Proceedings of the 2006 WSEAS International Conference on Mathematical Biology and Ecology, Miami, Florida, USA, January 18-20, 2006 (pp140-145)

Raking and Selection of Differentially Expressed Genes from Microarray Data

J. SHAIK and M. YEASIN
Computer Vision Pattern and Image Analysis (CVPIA) laboratory
Electical and Computer Engineering
University of Memphis, Memphis, TN- 38152, USA

Abstract: - This paper presents adaptive algorithms for ranking and selecting differentially expressed genes from microarray data. A ranking method originally proposed in [1] is adapted and supplemented with Hausdorff distance-based ranking method to improve the performance of the ranking algorithm. A weighted fusion scheme is developed to fuse the 'mean' and the Hausdorff distance-based ranking methods to develop a robust ranking method. The normalized consistency measure is used as the weight for the fusion of ranking methods. An adaptive subspace iteration (ASI) based selection algorithm is then applied on top ranked genes to select highly differentially expressed genes. To illustrate the utility of the proposed algorithms, a number of empirical analyses were conducted on both the simulated (400 simulated microarray dataset) and real microarray datasets (colon cancer dataset, gastric cancer dataset). From the empirical analysis it was observed that the proposed unified approach is robust against initialization and yields consistent selection of differentially expressed genes.

Key-Words: - Adaptive Sub-space Iteration, Clustering, Ranking, Differentially Expressed Genes and Micro-array Data Analysis.

1 Introduction

Real microarray data sets have small number of variables (in the order of 10^2 – 10^4) and samples/experimental conditions (in the order of 10^1-10^2). Several problems arise in analyzing microarray data that include (not limited to): (i) small sample size when compared to features; (ii) relative importance of individual samples; (iii) inadequate understanding of the underlying model distribution; (iv) experimental noise; (v) lack of ground truth information; (vi) redundancy among the high ranked genes. Several algorithms (e.g., using statistics [2-8], information theory [9-15], or on some functions of classifier outputs [4]) have been reported in ranking the microarray data. The key problems with most of the reported algorithms include (not limited to) (i) sensitivity to the initialization; (ii) lack of adaptivity in ranking and selection of differentially expressed genes and (iii) absence of evaluation methodologies of the computed results.

To solve some of the above mentioned problems this paper presents a unified framework in finding differentially expressed genes using adaptive ranking and selection algorithms. The ranking algorithm originally proposed in [1] is adapted and supplemented with Hausdorff distance-based ranking method to improve the performance of the ranking algorithm. A weighted fusion scheme is developed to fuse the 'mean' and the Hausdorff distance-based ranking methods to develop a robust ranking method. The normalized consistency measure (cf. equation 2) has been used as the weights for the fusion of ranking methods. An adaptive subspace iteration (ASI) based selection algorithm is then applied on top ranked genes to select highly differentially expressed genes [8, 16]. The computed results were validated using the silhouette index of the clusters.

The problem relating the mean method can be alleviated using Hausdorff distance measure. It works with unequal number of samples in both cases and random selection of samples is not necessary. The Hausdorff distance may also be influenced by the outlier sample(s). This problem can be addressed by using the K^{th} Hausdorff distance. Also the samples themselves are involved in finding the difference of expression rather than a single statistic representing all the samples like in the case of 'mean' method. To improve the robustness both the mean and Hausdroff distance-based method are fused using the consistency measure.

Selection and validation of differentially expressed genes is performed using the ASI algorithm on a fixed number of top ranked genes. It is hypothesized that if the top ranked genes fall into the same cluster they may be highly differentially expressed. This assumption may not always hold as expected. The solution to this problem can be found in ASI

clustering process. The ASI algorithm provides the information about the role and relative importance of samples in cluster formation process.

The rest of the paper is organized as follows. Section II presents a short overview of reported literature. Following this (Section III) discusses the proposed unified ranking and selection of differentially expressed genes. Section IV presents empirical analysis on simulated and real microarray data. Section V discusses the merits and demerits of the proposed approach and finally section VI concludes the paper.

2. Literature Review

The rankings of the genes are influenced by gene dependencies and feature selection. Basic idea behind using second method is to utilize inter gene dependencies rather than individual gene information as in t-statistic. Different methods have different information sharing. Examples include Significance analysis of microarrays (SAM) [14], B-statistic, a Bayesian based approach [17] and ANOVA based approaches [18]. In this paper, a Monte Carlo based method is followed in ranking of genes previously proposed by [1] and supplement it with Hausdorff distance method. The ranking function is similar to the t-statistic function where the independent parameters involved are selected using Monte Carlo simulation.

After the genes are ranked according to some criterion, the next step is gene selection. It is the task of determining which genes are the ones that are significantly differentially expressed. Informal approaches involve Q-Q plots [5]. An expected proportion of genes in a set of genes called false discovery rate may be used to measure statistical significance of the genes [2]. The hypothesis that largely differentiated ones are not only the ones contributing to the process under study brings into picture Bayesian ANOVA for microarrays (BAM) which strike balance between false rejections and false non-rejections. Other categorization methods include clustering methods [6, 7, 9, 10, 13, 15, 16, 19-22], correlation based methods, wrappers and embedded methods [12], nested subset methods [6], supervised feature selection methods and other statistical methods [7, 23]. In this paper we propose an adaptive subspace based algorithm for selecting most differentially expressed genes [8, 16].

3. Proposed Approach

The proposed approach for finding differentially expressed genes from microarray data has three major steps, namely,

1. Preprocessing of data to remove noise and reduce the dynamic range of the data.
2. Ranking of the genes based on their differential expression.
3. Selection of most differentially expressed genes from highly ranked genes.

3.1 Preprocessing of Microarray Data

The genes are first log transformed to reduce the dynamic range of the data. This process enables that lower weighted samples are not ignored. If the data is not log transformed, the affect of the samples having smaller values is totally diminished and hence reducing the dynamic range of the data is essential. The data is then normalized along the samples by dividing all the samples with the highest sample value for that gene. The genes showing random pattern may be ignored based on statistical methods for example, serial correlation test (SCT) [24]. To reduce the computational effort involved and more meaningful ranking, SCT is used to discard the variables (genes) which show totally random pattern.

3.2 Unified Ranking Function

In this paper, the genes are ranked based on the weighted combination of the 'mean' method and a more robust distance measure called k^{th} Hausdorff distance method providing better ranking of the genes under different conditions. The number of common genes consistently occupying the first fifty slots (consistency) based on ranking in both cases (normal and tumor for example) is recorded for both ranking methods. The ranking order producing maximum consistency is selected for both ranking methods. The relative rankings are then weighted by their average consistency[1] to come up with a new ranking method for the genes. This ranking method is robust to the outliers that may exist in the data.

Such a comparison of genes under different conditions requires a suitable selection of the ranking function to find differentially expressed genes. A suitable ranking function may be of the form given by equation (1) [1].

$$f(\theta_1, \theta_2, \theta_3) = \frac{d + \theta_1}{\theta_2 * \hat{\sigma}_i + \theta_3}. \qquad [1]$$

Here, $i=1$ corresponds to normal case and $i=2$ corresponds to abnormal case, 'd'is some kind of measure of similarity or dis-similarity between different samples and 'σ_i'is the standard deviation of the samples for each variable involved in

[1] Consistency is considered reliability in ranking result provided by a particular ranking method and hence used as weight for the construction of new ranking method.

experimentation for two cases. If 'd' is the distance between means, then the ranking function is similar to t-statistic. The parameters $[\theta_1, \theta_2, \theta_3]$ are estimated by Monte Carlo simulation where they are uniformly sampled between [-1, 1]. Although the probability of occurrence of zero in the denominator is very less, statistic is ignored in case of division by zero in (1). This paper supplements the idea of using 'd' as difference between the means with 'd' as k[th] Hausdorff distance between the samples by combining them using the consistency as weight. Let 'R_1' be the ranking obtained by difference between the means and 'C_1' be the corresponding consistency and let 'R_2' be the ranking obtained by k[th] Hausdorff distance and 'C_2' be the corresponding consistency respectively then, the new ranking function is obtained by equation (2).

$$f' = \frac{C_1 * R_1 + C_2 * R_2}{C_1 + C_2}. \qquad [2]$$

Where, consistency is defined as the number of variables in common between the sets 'S_1' and 'S_2' respectively. 'S_1' and 'S_2' are the highest ranked genes obtained using equation (1) from the normal and tumor sample cases respectively. Consistency ('C') is represented in mathematical form as given in equation (3) [1].

$$C(R, N, D) = \left| S_1 \cap S_2 \right|. \qquad [3]$$

Here, 'N' is the number of highest ranked genes to be selected and 'D' is the colon microarray data used for the experimentation. The ranks 'R_1' and 'R_2' are obtained by repeating the ranking procedure given by equation (1) 10 times for both 'mean' method and Hausdorff distance method and taking the average ranking for both respectively. Each time, the algorithm is run for 1000 iterations.

There is a high possibility that differentially expressed genes fall into different clusters. Hence ASI may be used as a metric to quantify the performance of ranking algorithm. ASI algorithm also returns the relative importance of the samples used in clustering as weights [8, 16].

4. Analysis on Microarray Data

A number of empirical analyses to evaluate the performance of the proposed algorithms using both the synthetic and real microarray data have been conducted. Four hundred simulated microarray datasets [1, 17] and two real microarray data, namely, colon cancer dataset [19] and Gastric cancer dataset [25] are used for the experimentation. The ranking function of equation (1) offers highest consistency between normal and abnormal cases for various combinations of thetas. The basic assumption is that relatively small number of genes express differentially.

4.1 Simulated Microarray Data Analysis

It has been found that microarray data follow lognormal distribution [1]. The artificial microarray datasets used in this paper are based on the ideas presented in [1, 17]. The simulated datasets are generated using a hierarchical model where data in each of the classes are drawn from normal distributions with prior distributions of means being normal and variances following a gamma [1]. First 200 datasets are generated with equal variance parameters for both the cases and next 200 datasets with unequal variance parameters for both the cases. Equal and unequal variance parameters provide different nature of microarray data that may be available. With unequal variance parameters, the variance under one case may be made larger than the other case such that one case is more wide spread than the other. The idea behind using the simulated datasets is that we have access to the ground truth information about genes that are highly differentially expressed and are available for comparison with the experimental results. A large data set also allows us to compute the ROC of the proposed method. The number of false positives and number of true positives for a particular experimental result are calculated for various thresholds of number of variables selected until all the differentially expressed are identified by the ranking method. The ROC curves provide information about how a particular ranking function performed whilst providing comparison of different ranking methods.

4.1.1 Simulated Data with Equal Variance

This dataset has same variance parameters for differentially expressed genes (DEGs) and non-differentially expressed genes (NDEGs). As experimented in [1], the variance parameter was not set to zero in the ranking function. We assumed the underlying data distribution is unknown and conducted the experiments on 200 different datasets under the condition mentioned. The ranking is performed using 'mean' method [1], Hausdorff distance method [26] and adaptive ranking method as given by (2).

As proposed in [1], all the 200 datasets contained 1025 genes with 25 of them being differentially expressed. All the genes have the sample size of 10, each case having sample size of 5. From the figure 1, it is evident that 'mean' method performed better than Hausdorff distance method as the data is

essentially governed by the mean. The average consistency of the 'mean' method (54.9) is found to be greater than the Hausdorff distance method (49.87). The adaptive ranking method as given by (2) performed better than 'mean' ranking method and Hausdorff distance ranking considered alone.

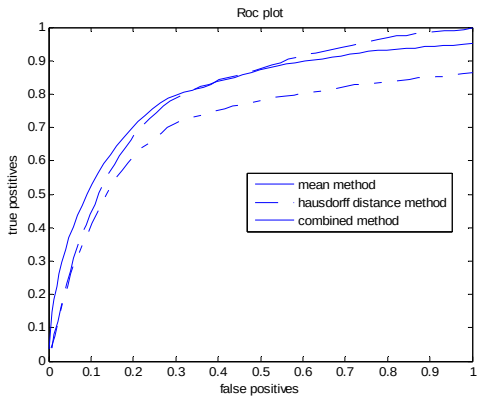

Fig. 1 Results for simulated dataset1, Roc curves showing the performance of the adaptive ranked method when compared to other ranking methods.

4.1.2 Simulated Data with Unequal Variance

The variance parameters of two cases are different for this case. The variance parameters of DEGs are made higher than that of NDEGs. This mimics the situation where the genes of interest show higher variability. Figure 2 shows the results using simulated dataset 2. From the Fig. 2, it is evident that Hausdorff distance ranking method performed better than the 'mean' ranking method. The adaptive ranking method performed better again when compared to both the ranking methods considered individually.

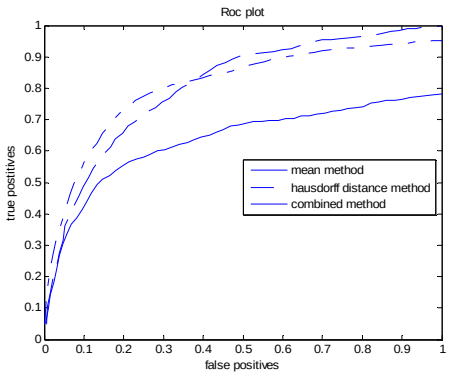

Fig. 2 Similar result as shown in Fig. 1 using artificial microarray data with unequal variances.

The average consistency of the Hausdorff distance method (55.3%) is found to be greater than the 'mean' method (52.4%), respectively. The adaptive ranking method as given by (2) however performed better than 'mean' ranking method and Hausdorff distance ranking considered alone.

4.2 Real Microarray Data Analysis

To further illustrate the utility of the proposed adaptive raking and selection of differentially expressed genes the proposed approach is used on two different microarray data sets. It is assumed that genes under different conditions fall into different clusters; they are more likely to be different than the ones falling into the same cluster.

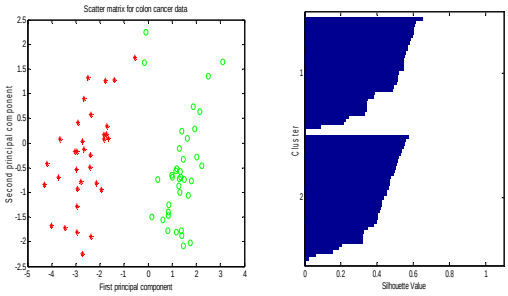

Fig. 3 Differentially expressed genes for colon cancer dataset (a) Clustering result using ASI algorithm (b) Silhouette index for the clusters formed by ASI.

4.2.1 Colon cancer dataset

Affymetrix oligonucleotide array complementary to more than 6,500 human genes are used to analyze the colon cancer tissues. The data to be ranked has 40 samples for tumor case and 22 samples for normal case.

Natural logarithm is first applied on raw colon data and then normalized to have sum of samples equal to 1. Randomly, 20 samples are chosen in both cases for experimentation and ranking method illustrated in section II is implemented. Monte Carlo method is used to estimate the parameters of the equation (1). Thousand samples are considered for each parameter (thetas) uniformly sampled in the range [-1, 1] and hence each experiment is run for 1000 iterations. The first 50 highest ranked genes are recorded every iteration separately for two cases (normal and tumor) and number of genes common in first fifty slots is used to calculate the consistency between two cases respectively as defined in equation (3).

The ordering of the genes producing highest consistency is then recorded for both ranking methods. Normalized consistencies are then used in equation (2) as C_1 and C_2 respectively and new ranking is calculated for all the genes. The first 50 highest ranked genes by the new method are selected and are estimate of highly differentiated genes.

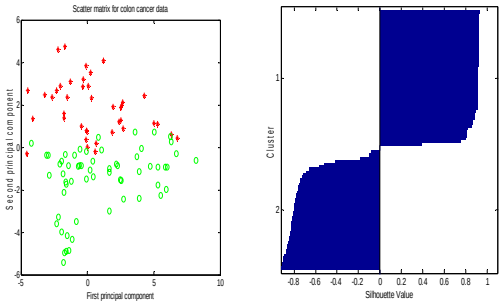

Fig. 4 similar results for gastric cancer dataset (a) Clustering result using ASI algorithm (b) Silhouette index for the clusters formed by ASI.

The ASI algorithm is then applied onto the highest selected 35 genes. From the empirical analysis it was found that these 35 are highly differentiated genes. Since the normal and tumor cases have unequal number of samples, 20 random samples from both the cases are selected for experimentation. Different classes are then projected in PCA space. Figure 3(a) shows the clustering result obtained using ASI algorithm. From the figure, it is evident that samples falling into different clusters are indeed differentially expressed. Figure 3(b) shows the silhouette index for the clusters formed. From the fig 3(b), it is evident that elements falling into the same cluster are highly similar.

4.2.2 Gastric cancer dataset

Gastric cancer is the world's second most common cause of cancer death. For the gastric cancer data, the data to be clustered has 90 samples for tumor case and 22 samples for normal case and 14 metastatic gastric cancers. Approximately 30300 genes are used to analyze these tissues.

The expectation of the experiment is to find expression of the genes correlated to patient survival, which further suggest differences in pathogenetic pathways and potential therapeutic strategies. In this case 20 samples are randomly chosen from both cases for experimentation and ranking method described in section II is implemented. Ranking parameters and number of iterations were same as that used for colon cancer data. As before, since the access to the ground truth is not available, we rely on ASI algorithm to see if the genes indeed show differential expression after ranking by adaptive ranking method. Fig. 4(a) shows the clustering result of ASI algorithm where different classes are visualized in PCA space. Fig. 4(b) shows that elements falling into the same cluster are highly similar.

5. Conclusions

A unified ranking and selection method for finding differentially expressed genes is presented in this paper. Assumption that only a few genes are differentially expressed is made. The ranking method proposed in [1] is adapted and supplemented with k^{th} Hausdorff distance measure under the assumption that 'Mean' method might be governed by the outliers in the experimental data. A variant of ASI algorithm proposed in [16] is implemented for selecting the differentially expressed genes from micro-array data. ASI algorithm is used to filter out the highly ranked genes that may not be highly differentially expressed. The adaptive ranking method proposed in this paper is applied on four hundred simulated [1, 17] and two real datasets [19, 25]. ROC curves show that adaptive ranking method performed better than the mean or Hausdorff ranking method considered alone for simulated datasets. For the real datasets, ranking of genes using adaptive ranking method is performed and then highly expressed genes are selected using ASI algorithm. Quality of clusters is analyzed using Silhouette index. From the empirical analysis, it is observed that the proposed unified approach produces high quality clusters.

Acknowledgements: Authors wish to thank Sach Mukherjee [1] for his suggestions related to simulated microarray data. This work is partially supported by the start up grants and Herff College of engineering fellowship from The University of Memphis.

References

[1] S. Mukherjee, S. J. Roberts, and M. J. v. d. Laan, "Data Adaptive test statistics for microarray data," *Bioinformatics*, 2005.

[2] J. Aubert, A. Bar-Hen, J. J. Daudin, and S. Robin, "Determination of the differentially expressed genes in microarray experiments using local FDR," *BMC bioinformatics*, vol. 5, 2004.

[3] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple

testing," *J. R. Statist. Spc.*, vol. 57, pp. 289-300, 1995.

[4] W. Duch, J. biesiada, T. Winiarski, T. Grudzinski, and K. Grabczewski, "Feature Ranking, Selection and Discretization," presented at Int. Conf. on Artificial Neural Networks (ICANN) and int. conf. on Neural Information Processing (ICONIP), 2003.

[5] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed, "Statistical methods for identifying differentially expressed genes in replicated CDNA microarray experiments," Department of Biochemistry, Stanford University August 2000.

[6] I. Guyon, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, pp. 1157-1182, 2003.

[7] W. Pan, "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 18, pp. 546-554, 2002.

[8] J. S. Shaik and M. Yeasin, "Unified Framework for Ranking and Clustering of Microarray Data Using Adaptive Sub-space Iteration," *RECOMB*, 2006.

[9] J. Biesiada, W. Duch, A. Kachel, K. Maczka, and S. Palucha, "Feature ranking methods based on information entropy with Parzen windows," *Intl. conf. on Research in Electrotechnology and Applied Informatics*, pp. 1-9, 2005.

[10] C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," *Bioinform Comput Biol.*, pp. 185-205, 2005.

[11] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, D. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, pp. 531-537, 1999.

[12] R. Kohavi and G. John, "Wrappers for feature selection," *Artifcial Intelligence*, vol. 97, pp. 273-324, 1997.

[13] X. Liu, A. Krishnan, and A. Mondry, "An Entropy based gene selection method for cancer classification using microarray data," 2005.

[14] Tusher, Tibshirani, and K.-M. Chu, "Significance analysis of microarray applied to the ionizing radiation response," *PNAS 2001*, vol. 98, pp. 5116-5121, 2001.

[15] M. Zaffalon and M. Hutter, "Robust Feature Selection by Mutual Information Distributions," *Proc. Of the 14th Intl. conf. on Uncertainty in Artificial Intelligence (UAI- 2002)*, 2002.

[16] T. Li, S. Ma, and M. Ogihara, "Document Clustering via Adaptive Subspace Iteration," presented at SIGIR, South yokshire, 2004.

[17] I. Lonnstedt and T. Speed, "Replicated MicroArray Data," *Statistica Sinica*, vol. 12, pp. 31-46, 2002.

[18] M. K. Kerr, M. Martin, and G. A. Churchill, "Analysis of variance for gene expression microarray data," *Journal of computational Biology*, vol. 7, pp. 819-837, 2000.

[19] U. Alon, N. Barkai, D. A. Notterman, K.Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc.Natl. Acad. Sci*, vol. 96, 1999.

[20] K. J. Antonellis, "Optimization of an External Standard for the Normalization of Affymetrix GeneChip Arrays," GeneLogic Inc 2002.

[21] B. Krishnapuram, L. Carin, and A. J. Hartemink, "Joint classifier and feature optimization for comprehensive cancer diagnosis using gene expression data," *J. comput. Biol.*, vol. 11, pp. 227- 242, 2004.

[22] Y. Zhao and G. Karypis, "Evaluation of hierarchical clustering algorithms for document datasets," Dept. Of comp. Sc., University of Minnesota.

[23] T. G. Dietterich, "Approximate statistical test for comparing supervised classifcation learning algorithms," *Neural Computation*, vol. 10, pp. 1895-1924, 1998.

[24] G. K. Kanji, *100 statistical tests*. New Delhi: SAGE Publications, 1999.

[25] X. Chen, S. Y. Leung, S. T. Yuen, K.-M. Chu, J. Ji, R. Li, A. S. Y. Chan, S. Law, O. G. Troyanskaya, J. Wong, S. So, D. Botstein, and P. O. Brown, "Variation in Gene Expression Patterns in Human Gastric Cancers," *Mol Biol Cell*, vol. 14, pp. 3208–3215, 2003.

[26] D. Vignon, B. C. Lovell, and R. J. Andrews, "General Purpose Real-Time Object Tracking Using Hausdorff transforms," presented at IPMU2002 Special Session on Intelligent Systems for Video Processing.