International Journal of Computer Applications Technology and Research
Volume 2– Issue 5, 609 - 613, 2013, ISSN: 2319–8656

Gene Selection for Cancer Classification using Microarrays

T.Shanmugavadivu,
Karpagam University,
Coimbatore-21

T.Ravichandran
Hindusthan Institute of Technology,
Coimbatore-32.

Abstract Microarrays allow biologists to better understand the interactions between diverse pathologic states at the gene level. However, the amount of data generated by these tools becomes problematic. New techniques are then needed in order to extract valuable information about gene activity in sensitive processes like tumor cells proliferation and metastasis activity. Recent tools that analyze microarray expression data have exploited correlation-based approach such as clustering analysis. Here wed scribe a novel GA/ANN distributed approach for assessing the importance of genes for sample classification based on expression data. Several different approaches have been exploited and a comparison has been given. The developed system was employed in the classification of ER+/- metastasis recurrence of breast cancer tumors and results were validated using a real life database. Further validation has been carried out using Gene Ontology based tools. Results proved the valuable potentialities and robustness of similar systems.

Key Words : Diagnosis, diagnostic tests, drug discovery, Support Vector Machines.

1. INTRODUCTION

Introduced for the first time in 1989, microarrays have gained in this time a great fame thanks to their ability to give biologists a quite detailed snapshot of cellular and genomic activity in particular states of the examined organism. Recent advances in microarray technology have allowed studying the expression patterns of thousands of genes in parallel. The principles these devices are based on are really few and simple. Microarrays use hybridization-based methodology that allows mRNA molecules to bind to their complementary parts (genes). Several probes for each gene are placed on a coated quartz surface (1.28 cm x 1.28/ cm); mRNA segments hybridize with probes according to A-T C-G base pairing principle and this allows the monitoring of the expression levels of thousands of genes simultaneously. This enables the measurement of the levels of mRNA molecules inside a cell and, consequently, the proteins being produced. Hence, the role of the genes in a cell at a given moment can be better understood by analyzing their expression levels. In this context, the comparison between gene expression patterns through the measurement of the levels of mRNA in healthy versus unhealthy cells can supply important information about pathological states, as well as information that can lead to earlier diagnosis and more efficient treatment.

Many genes are strongly regulated and only transcribed at certain times, in certain environmental conditions, and in certain cell types. Microarrays simultaneously measure the mRNA expression level of thousands of genes in a cell mixture. By comparing the expression profiles of different tissue types we might find the genes that best explain a perturbation or might even help clarify how cancer is developing. Given a series of microarray experiments for a specific tissue under different conditions. To find the genes most likely differentially expressed under these Conditions. In other words, To find the genes that best explain the effects of these conditions. This task is also called feature selection, a commonly addressed problem in machine learning, where one has class-labeled data and wants to figure out which features best discriminate among the classes. If the genes are the features describing the cell, the problem is to select the features that have the biggest impact on describing the results and to drop the features with little or no effect. These features can then be used to classify unknown data. Noisy or irrelevant attributes make the classification task more complicated, as they can contain random

correlation. Therefore we want to filter out these features.

2. GENE IDENTIFICATION:

The dataset contains gene expression information extracted from DNA microarrays. This microarray dataset is used to distinguish tumor and normal tissues. There are 62 tissue samples, of which 22 are normal and 40 are cancer tissues, each having 2000 genes with highest minimal intensity across the 62 issues. The data set was divided into a training set with 32 samples and a test set with 30 samples total number of 5157 subsets of genes that correctly classify all training samples are obtained using our bootstrapped GA/SVM algorithms. Each subset consists of five genes. Genes are then ordered based on the number of occurrences with which genes are selected. Figure 1shows the number of occurrences for each gene.

3. GENE CLASSIFICATION:

Classification problems where the input is a vector that call a "pattern" of n components and call a "features". F the n dimensional feature space. In the case of the problem at hand, the features are gene expression coefficients and patterns correspond to patients. The limit ourselves to two-class classification problems. To identify the two classes with the symbols (+) and (-). A training set of a number of patterns $\{x_1, x_2, \dots x_k, \dots x_l\}$ with known class labels $\{y_1, y_2, \dots y_k, \dots y_l\}$, y_kî{-1,+1}, is given. The training patterns are used to build a decision function (or discriminate function) D(x), that is a scalar function of an input pattern x. New patterns are classified according to the sign of the decision function:

D(x) > 0 =>class (+)
D(x) < 0 => class (-)
D(x) = 0, decision boundary.

Decision functions that are simple weighted sums of the training patterns plus a
bias are called linear discriminate functions In our notations:

D(x) = w.x+b, (1)
where w is the weight vector and b is a bias value.
A data set is said to be "linearly separable" if a linear discriminate function can
separate it without error.

3. SPACE DIMENSIONALITY REDUCTION AND FEATURE SELECTION:

A known problem in classification specifically, and machine learning in general, is to find ways to reduce the dimensionality n of the feature space F to overcome the risk of "over fitting". Data over fitting arises when the number n of features is large (in our case thousands of genes) and the number l of training patterns is comparatively small (in our case a few dozen patients). In such a situation, one can easily find a decision function that separates the training data (even a linear decision function) but will perform poorly on test data. Training techniques that use regularization avoid over fitting of the data to some extent without requiring space dimensionality reduction. Such is the case, for instance, of Support Vector Machines (SVMs) benefit from space dimensionality reduction.

4. SUPPORT VECTOR MACHINE (SVM) ALGORITHM:

The SVM learning algorithm is fairly simple. Our implementation follows the formulation .This approach differs slightly from that the geometric interpretation remains the same.

$$\mathbf{L(x)}=\sum_{i=1}^{n} y_i \propto_i k(\mathrm{x}, \mathrm{x}_i)$$

The goal is to learn a set of weights that maximize the following objective function:

$$J(\propto) = \sum_{i=1}^{n} \propto_i (2 - y_i L(x_i))$$

$$= f\left(\frac{1 - y_i L(x_i) + \propto_i k(x_{i,} x_i)}{k(x_{i,} x_i)}\right)$$

This maximum can be obtained by iteratively updating the weights using the following update rule:

$$\propto_i \leftarrow f\left(\frac{1 - y_i L(x_i) + \propto_i k(x_{i,} x_i)}{k(x_{i,} x_i)}\right)$$

where $f(x)=x$ for $x > 0$ and $f(x)=0$ for $x <= 0$. difference arises because we implement the soft margin by

modifying the diagonal of the kernel matrix, rather than by truncating the weights.

5. FEATURE RANKING WITH SUPPORT VECTOR MACHINES:

To test the idea of using the weights of a classifier to produce a feature ranking, we used a state-of-the-art classification technique: SVMs have recently been intensively . They are presently one of the best-known classification techniques with computational advantages over their contenders SVMs. The handle non-linear decision boundaries of arbitrary complexity, we limit ourselves, in this paper, to linear SVMs because of the nature of the data sets under investigation. Linear SVMs are particular linear discriminate classifiers An extension of the algorithm to the non-linear If the training data set is linearly separable, a linear SVM is a maximum margin classifier. The decision boundary (a straight line in the case of a two-dimensional separation) is positioned to leave the largest possible margin on either side.

6. COLON CANCER DIAGNOSIS:

Gene expression information was extracted from DNA micro-array data . The 62 tissues include 22 normal and 40 colon cancer tissues. The matrix contains the expression of the 2000 genes with highest minimal intensity across the 62 tissues. Some genes are non-human genes provides an analysis of the data based on top down hierarchical clustering, a method of unsupervised learning. They show that most normal samples cluster together and most cancer samples cluster together. They explain that “outlier” samples that are classified in the wrong cluster differ in cell composition from typical samples. They compute a so-called “muscle index” that measures the average gene expression of a number of smooth muscle genes.

7. RESULTS :

Our experiments show the benefits of classifying genes using support vector machines trained on DNA microarray expression data. We begin with a comparison of SVMs versus four non-SVM methods and show that SVMs provide superior performance. We then examine more closely the performance of several different SVMs and demonstrate the superiority of the radial basis function SVM. Finally, we examine in detail some of the apparent errors made by the radial basis function SVM and show that many of the apparent errors are in fact biologically reasonable classifications.

SVM performance using various kernels. SVMs were trained using four different kernel functions on five different random three-fold splits of the data, training on two-thirds and testing on the remaining third. The first column contains the class . The second column contains the kernel function.The next five columns contain the threshold-optimized cost (i.e., the number of false positives plus twice the number of false negatives) for each of the five random three-fold splits. The final column is the total cost across all five splits.

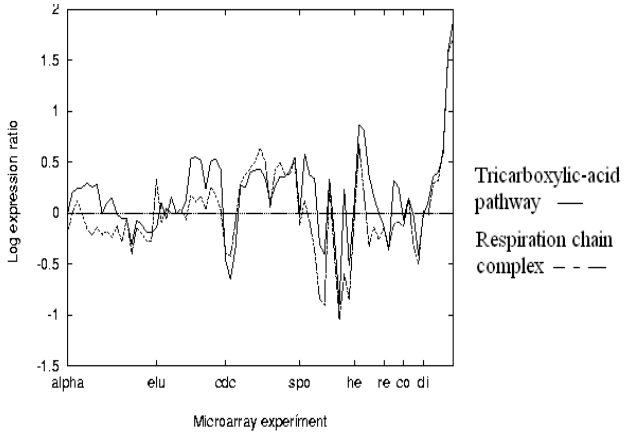
Class	Kernel	Cost for each split					Total
Tricarboxylic acid	Radial	18	21	15	22	21	97
	Dot-product-1	15	22	18	23	22	100
	Dot-product-2	16	22	17	22	22	99
	Dot-product-3	16	22	17	23	22	100
Respiration	Radial	16	18	23	20	16	93
	Dot-product-1	24	24	29	27	23	127
	Dot-product-2	19	19	26	24	23	111
	Dot-product-3	19	19	26	22	21	107
Ribosome	Radial	8	12	15	11	13	59
	Dot-product-1	13	18	14	16	16	77

	Dot-product -2	11	16	14	16	15	72
	Dot-product -3	9	15	11	15	15	65
Proteasome	Radial	14	10	9	11	11	55
	Dot-product -1	16	12	12	17	19	76
	Dot-product -2	16	13	15	17	17	78
	Dot-product -3	16	13	16	16	17	79
Histone	Radial	4	4	4	4	4	20
	Dot-product -1	4	4	4	4	4	20
	Dot-product -2	4	4	4	4	4	20
	Dot-product -3	4	4	4	4	4	20

To demonstrating the superior performance of SVMs relative to non-SVM methods, the radial basis SVM performs better than SVMs that use a scaled dot product kernel. In order to verify this difference in performance, we repeated the three-fold cross-validation experiment four more times, using four different random splits of the data. The total cost in all five experiments is reported in the final column of the table. The radial basis SVM performs better than the scaled dot product SVMs for all classes except the histones, for which all four methods perform identically.

The number of genes that a radial basis SVM misclassifies only once in the five experiments. The right-most column lists the number of genes that are consistently misclassified in all five experiments. These latter genes are of much more interest, since their misclassification cannot be attributed to an unlucky split of the data

Each series represents the average log expression ratio for all genes in the given family plotted as a function of DNA microarray experiment.



These are genes for which the radial basis support vector machine consistently disagrees with the classification. Many of these disagreements reflect the different perspective provided by the expression data concerning the relationships between genes. The microarray expression data represents the genetic response of the cell to various environmental perturbations, and the SVM classifies genes based on how similar their expression pattern is to genes of known function. The definitions of functional classes have been arrived at through biochemical experiments that classify gene products by what they do, not how they are regulated. These different perspectives sometimes lead to different functional classifications. The genes that are regulated at the translational level or protein level, rather than at the trancriptional level measured by the microarray experiments, cannot be correctly classified by expression data alone. Third, genes for which the microarray data is corrupt cannot be correctly classified. Disagreements represent the cases

where the different perspectives of the SVM lead to different functional classifications and illustrate the new information that expression data brings to biology.

8. CONCLUSIONS:

SVMs lend themselves particularly well to the analysis of broad patterns of gene expression from DNA micro-array data. They can easily deal with a large number of features (thousands of genes) and a small number of training patterns (dozens of patients). They integrate pattern selection and feature selection in a single consistent framework.

The top ranked genes found by SVM all have a plausible relation to cancer. In contrast, other methods select genes that are correlated with the separation at hand but not relevant to cancer diagnosis. This simple method allows us to find nested subsets of genes that lend themselves well to a model selection technique that finds an optimum number of genes. Our explorations indicate is much more robust to data overfitting than other methods, including combinatorial search.

Further work includes experimenting with the extension of the method to nonlinear classifiers, to regression, to density estimation, to clustering, and to other kernel methods. We envision that linear classifiers are going to continue to play an important role in the analysis of DNA micro-array because of the large ratio number of features over number of training patterns. generally, the simultaneous choice of the learning machine and the feature subset should be addressed, an even more complex and challenging model selection problem.

9. REFERENCES:

[1] (Aerts, 1996) Chitotriosidase - New Biochemical Marker. Hans Aerts. Gauchers News, March, 1996.

[2] (Alizadeh, 2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Ash A. Alizadeh *et al*, Nature, Vol. 403, Issue 3, February, 2000.

[3] (Alon, 1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon cancer tissues probed by oligonucleotide arrays. Alon *et al*, PNAS vol. 96 pp. 6745-6750, June 1999, Cell Biology. The data is available on-line at http://www.molbio.princeton.edu/colondata.

[4] (Aronson, 1999) Remodeling the Mammary Gland at the Termination of Breast Feeding: Role of a New Regulator Protein BRP39, The Beat, University of South Alabama College of Medecine, July, 1999.

[5] (Ben Hur, 2000) A support vector method for hierarchical clustering. A. Ben Hur, D. Horn, H. Siegelman, and V. Vapnik. Submitted to NIPS 2000. (Boser, 1992) An training algorithm for optimal margin classifiers. B. Boser, I. Guyon, and V. Vapnik. In Fifth Annual Workshop on Computational Learning Theory, pages 144--152, Pittsburgh, ACM. 1992.

[6] (Perou, 1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers, Charles M. Perou et al Proc. Natl. Acad. Sci. USA, Vol. 96, pp. 9212–9217, August 1999, Genetics.

[7] (Schölkopf, 1998) Non-linear component analysis as a kernel eigenvalue problem. B. Schölkopf, A. Smola, K.-R. Muller. Neural computation, vol. 10, pp. 1299-1319, 1998.

[8] (Smola, 2000) Sparce greedy matrix approximation for machine learning. A. Smola and B. Schölkopf. Proceedings of the 17th international conference onmachine learning, pp 911-918. June, 2000.