Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Gene selection for Brain Cancer Classification

Y. Y. Leung[a], C. Q. Chang[a], Y. S. Hung[a], P. C. W. Fung[a,b]
[a]Department of Electrical and Electronic Engineering, [b]Department of Medicine,
The University of Hong Kong, Pokfulam Road, Hong Kong

***Abstract*—With the introduction of microarray, cancer classification, diagnosis and prediction are made more accurate and effective. However, the final outcome of the data analyses very much depend on the huge number of genes with relatively small number of samples present in each experiment. It is thus crucial to select relevant genes to be used for future specific cancer markers. Many feature selection methods have been proposed but none is able to classify all kinds of microarray data accurately, especially on those multi-class datasets. We propose a one-versus-one comparison method for selecting discriminatory features instead of performing the statistical test in a one-versus-all manner. Brain cancer is chosen as an example. Here, 3 types of statistics are used: signal-to-noise ratio (SNR), *t*-statistics and Pearson correlation coefficient. Results are verified by performing hierarchical and *k*-means clustering. Using our one-versus-one comparisons, best performance accuracies of 90.48% and 97.62% can be obtained by hierarchical and *k*-means clustering respectively. However best performance accuracies of 88.10% and 80.95% can be obtained respectively when using one-versus-all comparison. This shows that one-versus-one comparison is superior.**

***Index Terms*—Biomedical signal processing**

I. Introduction

Microscopic histology and tumor morphology are used to be important criteria for classifying different cancers. Microarray has emerged as one of the most potential tools for assisting clinicians to diagnose the disease since the last decade. Gene expression profiles may offer more information on how to classify cancer samples accurately. This can be used not only for prediction, but also for diagnosis, understanding and prognosis of disease [1]-[2].

Microarrays have successfully been applied to differentiate between unknown types of cancers in a parallel, rapid and efficient manner [3]-[5]. Here we focus on finding genes that may contribute to the development of brain cancer. Brain structure is the most complex inside our body for it expresses the highest proportion of the genome [6]. Classifications of cancer are based on the tumors' originalities but not their locations. Tumors can develop wherever they like in any types of cells and this contribute mainly as to why classification of brain cancer is so difficult [7]. Recent studies have identified some genetic markers in glioblastoma survival [8]. Genes present in different cells in our 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 to identify those genes that are representative of the brain [4]. The answer lies in the essence of gene selection [9].

In the following sections, the proposed one-versus-one comparison method as well as the original one-versus-all comparison method will be described first, followed by basic description of the statistics and clustering tools used. Results obtained from hierarchical and *k*-means clustering are given. Performance is measured based on the accuracy of clustering samples in classes using the selected genes.

II. Information on Datasets

The dataset used in this study is obtained from the website http://www.broad.mit.edu/mpr/publications/projects/CNS [10] which contains 92 brain cancer expression profiles consisting of 7129 genes using an Affymetrix oligonucleotide array. These samples are grouped into 6 classes: 46 samples of classic medulloblastoma (CMD); 14 of desmoplastic medulloblastoma (DMD); 10 of malignant gliomas (MG); 10 of atypical teratoid/rhabdoid tumours (AR); 4 of normal cerebellum (NC) and 8 of supratentorial primitive neuroectodermal tumours (PN). Ideally we should include all cancer and non-cancer subtypes in the dataset for classification. After performing some preliminary studies, we found that the expression levels of the NC group varies significantly from the other cancer-related sub-groups and these 4 samples can often be correctly classified into one single group. We exclude this NC group first and are left with 5 groups where only the sample size of the group CMD is much greater than the others. Due to statistical reasons, we leave this group out also for the huge sample size will affect the classification accuracy of the remaining samples. Consequently we will be distinguishing the 4 subtypes of cancer where two of them, when classified solely by morphological characteristics, are in controversy of whether they belong to the same group or not [8].

III. Methodologies Used

In general we have far more genes than samples in microarray datasets, for the purpose of performance and clinical benefits, we should first select a small set of informative genes that can effectively discriminate samples in different classes before performing classification. Most gene selection methods are developed for the case of 2 classes,

Manuscript received April 24, 2006. This work was supported in part by Hong Kong RGC grant under HKU7180/03E.
E-mail: {yyleung, cqchang, yshung}@)eee.hku.hk, hrspfcw@hkucc.hku.hk

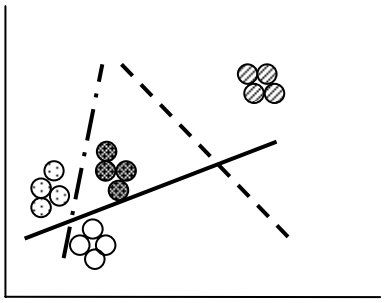

Fig. 1. Diagram representation on the limitation of one-over-all comparison in multi-class datasets.

while in this paper we emphasize on gene selection for multi-class discrimination. For ease of comparison, the samples are grouped accordingly to the predefined groups before carrying out any of the analyses.

A. One-versus-all method

The one-versus-all approach [11]-[12] divides the classes into 2 groups each time, with one group consisting of a single class and the other group consisting of samples in all the other classes. Then, a 2-class gene selection method can be applied to find a set of informative genes. In our study, if we take DMD as one group, the other group will contain the remaining samples from MG, AR and PN. Genes representative of the group DMD with respect to the others are selected by applying a 2-class gene ranking method to these 2 groups. This process is performed 4 times and each time 7 genes are selected to represent one of the 4 classes DMD, MG, AR and PN.

B. One-versus-one method

The problem with the one-versus-all 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. The diagram below illustrates the drawback of using the one-versus-all algorithm. Fig. 1 shows an example of the projected 2-dimensional scatter plot of the samples divided into 4 classes labeled with different methods. If we perform the one-versus-all comparison between line-shadowed cluster and the remaining ones, the 2 groups are well separated and no doubt we can select a set of genes that discriminate well the line-shaded cluster from the group of remaining classes. The genes selected for the line-shaded cluster is representative of the distinction between this cluster and the others. Next, consider applying one-versus-all comparison to the dotted cluster. This will identify genes discriminating the dotted cluster from the average of the other 3. Since the line-shaded cluster is well separated from the others, the difference between the dotted cluster and the average of the other 3 clusters is dominated by this line-shaded cluster. Hence the genes selected will tend to discriminate between the dotted cluster and the line-shaded cluster. The same argument applies for the black and white clusters. This means that all 4 one-versus-all comparisons end up selecting genes representing the distinction between the line-shaded cluster and the other 3 clusters, and no selected genes can discriminate between the 3 closely located clusters.

To remedy the situation, we use one-versus-one comparison instead. This approach involves performing gene selection for each of the pairs of classes available in the dataset. This ensures that each of these groups is compared with each of the remaining groups one by one, and the most significant differences can be represented by the corresponding selected genes. As in Fig. 1, since the 4 clusters are well separated from each other, there exists at least one gene to well discriminate between each pair of clusters, and such genes can indeed be selected by the one-versus-one comparison approach. Therefore the selected genes can more effectively discriminate the 4 clusters.

For our study, genes that are able to discriminate group DMD from the others can first be found by performing the statistics 3 times, each between DMD and one of the remaining groups (i.e. first time with MD, second time with AR and third time with PN). 5 genes are selected in each one-versus-one comparison. A total of 15 genes are representative of the DMD group. The procedure is repeated for all 4 groups and due to overlapping, a total of 30 genes are chosen finally to distinguish these groups. Note that the number of genes included for one-versus-all and one-versus-one comparison is different. 24 or 48 genes should be chosen in each method for better comparison purpose; however, the number of genes is either too small or too large for classification purposes. The minor difference in number of genes selected will not pay much effect to the final accuracy.

C. Gene ranking metrics and clustering methods

As mentioned above, different feature selection statistics are employed in both comparison methods which are used to determine whether the genes act variably across samples [13]. Here we describe 3 such metrics. Assume that there are 2 classes within a large dataset. All expression values are normalized to mean zero and have variance equal to one [14]. Any types of ranking metrics can be applied for all samples are of normal distribution. If we want to determine whether gene one in the dataset is useful for discriminating between the 2 classes, we can apply either of the statistics below, where X and Y are the gene expression datasets for each of the 2 classes for comparison, $\overline{X}$ and $\overline{Y}$ are the sample means, $\sigma(X)$ and $\sigma(Y)$ are the standard deviations and N is the total number of genes. Please refer to [14]-[15] for details.

- *t*-statistics (TS):

$$t = (\overline{X} - \overline{Y}) / \sqrt{\sigma^2(X)/N + \sigma^2(Y)/N} \quad (1)$$

- Signal-to-noise Ratio(SNR) score

$$SNR = (\overline{X} - \overline{Y}) / (\sigma(X) + \sigma(Y) \quad (2)$$

- Pearson Correlation Coefficient (PC)

$$r_{Pearson} = \frac{\sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N} (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}} \quad (3)$$

After the potential genes are chosen using one of the above three ranking metrics and one of the two comparison approaches (one-versus-one and one-versus-all), clustering tools using these selected genes as features will then be applied to the microarray datasets for verification. Two

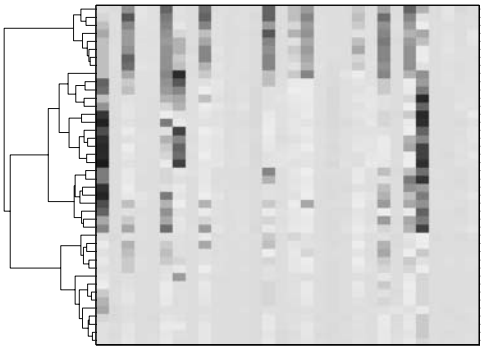

Fig. 2. Best clustergram obtained using one-versus-one comparison method, euclidean as distance metric, ward as linkage metric, *t*-statistics as gene selection criterion. The performance accuracy is

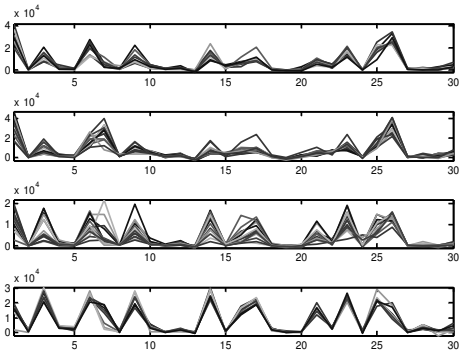

Fig. 3. Best expression profiles obtained using one-versus-one comparison method, euclidean as distance metric, ward as linkage metric, *t*-statistics as gene selection criterion.

popular clustering tools are used in our study: the hierarchical and *k*-means clustering. Please refer to [16]-[18] for details. The main parameter for both clustering methods is the distance measure. For hierarchical clustering we also need to specify the type of linkage. Each of these differs in the way in which distances are calculated between the growing clusters and the remaining members of the dataset, and average linkage is suggested to be used in general [19]. However our analyses show that ward linkage is the best linkage to be used [20].

IV. Results

In order to determine whether the selected genes are discriminative enough, we verify the results using 2 clustering techniques. Performance accuracy is determined by the number of correctly classified samples over the total number of samples used for analysis.

A. Verification using Hierarchical Clustering

Hierarchical clustering is first done on these selected genes using different comparison method and gene selection statistics mentioned above. The table below summarizes the results:

TABLE I.
HIERARCHICAL CLUSTERING PERFORMANCE ACCURACY RESULTS BY DIFFERENT TYPES OF FEATURE SELECTION STATISTICS, USING EUCLIDEAN AS DISTANCE MEASURE, WARD AS LINKAGE METRIC.

	One-versus-one	One-over-the-rest
SNR	85.71 %	76.19 %
TS	95.24 %	69.05 %
PC	85.71 %	88.10 %

The performance accuracies of the one-versus-one comparison are higher in 2 of the 3 feature selection statistics chosen. Genes selected by *t*-statistics using one-versus-one comparison method seem to classify the samples best. Genes selected by the SNR score and Pearson coefficient are comparable in both methods. Fig. 2 shows the best clustergram on genes obtained. Here rows represent samples; columns represent the 30 chosen genes. The performance accuracy is 95.24%. Darker color indicates that the gene is over-expressed[21]. 4 groups can be identified by drawing 3 horizontal lines across the clustergram. The first cluster has eight samples (reading Fig. 2 from top) which represents group PN. The second cluster represents group MG of which 2 extra samples from group AR are found. The third cluster represents group AR and 2 samples inside this original group are missing. The fourth cluster represents group DMD and all are classified accordingly.

In order to visualize the gene expression levels as patterns, Fig. 3 shows the expression profiles across all samples on these 30 selected genes using one-versus-one comparison method, *t*-statistics, Euclidean as distance and ward as linkage. X-axis represents the sample number while y-axis represents the gene expression values. Each subplot represents the expression profile of each group. The first represents group PN, second corresponds to group AR, the third represents group MG and the remaining represents group DMD. Peaks and valleys are at different locations across samples, meaning all 4 groups have different expression profiles of different magnitudes. Some genes are excited at some particular groups but not the others.

B. Verification using k-means clustering

k-means clustering is subsequently done on these selected genes using different comparison method and gene selection statistics mentioned above. Table II summarizes the results:

TABLE II
K-MEANS CLUSTERING PERFORMANCE ACCURACY RESULTS BY DIFFERENT TYPES OF FEATURE SELECTION STATISTICS, USING CITYBLOCK DISTANCE AS DISTANCE METRIC.

	One-versus-one	One-over-the-rest
SNR	85.71 %	64.29 %
TS	97.62 %	80.95 %
PC	76.19 %	66.67 %

The performance accuracies of the one-versus-one comparison are better for all the statistics chosen. Best performance accuracy is obtained from genes selected by *t*-statistics using either one-versus-one or one-versus-all comparison method. Yet genes selected by correlation coefficient are on average the worst discriminative.

Fig. 4 shows the stem-plot on genes selected by *t*-statistics using one-versus-one comparison method and cityblock as distance measure. X-axis represents the number of samples where y-axis represents the group number in which one

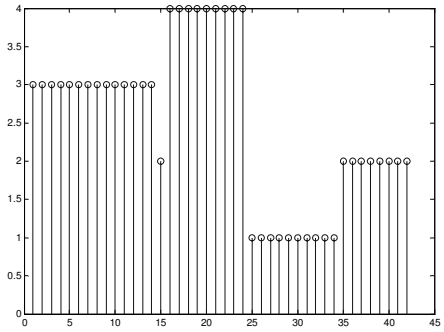

Fig. 4. Best *k*-means clustering obtained using One-versus-one comparison method, Cityblock distance as distance metric, *t*-statistics as gene selection criterion. The performance accuracy is 97.62%.

sample is clustered into. As the samples are grouped according into their predefined classes before analyses, the ideal case involves 14 ones at the far left hand part of the stem-plot representing group DMD, followed by 10 consecutive twos representing group MG, 10 consecutive threes and 8 consecutive fours representing group AR and PN respectively. These ones, twos, threes, and fours are arbitrary numbers for the group numbers. Once the correct number of samples attained a certain integer value (between 1 and 4) is lined up consecutively, the pattern can be treated as an ideal case. As can be seen in Fig. 4, 14 samples of group DMD are aligned on the far left, followed by group MG and AR, each contains 10 samples. The last 8 samples are from group PN. Only one sample originated from group 2 (group MG) is misclassified into group 4 (PN). The overall performance accuracy is thus 97.62 %.

V. DISCUSSION

We have demonstrated that the choice of feature selection statistics, comparison method between groups and clustering types can all affect the interpretation of final results of microarray data. With regard to which feature selection statistics performs better, genes selected by simple *t*-statistics seem to better classify all the samples with an accuracy of over 90% in one-versus-one comparison cases.

According to Tables I and II, one-versus-one comparison method performs better or equally well in these cases experimented. This shows that when dealing with multi-class cancer subtype datasets, one-versus-one comparison method can give better performance accuracy than the commonly used one-versus-all algorithm. The former looks into each of the within pairs differences but the latter emphasizes on how one group can be distinguished from the other remaining groups.

VI. CONCLUSION

Gene selection process lies in the heart of microarray analysis. Results show that our proposed one-versus-one comparison method, in most of the cases, outperforms the original one-versus-all comparison method irrespective of whether hierarchical or *k*-means clustering is used. These serve as the basis for further investigations into the real relationships hidden inside these gene expression datasets [19].

REFERENCES

[1] 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, no. 1-2, pp. 22-30, Oct. 2001.

[2] P.J. French *et al*., "Gene Expression Profiles Associated with Treatment Response in Oligodendrogliomas," *Cancer Research*, vol. 65, pp. 11335-11344, 2005.

[3] T.R. Golub *et al*., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, no. 5439, pp. 531-537, Oct. 1999.

[4] D.J. Duggan *et al*., "Expression profiling using cDNA microarrays," *Nature Genetics*, vol. 21, pp. 10-14, 1999.

[5] S. Ramaswamy S and T.R. Golub, "DNA microarrays in Clinical Oncology," *Journal of Clinical Oncology*, vol. 20, no. 7, pp. 1932-1941, April. 2002.

[6] S.J. Watson SJ, F. Meng, R.C. Thompson and H. Akil, "The 'Chip' as a Specific Genetic Tool," *Biol Psychiatry,* vol. 48, pp. 1147-1156, 2000.

[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, no. 10, pp. 782-792, Oct. 2004.

[8] J.N. Rich *et al*., "Gene Expression Profiling and Genetic Markers in Glioblastoma Survival," *Cancer Research*, vol. 65, pp. 4051-4058, 2005.

[9] Tamayo P and Ramaswamy S, "Cancer Genomics and Molecular Pattern Recognition, in *Expression profiling of human tumors: diagnostic and research applications*, Humana Press, 2003.

[10] S.L. Pomeroy *et al*., "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436-442, 2002.

[11] J.W. Lee, B. J. Lee, 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, no. 3, 2006.

[13] L.J. Heyer, S. Kruglyak and S. Yooseph, "Exploring Expression Data: Identification and Analysis of Coexpressed Genes," *Genome Research*, vol. 9, no. 11. pp. 1106-1114, Nov. 1999.

[14] H. Liu, J. Li and L.A. Wong, "Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns," *Genome Informatics*, vol. 13, pp. 51-60, 2002.

[15] T.D. Wu, "Analysing gene expression data from DNA microarrays to identify candidate genes," *Journal of Pathology*, vol. 195, no. 1, pp. 53-65, Sep. 2001.

[16] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. USA*, vol. 95, no. 25, pp. 4863–14868, Dec. 1998.

[17] A. Brazma and J. Vilo, "Gene expression data analysis," *FEBS Letters*, vol. 480, pp. 17-24, June. 2000.

[18] T. Jirapech-Umpai and S. Aitken, "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes," *BMC Bioinformatics*, vol. 6, pp. 148-158, June. 2005.

[19] J. Quackenbush, "Computational Analysis of Microarray Data," *Nature Reviews Genetics*, vol. 2, no. 418, pp. 418-427, June. 2001.

[20] Y.Y. Leung, C.Q. Chang and Y.S. Hung, "Microarray Data Analysis for Acute Leukemia Classification – Is Gene Selection the Only Factor? (Presented Conference Paper style)," presented at the RIUPEEE, Macau, July 13–14, 2006.

[21] P. F. Macgregor and Squire JA, "Application of Microarrays to the Analysis of Gene Expression in Cancer, "*Clinical Chemistry*, vol. 48, no.4, pp. 1170-1177, 2002.