Multi-gene linear separability of gene expression data in linear time

Md. Shafiul Alam, Satish Panigrahi, Puspal Bhabak, and Asish Mukhopadhyay*

School of Computer Science, University of Windsor, 401 Sunset Avenue, Windsor, ON, N9B 3P4, Canada
{alam9, panigra, bhabak, asishm}@uwindsor.ca

Abstract. In [9] Unger and Chor showed how to test for linear separability of gene expression data with respect to pairs of genes. Their method however is not amenable to an efficient test when more than 2 genes are involved. The main contribution of this note is to show how to use linear programming to check for linear separability of gene expression data with respect to any number of genes in $O(n)$ time where n is the sample size. The hidden constant in the $O(n)$ term depends exponentially on the number of genes (the dimensionality of the problem). So, this makes for an efficient test when the number of genes is a small constant. To test the effectiveness of our algorithm, as an initial step, we have implemented this algorithm for gene pairs and are working on extending this implementation to larger groups of genes.

Key words: gene expression analysis, linear separability, tissue classification, cancer diagnosis

1 Introduction

According to Ben-Dor *et al.* [2], the correct diagnosis of a cancer type is often crucial to a successful treatment. As normal cells can evolve into cancerous cells through mutations in genes, it is believed that the gene expression data can be exploited for more effective diagnosis and treatment of cancer. For this, it is necessary to identify groups of genes that play important roles in various types of cancers. Once the genes are identified, it is possible to diagnose the presence of or the type of a cancer and determine the course of treatment [6].

In [9] Unger and Chor showed how to test for linear separability of gene expression data with respect to pairs of genes. Interestingly enough, they were able to show that 7 out of the 10 datasets of two types of cancerous cells that they studied are linearly separable with respect to a pair of genes. From this they concluded that this linear separability of gene expression datasets is strongly correlated to some underlying molecular mechanism of these gene pairs. Their

* Research Supported by an NSERC Individual Discovery Grant to this author

method of linear separability, applicable to a pair of genes only, checks for separability incrementally. When the dataset is linearly separable its running time is in $O(n^2)$, where n is the sample size.

However, checking just pairs of genes for classification may not be enough. Indeed, van't Veer *et al.* [10] found that 231 genes are significantly related to breast cancer. Among these, they identified 70 genes as optimal marker genes for classification of breast cancer for prognosis purposes. Khan *et al.* [5] found 96 genes to classify small, round, blue-cell cancers. Ben-Dor *et al.* [2] used 173-4,375 genes to classify various cancers. Golub *et al.* [3] selected 50 genes to classify leukemias. Some researchers used far more genes to classify cancers. For example, Alon *et al.* [1] used 2,000 genes to classify colon cancers.

A major bottleneck with any classification scheme based on gene expression data is that while the number of samples are limited, numbering in hundreds, the feature space is much bigger, running into tens of thousands of genes. Using too many genes as classifiers will result in over fitting, while on the other hand using too few may result in under fitting. The common consensus is that genes numbering between 10 and 50 genes may be sufficient for good classification [3, 6]. Note that even using the minimum number of 10 genes for linear classification of 100 samples that involve 20,000 genes is also an enormous task. This calls for a very efficient algorithm with an incremental feature that allows for early termination if the samples are not separable for some combination of 10 genes.

The main contribution of this note is to show how to use the linear programming algorithm of [7, 8] to check for linear separability of gene expression data with respect to any number of genes in $O(n)$ time where n is the sample size. The hidden constant in $O(n)$ depends exponentially on the number of genes. Thus, the test is efficient when the number of genes is reasonably small. To test the effectiveness of our algorithm, as an initial step, we have implemented this algorithm for gene pairs and are working on extending this implementation to larger groups of genes.

2 Multi-gene Linear Separability

We have a set of n $(= m_1 + m_2)$ samples, m_1 from a cancer type C_1 and m_2 from a cancer type C_2 (for example, m_1 from ALL and m_2 from AML) in a d-dimensional space, where d is the size of the gene set that is being tested as a linear classifier. If there exists a hyperplane that separates the m_1 samples of C_1 from the m_2 samples of C_2 then the considered group of d-genes is a linear separator.

We reformulate the original separability problem in primal space as a linear program in dual space. This involves mapping each sample point to a hyperplane in dual space. Suppose there is a separating hyperplane in primal space and, say, the sample points of C_1 are above this plane, while the sample points of C_2 are below. The dual mapping preserves this above-below relationship in the sense that if a sample point p lies above (below) a hyperplane H in the primal space then the dual of p, viz. the hyperplane p^*, lies below (above) the dual of H, viz. the point H^*, in the dual space. One such above-below preserving dual map from the primal plane (x, y) to the dual plane (u, v) is:

$$\begin{gathered} p = (p_x, p_y) \rightarrow p^* : v = p_x u - p_y \\ l : y = l_u x - l_v \rightarrow l^* = (l_u, l_v) \end{gathered}$$

Thus, there is a separating hyperplane in primal space if the resulting linear program in dual space has a feasible solution (see Fig. 1 for the $2d$ case). Note that we will have to solve 2 linear programs since it is not known if the m_1 samples of C_1 lie above or below the separating hyperplane H. Formally one of these linear programs in the dual space $(u_1, u_2, \ldots, u_n)$ is shown below:

$$p_1^i u_1 + ... + p_{n-1}^i u_{n-1} - u_n - p_n^i < 0, i = 1, ..., m_1 \tag{1}$$

$$p_1^{'i} u_1 + ... + p_{n-1}^{'i} u_{n-1} - u_n - p_n^{'i} > 0, i = 1, ..., m_2 \tag{2}$$

where $(p_1^i, p_2^i, \ldots, p_n^i)$ is the i-th sample point from C_1, and the first set of inequalities express the conditions that these sample points are above the separating plane, while the second set of inequalities express the conditions that the sample points $(p_1^{'i}, p_2^{'i}, \ldots, p_n^{'i})$ from C_2 are below this plane.

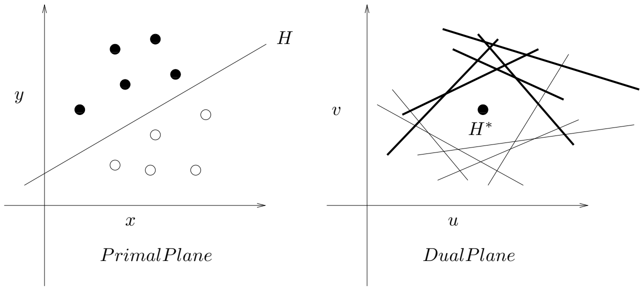

Fig. 1. *A separating line in primal space is a feasible solution in dual space*

3 Experimental Results

To test out our ideas, as a preliminary step, we have implemented our algorithm in 2-dimensions. This means testing pairs of genes as classifier. We ran our program on a Dell Inspiron laptop with an Intel Core2 Duo processor on the publicly available data set of Golub [3] with ALL and AML data testing for separability with respect to all pairs of 12,582 genes for a total of 79,147,071 separability tests. Out of these 249,567 pairs proved to be separable, representing just 0.32% of the total. The total running time was 5.07 hrs, which makes the case for having a very efficient separability test.

4 Conclusions

We believe that gene groups of size greater than 2 might be better classifiers. We are working on extending our algorithm to cover these cases. We are also going to test our algorithm on all data sets that have been tried by Unger and Chor [9] to see if our findings are consistent with their conclusions. The main contribution of our paper is to point out that separability tests can be carried out efficiently for groups of genes larger than 2.

Acknowledgements. Thanks to Dr. Alioune Ngom of the University of Windsor for bringing this problem to our attention.

References

1. Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotidearrays. *Proc. Natl. Acad. Sci.*, USA, 96, 6745-6750, 1999.
2. Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, Z. Tissue classification with gene expression profiles. In *Journal of Computational Biology*, 7(3-4),559-583, 2000.
3. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531-537, October 1999.
4. Gordon, G. J., Jensen, R. V., Hsiao, L-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J. and Bueno, R. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, 62(17):4963-4967, Sep 2002.
5. Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C. and Meltzer, P. S. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7(6):673-679, June 2001.
6. Kim, S., Dougherty, E. R., Barrera, J., Chen, Y., Bittner, M. L. and Trent, J. M. Strong feature sets from small samples. *Journal of Computational Biology*, 9(1), 127-146, 2002.
7. Megiddo, N. Linear-time algorithms for linear programming in R^3 and related problems. In *SIAM Journal of Computing*, 12(4), 759-776, 1983.
8. Megiddo, N. Linear Programming in Linear Time When the Dimension is Fixed. In *Journal of the ACM*, 31(1), 114-127, 1984.
9. Unger, G. and Chor B. (2007). Linear Separability of Gene Expression Datasets. To appear in *IEEE Transactions on Computational Biology and Bioinformatics.*
10. van't Veer, L. J., Dal, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. and Friend, S. H. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871), 530-536, Jan 2002.