/GEDatasets/Datasets.html

[12] Zheng, Z., Srihari, R,. (2003). *Optimally Combining. Positive and Negative Features for Text Categorization.* ICML,2003 Workshop.

[13] B. C. How and Narayanan K. An Empirical Study of Feature Selection for Text Categorization based on Term Weightage. Proc. of IEEE/WIC/ACM Intl Conf on Web Intelligence (WI'04). 2004