GENE SELECTION IN MICROARRAY DATA ANALYSIS FOR BRAIN CANCER CLASSIFICATION

Y. Y. Leung, C. Q. Chang, Y. S. Hung, P.C.W. Fung

Department of Electrical and Electronic Engineering, the University of Hong Kong, Pokfulam Road, Hong Kong
yyleung@eee.hku.hk, cqchang@eee.hku.hk, yshung@eee.hku.hk, hrspfcw@hku.hk

ABSTRACT

Cancer classification has been one of the most challenging tasks in clinical diagnosis. At present cancer classification is done mainly by looking through the cells' morphological differences, which do not always give a clear distinction of cancer subtypes. Unfortunately, this may have a significant impact on the final outcome of whether a patient could be cured effectively. Microarray technology can play an important role on diagnosing which type of disease one is carrying. The gene selection process is critical for developing gene markers for faster and more accurate diagnosis. In this paper, we develop a method using pairwise data comparisons instead of the one-over-the-rest approach used nowadays. Results are evaluated using available clustering techniques including hierarchical clustering and k-means clustering. Using pairwise comparison, the best accuracy achieved is 95% while it is only 83% when using one-over-the-rest approach.

1. INTRODUCTION

Instead of classifying cancers based on microscopic histology and tumor morphology, the introduction of microarray technology significantly improves the discovery rates of different types of cancers through monitoring thousands of gene expressions in a parallel, rapid and efficient manner. Several studies have been successful [1-4] in differentiating various cancer cell-types.

Genes present in different cells in our human body are responsible for carrying out unique functions at their specific locations. The problem is out of the 25,000 genes present in the human genome, how we can identify something representative of the brain [5] is a great challenge. The answer lies in the essence of genes selection. Gene expression patterns are useful for classification, diagnosis and understanding of diseases [6]. In the case of classifying microarray data which often contains far more genes than samples, we need to first select a small set of informative genes that can effectively discriminate samples into different classes. While most of the test-statistics are developed for the use of two classes, in the case of multiple classes which we are focusing on, we propose a gene selection procedure based on pairwise testing, although the one-versus-the-rest testing procedure is most often employed instead. The accuracy is greatly improved when using pairwise testing for we use all possible combinations of groups available in the dataset. Details of the two gene selection algorithms will be presented in the following section.

2. DATASET AND METHODS

2.1. Descriptions on Dataset

Brain cancer is chosen as a test case. Little is known about the biology of brain cancer and quite often there is controversy on diagnosis solely based on the cells' morphological differences. Classifications till now are based on the tumors' originalities (cell types) but not their locations. Tumors can develop in any types of cells, and that's why classification of brain cancer is so difficult [7]. A number of genes that may be involved in glial tumorigenesis [8] and in prediction of glioblastoma survival [9] have been identified recently. The dataset we use is obtained from the website [10] which contains 92 brain cancer samples grouped into 5 classes.

2.2. Two-group comparison approaches - pairwise versus one-over-the-rest algorithm

Our approach to select genes is to rank them by their discrimination power and select those that are most discriminative. Signal-to-noise score (SNR) is chosen both as our statistic and also in the paper. The conventional one-versus-the-rest multiple comparison method compares each data subgroup with the rest of the data and selects representative genes for each subgroup, as in [11-12]. To illustrate this, we have 5 groups of samples (g1-g5). Five comparisons are made: g1 versus g2-g5; g2 versus g1, g3-g5; g3 versus g1-g2, g4-g5; g4 versus g1-g3, g5; and g5 versus g1-g4. 10 genes are selected from each comparison so we have 50 in total [2, 10]. The problem with this approach is that it cannot find genes that have dissimilar expression profiles between the single group and each of the groups in the other group. Here we propose an alternative approach using pairwise testing. The proposed method involves using SNR across the pairs of groups one by one. SNR is performed between any two groups. Using the same illustration, representative genes on g1 can be found by running the SNR between g1 and g2 first. 5 genes, which attain the largest p-values out of all, are selected. Similarly, this SNR is run between g1 and each of the remaining groups (i.e. between g1 and g3; g1 and g4; g1 and g5). As a result, 20 genes altogether are selected to represent the group g1. After iterative computation, 50 different genes are selected for all five groups.

3. RESULTS

A total of 50 genes are selected using each of the two methods described in Section 2. Verification on whether the selected genes

This work is supported in part by Hong Kong RGC grant under HKU7180/03E.

can be used to properly classify the samples is done by clustering. Hierarchical and k-means clustering are chosen.

3.1. Hierarchical clustering results comparison

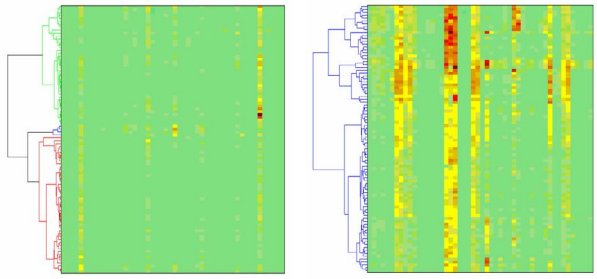

Fig. 1a. (Left) Hierarchical clustering on genes selected by one-over-the-rest method; **Fig. 1b.** (Right) Hierarchical clustering on genes selected by pairwise comparison method.

As shown from Fig. 1a, rows represent samples and columns represent genes. Very little difference can be observed across the expression levels of the genes selected by the one-versus-the-rest multiple comparison method, by which only 76 out of 92 samples are classified correctly into the 5 groups. Fig. 1b shows that the genes selected by our pairwise SNR method give much better differentiation of the samples from different classes. Using our pairwise method, only 4 samples are misclassified.

3.2. k-means clustering results comparison

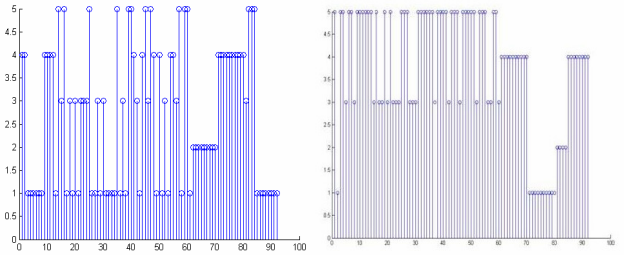

Fig. 2a. (Left) k-means clustering on genes selected by one-over-the-rest method; **Fig. 2b.** (Right) k-means clustering on genes selected by pairwise comparison method.

Results of k-means clustering are shown in Fig. 2. Samples are grouped in priori to their predefined groups before doing clustering. The x-axis represents the sample identification number while y-axis represents the defined group (numbered 1 to 5). Using the one-versus-the-rest multiple comparison method, only 65 out of 92 samples are classified correctly into the 5 groups. In Fig. 2b, 72 of them are classified accordingly using pairwise approach.

Results of the two clustering techniques applied to the expressions of genes selected by the one-versus-the-rest and the pairwise testing methods are summarized in Table 1.

Table 1. Summary on accuracies of different clustering algorithms based on two comparison methods

	one-versus-the-rest	pairwise
Hierarchical clustering	83%	95%
k-means clustering	71%	78%

4. DISCUSSIONS & CONCLUSIONS

Accuracy of pathological diagnoses on patients can be improved by the technique of microarray analysis. No doubt gene selection process lies in the heart of this technique. The results given in Table 1 show that our proposed pairwise comparison approach used in multiple classes classification outperforms the original one-over-the-rest method irrespective of whether classification is performed using hierarchical or k-means clustering. Genes cluster together are those with common expression profiles which means they may share regulatory pathways, and further studies into these may clarify the mechanisms of this common co-regulation [13]. Unknown gene functions and regulations can then be inferred more efficiently using standard molecular approaches.

5. REFERENCES

[1] P.J. French, S.M.A. Swagemakers, J.H.A. Nagel, M.C.K. Kouwenhoven, E. Brouwer, P. Spek, T.M. Luider, J.M. Kros, M.J. Bent and P.A.S Smitt, "Gene Expression Profiles Associated with Treatment Response in Oligodendrogliomas," *Cancer Research*, vol. 65, pp. 11335-11344, 2005

[2] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J. 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

[3] J. Khan, J.S. Wei, M. Ringnér, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson and P.S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, pp.673-679, 2001

[4] G.J. Gordon, R.V. Jensen, L.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*, vol. 62, pp. 4963-4967, 2002

[5] D.J. Duggan, M. Bittner, Y. Chen, P. Meltzer and J.M. Trent, "Expression profiling using cDNA microarrays," *Nature Genetics* , vol. 21, pp. 10-14, 1999

[6] N. Dhiman, R. Bonilla, D.J. O'Kane and G.A. Poland, "Gene expression microarrays: a 21st century tool for directed vaccine design," *Vaccine*, vol. 20, pp. 22-30, 2001

[7] P.S. Mischel, T.F. Cloughesy and S.F. Nelson, "DNA-Microarray Analysis of Brain Cancer: Molecular classification for Therapy," *Nature Reviews Neuroscience*, vol. 5, pp. 782-792, 2004

[8] Y. Liang, M. Diehn, N. Watson, A.W. Bollen, K.D. Aldape, M.K. Nicholas, K.R. Lamborn, M.S. Berger, D. Botstein, P.O. Brown and M.A. Israel, "Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme," *Proc Natl Acad Sci US A* , vol. 102, pp. 5814-5819, 2005

[9] J.N. Rich, C. Hans, B. Jones, E.S. Iversen, R.E. McLendon, B.K.A. Rasheed, A. Dobra, H.K. Dressman, D.D. Bigner, J.R. Nevins and M. West, "Gene Expression Profiling and Genetic Markers in Glioblastoma Survival," *Cancer Research*, vol. 65, pp. 4051-4058, 2005

[10] http://www.genome.wi.mit.edu/MPR/CNS

[11] J.W. Lee, L.B. Jung, M. Park and S.K. Song, "An extensive comparison of recent classification tools applied to microarray data," *Computational Statistics & Data Analysis*, vol. 48, pp. 869 – 885, 2005

[12] R. Díaz-Uriarte and S.A. Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol.7, 2006

[13] S. Ramaswamy and T.R. Golub, "DNA microarrays in clinical oncology," *Journal of Clinical Oncology*, vol. 20, pp. 1932-1941, 2002