Gene Selection via a Spectral Approach

L. Wolf

The McGovern institute
for brain research
M.I.T
Cambridge, MA

A. Shashua

School of Engineering
and Computer Science
The Hebrew University
Jerusalem, Israel

S. Mukherjee

Institute of Statistics
and Decision Sciences
Duke University
Durham, NC

Abstract

Array technologies have made it possible to record simultaneously the expression pattern of thousands of genes. A fundamental problem in the analysis of gene expression data is the identification of highly relevant genes that either discriminate between phenotypic labels or are important with respect to the cellular process studied in the experiment. Examples include: cell cycle or heat shock in yeast experiments, chemical or genetic perturbations of mammalian cell lines, and genes involved in class discovery for human tumors. We focus on the task of unsupervised gene selection. Selecting a small subset of genes is particularly challenging as the data sets involved are typically characterized by a small sample size and a very large feature space. We propose a model independent approach which scores candidate gene selections using spectral properties of the candidate affinity matrix. The algorithm is simple to implement, yet contains a number of remarkable properties which guarantee consistent sparse selections.

We applied our algorithm on five different data sets. The first consists of time course data from four well studied Hematopoietic cell lines (HL-60, Jurkat, NB4, and U937). The other four data sets include three well studied treatment outcomes (large cell lymphoma, childhood medulloblastomas, breast tumors) and one unpublished data set (lymph status). We compared our approach both with other unsupervised methods (SOM,PCA,GS) and with supervised methods (SNR,RMB,RFE). The results show that our approach considerably outperforms all the other unsupervised approaches in our study, is competitive with supervised methods and in some cases even outperforms supervised approaches.

Keywords: gene selection, spectral methods, microarray analysis.

1 Introduction

In DNA microarray expression studies, estimated abundances of thousands of mRNA species in tissue samples are obtained through hybridization to oligonucleotide or cDNA arrays. Biological class differences are manifested as significant differences in the expression levels of a *small* set of genes, resulting in the observed overabundance of mRNA. The set of relevant genes is typically small, since the majority of the active cellular mRNA is not affected by the biological differences. In other words, a significant difference in biological characteristics (e.g a normal cell versus a tumor cell from the same tissue) does have a gene expression level manifestation, but the set of genes involved can be rather small. For example, previous work on classification of tumor tissue samples based on gene expression profiles has shown that in many cases, cancer types can be discriminated using only a small subset of genes whose expression levels are strongly correlated with the class distinction [9, 5]. Identifying highly relevant genes from the data is therefore a fundamental problem in the analysis of expression data.

Relevant genes can be selected either in a supervised or unsupervised fashion. A tissue sample consists of a vector in $\mathbb{R}^n$ describing the expression values of n genes/clones. In a *supervised* setting each tissue sample is associated with a label - typically binary or trinary (negative, positive, control) - denoting its class membership. In an *unsupervised* setting, the class labels are omitted or unknown. A variety of algorithms exist for supervised gene selection: signal-to-noise [9], recursive feature elimination [10], t-test metrics [13], Wilcoxon rank sum test [13], and gene shaving [11]. These studies make an implicit assumption that relevant genes are discriminative genes.

However, discrimination is not the only measure of relevance and there are many studies where the objective does not necessarily consist of some measure of discrimination. These include: finding genes relevant for cell cycle [1], analyzing a compendium of expression profiles with different mutations [12], and class discovery [21].

Unsupervised methods for selecting relevant features have been applied in these types of problems using singular value decomposition (SVD) [1], principle components analysis (PCA) and iterative principle components analysis, a.k.a gene shaving [11, 17], max-surprise [2], self organizing maps [21], and hierarchical clustering [6].

In this paper we focus on the task of unsupervised gene selection. Gene selection, unlike other applications of feature selection in the machine learning literature, is characterized first and foremost by a very small sample size q — typically in the order of few tens of tissue samples — and by a relatively very large feature space $\mathbb{R}^n$ as the number of genes tend to be in the thousands ($n \approx 10^4$). Coupled with the notion that applications in the domain of unsupervised gene selection (such as ”class discovery”) require one to discover things which are unknown or unexpected, it follows that the unsupervised gene selection process should be model-independent as much as possible. Motivated by this fact, we approach the gene selection task as a process of dividing the tissue samples into k (typically $k = 2, 3$) clusters where the goal is to find the gene subset which maximizes the clusters coherency. In other words, we assume that if one knew which were the relevant genes to begin with then the tissue sample values corresponding to the selected genes would be naturally clustered in into k sets. Our task is therefore to find that subset which maximizes the cluster coherency.

The task of gene selection is somewhat different from that of interpreting patterns of gene expressions [21, 4]. In the latter case, by regarding the quantitative expression levels of n genes over q samples as defining n points in $\mathbb{R}^q$, one employs a clustering technique for grouping together genes which have similar expression profiles and use the cluster averages as expression profiles. Various clustering techniques have been proposed, from direct visual inspection [4] to employing self organization maps (SOM) [21]. The visual inspection approach does not scale well to larger data sets, is best suited for data with an expected pattern (like cyclic cell lines) and is therefore less appropriate for discovering unexpected patterns. The SOM clustering approach, as all exploratory data analysis tools, involves manual inspection of the data to extract insights. Gene selection in comparison is an open problem. Rather than grouping the gene expression levels into clusters, one seeks to distinguish a small set of genes which are relevant to the biological classification of the tissue samples. As a result of this distinction, rather than grouping the gene expression levels we look for a subset of genes for which the corresponding tissue sample values are coherently divided into k (2 or 3) clusters. The notion of clustering is still there but in an indirect manner — *the goodness of clustering is used as a score for the gene selection process*.

The clustering score in our approach is measured indirectly. Rather than explicitly performing a clustering phase per gene selection candidates, we employ spectral information in order to measure the cluster arrangement coherency. Spectral algorithms have been proven to be successful in clustering, manifold learning or dimensionality reduction, and approximation methods for NP-hard graph theoretical problems. In a nutshell, given a selection of genes, the strength (magnitude) of the leading k eigenvalues of the affinity matrix constructed from the corresponding expression levels of the selected genes is directly related to the coherence of the cluster arrangement induced by the subset of selected genes. More details are described in Section 2.

It is worthwhile to note that unsupervised gene selection differs from dimensionality reduction in that it only selects a handful of genes (features) which are “relevant” with respect to some inference task. Dimensionality reduction algorithms, PCA for example, generate a small number of features, each of which is a combination of all the original features. A main purpose of expression analysis is to extract a set of genes that are of interest from the perspective of the biological process being studied. In general, it is assumed that each such process involves a limited number of genes. For this reason, feature combination methods are not as desirable as methods that extract a small subset of genes. The challenge is to overcome the computational burden of pruning an exponential amount of gene subsets. The $Q - \alpha$ algorithm [24, 19, 25] which we propose to use as a basis for our approach handles the exponential search space by harnessing the spectral information (the sum of eigenvalues of the candidate affinity matrix) in such a manner where a computationally straightforward optimization guarantees a sparse solution, i.e., a selection of genes rather than a combination of the original genes.

2 Methods: Selecting Genes with a Spectral Approach

The array based technologies, cDNA and oligonucleotide, for studying gene expression levels provide static information about gene expression (i.e. in which tissue(s) the gene is expressed) and dynamic information (i.e. how the expression pattern of one gene relates to those of others). In general, the raw data has to be corrected for different experimental conditions by a normalization procedure sometimes followed by a logarithmic transformation to the absolute intensities or ratios. This results in a data matrix whose rows correspond to genes and whose columns correspond to tissue samples. We assume that exactly one value for each gene/sample is given, which may be achieved over repeated measurements for samples or genes.

Let the microarray data matrix be denoted by M. The gene expressions levels that form the rows of M are denoted by $\mathbf{m}_1^\top, ..., \mathbf{m}_n^\top$ and are normalized to unit norm $\|\mathbf{m}_i\| = 1$. Each row vector represents a gene sampled over the q trials. The column vectors of M represent the q samples (each sample is a vector in $\mathbb{R}^n$). As mentioned in the previous section, our goal is to select rows (genes) from M such that the corresponding candidate data matrix (containing only the selected rows) consists of columns that are coherently clustered in k groups. The value of k is user specified and is typically 2 or 3 denoting the expected number of different biological classes in the tissue samples.

Mathematically, to obtain a clustering coherency score, we compute the "affinity" matrix of the candidate data matrix defined as follows. Let $\alpha_i \in \{0, 1\}$ be the indicator value associated with the i'th gene, i.e., $\alpha_i = 1$ if the i'th gene is selected and zero otherwise. Let A_α be the corresponding affinity matrix whose (i, j) entries are the inner-product (correlation) between the i'th and j'th columns of the resulting candidate data matrix: $A_\alpha = \sum_{i=1}^n \alpha_i \mathbf{m}_i \mathbf{m}_i^\top$ (sum of rank-1 matrices). From algebraic graph theory, if the columns of the candidate data matrix are coherently grouped into k clusters, we should expect the leading k eigenvalues of A_α to be of high magnitude [14, 7]. The resulting scheme maximizes the sum of eigenvalues of the candidate data matrix over all possible settings of the indicator variables α_i.

What we do in practice, in order to avoid the exponential growth of assigning binary values to n indicator variables, is to allow α_i to receive real values. A least-squares energy function over the variables α_i is formed and its optimal value is sought. What makes this approach different from the "garden variety" soft-decision-type algorithms is that in this particular formulation, optimizing over spectral properties guarantees that the α_i *always come out positive and sparse* over all local maxima of the energy function. The energy function takes the following form:

$$\max_{Q,\alpha_i} \ \text{trace}(Q^\top A_\alpha^\top A_\alpha Q) \qquad (1)$$

$$\text{subject to} \quad \sum_{i=1}^n \alpha_i^2 = 1, \ \ Q^\top Q = I$$

Note that the matrix Q holds the first k eigenvectors of A_α and that trace$(Q^\top A_\alpha^\top A_\alpha Q)$ is equal to the sum of squares of the leading k eigenvalues: $\sum_{j=1}^k \lambda_j^2$. A local maximum of the energy function is achieved by interleaving the "orthogonal iteration" scheme [8] within the computation of α as follows:

Definition 1 (Spectral Gene Selection) *Let M be an $n \times q$ input matrix with rows $\mathbf{m}_1^\top, ..., \mathbf{m}_n^\top$, and let there be some orthonormal $q \times k$ matrix $Q^{(0)}$, i.e., $Q^{(0)^\top} Q^{(0)} = I$. Perform the following steps through a cycle of iterations with index $r = 1, 2, ...$*

1. *Let $G^{(r)}$ be a matrix whose (i, j) components are $(\mathbf{m}_i^\top \mathbf{m}_j)\mathbf{m}_i^\top Q^{(r-1)} Q^{(r-1)^\top} \mathbf{m}_j$.*
2. *Let $\alpha^{(r)}$ be the leading eigenvector of $G^{(r)}$.*
3. *Let $A^{(r)} = \sum_{i=1}^n \alpha_i^{(r)} \mathbf{m}_i \mathbf{m}_i^\top$.*
4. *Let $Z^{(r)} = A^{(r)} Q^{(r-1)}$.*
5. *$Z^{(r)} \xrightarrow{QR} Q^{(r)} R^{(r)}$, that is, $Q^{(r)}$ is determined by the "QR" factorization of $Z^{(r)}$.*
6. *Increment index r and go to step 1.*

Note that steps 4 and 5 of the algorithm consist of the "orthogonal iteration" module, i.e., if we were to repeat *only* these we would converge onto the eigenvectors of $A^{(r)}$. However, the algorithm does not repeat these steps in isolation and instead recomputes the weight vector α (steps 1,2,3) before applying another cycle of steps 4 and 5.

The algorithm is meaningful provided that three conditions are met: (1) the algorithm converges to a local maximum. (2) at the local maximum $\alpha_i \geq 0$ (since negative weights are not admissible), and (3) the weight vector α is *sparse* (since without it the soft decision does not easily translate into a hard gene selection).

Conditions (2) and (3) are not readily apparent in the formulation of the algorithm (the energy function lacks the explicit inequality constraint $\alpha_i \geq 0$ and an explicit term to "encourage" sparse solutions) but are nevertheless satisfied. The key for having sparse and non-negative weights is buried in the matrix G (step 1). Generally, the entries of G are not necessarily positive (otherwise α would have been non-negative due to the Perron-Frobenious theorem). However, it can be shown that in a probabilistic manner the leading eigenvector of G is positive with probability $1 - o(1)$ (i.e, as the number of genes n grows larger the chances that the leading eigenvector of G is positive increases rapidly to unity). Fig. 1 shows the (sorted) α values for the Hematopoietic differentiation cell lines (details about this data set are found below). The details of why the makeup of G induces such a property, the convergence proof and the proof of the "Probabilistic Perron-Frobenious" claim can be found in [24].

Finally, it is worth noting that the scheme can be extended to handle the supervised situation; that the scheme can be applied also to the Laplacian affinity matrix; and that the scheme readily applies when the spectral gap $\sum_{i=1}^k \lambda_i^2 - \sum_{j=k+1}^q \lambda_j^2$ is maximized rather than $\sum_{i=1}^k \lambda_i^2$ alone. Details can be found in [24].

3 Data sets

We evaluated our proposed approach for gene selection on five data sets — one of which is a time course data set and the remaining four data sets with outcome or status labels. With the four data sets with label information we applied supervised approaches to compare with our unsupervised gene selection algorithm.

The first data set consisted of time course data from four Hematopoietic cell lines [21]: HL-60, Jurkat, NB4, and U937. The dimensionality of the expression data was $7,229$ genes. The HL-60, U937, and Jurkat cell lines were stimulated with phorbol 12-myristate 13-acetate (PMA) for $(0,\ .5,\ 4,\ 24)$ hours. The NB4 cell line was stimulated with all trans-retinoic acid (ATRA) for $(0,\ 6,\ 24,\ 48,\ 72)$ hours.

The remaining four data sets were treatment outcome or status studies. The first was a study of treatment outcome of patients with diffuse large cell lymphoma (DLCL), referred to as ”lymphoma” [20]. The dimensionality of this data set was $7,129$ and there were 32 samples with good successful outcome and 26 with unsuccessful outcome. The second was a study of treatment outcome of patients with childhood medulloblastomas [15], referred to as ”brain”. The dimensionality of this data set was $7,129$ and there were 39 samples with good successful outcome and 21 with unsuccessful outcome. The third was a study of the metastatis status of patients with breast tumors [22], referred to as ”breast met”. The dimensionality of this data set was $24,624$ and there were 44 samples for which the patients were disease-free for 5 years after onset and 34 samples where the tumors metastasized within five years. The fourth is an unpublished study of of breast tumors [16] for which corresponding lymph nodes were either cancerous or not, referred to as ”lymph status”. The dimensionality of this data set is $12,600$ with 47 positive samples for lymph status and 43 negative samples.

4 Results

The above detailed data sets which were used for our experiments consist of thousands of genes (in the order of 10^4). Many of the techniques presented in the past start with a pre-filtering step aiming at reducing the number of genes from thousands to hundreds. For example, [21] passes the gene expression vectors through a variation filter before applying the SOM code for clustering the remaining gene expression vectors. The variation filter eliminates those genes with no significant change across the samples.

One of the strengths of our approach is the ability to handle large amounts of data. Any preprocessing filtering step of the data imposes a prior which very likely has a dramatic effect on the final results. In many cases, the final results depend not so much on the strength of the main algorithm but on the type and care placed on the pre-filtering step. Here, we applied our algorithm on the original data set without performing pre-filtering steps. The results reported below start with data matrices consisting of thousands of genes and produce tens of relevant genes.

4.1 Comparison with SOM on Time Course Data

A significant amount of expression data is time course data. Finding relevant genes in these types of data sets is an open problem. PCA is a reasonable approach when the underlying factor of the study is cyclical, for example cell cycle [1] or circadian rhythms [18]. However, for many studies the underlying process of interest is not cyclical. One may wish to find genes that increase in expression over time in one cell line but decrease in expression for another cell line. A standard approach to address this is to cluster genes and use the clusters as expression profiles. Using our unsupervised gene selection procedure we can find the relevant genes in time course data directly, without having to cluster.

In [21] Hematopoietic differentiation was studied across four cell lines. Two myeloid cell lines HL-60 and U937 were examined, a T cell line called Jurkat was examined, and an acute promyelocytic leukemia cell line was examined. Time course data for these four cell lines was concatenated into a data set with 17 samples and 7229 genes. A 6×4 self-organizing map (SOM) was used to cluster this data set after preprocessing with a variation filter. The 24 clusters are displayed in Fig. 2(a). We applied our algorithm to this data set and found that the set of relevant genes was sparse (meaning it contained a small number of relevant genes), as shown in Fig. 1. Of the genes corresponding to the top 40 α values, we display the time course signatures of 6 genes if Fig.2(b) for brevity. The signature of all 40 genes will be available on our web page. The time course of these 6 genes correspond to clusters 20, 1, $22/23$, 4, 15, 21 in Fig.2(a). For a biological explanation of these genes or corresponding clusters see [21]. Using our algorithm, we were able to recapitulate the time courses of [21] with *individual genes* rather than gene clusters and also find those genes that are relevant.

4.2 Comparison with Other Supervised and Unsupervised Methods using Labeled Data

For the four data sets with label information, classification accuracy was used as a measure of the goodness of our (unsupervised) algorithm. We compared the leave-one-out error on these data sets with the one achieved by both supervised and unsupervised methods of gene selection. For both supervised and unsupervised methods, only the training examples were used in the process of feature selection and dimensionality reduction. The supervised methods used were signal-to-noise (SNR) [9], radius-margin bounds (RMB) [3, 23], and recursive feature elimination (RFE) [10]. The unsupervised methods used were PCA and gene shaving (GS) [11]. In the unsupervised mode, the class labels were ignored — and therefore in general one should expect the supervised approaches to produce superior results than the unsupervised ones. A linear support vector machine classi-

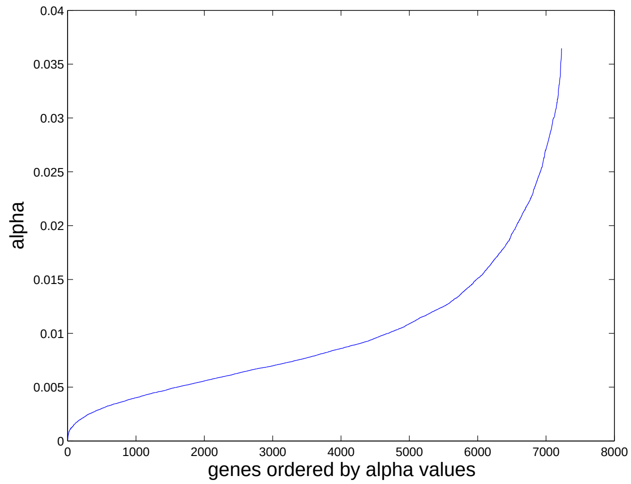

Figure 1: A plot of the sorted α-values for the Hematopoietic differentiation cell lines. As noted, all values come out positive despite the fact that positivity is not explicitly constrained in the energy function. The profile of the values indicates sparsity meaning that around 95% of the values are of an order of magnitude smaller than the remaining 5%.

fier was used for all the gene selection methods. Parameters for SNR, RFE, and RMB were chosen to minimize the leave-one-out error. A summary of the results appears in table 1.

Our algorithm considerably out-performs all other unsupervised methods. Furthermore, and somewhat intriguing, is that our algorithm is competitive with the other supervised algorithm (despite the fact that the labels were not taken into account in the course of running the algorithm) and performs *significantly better* on the lymph status of breast tumors as compared to all other gene selection approaches — including the supervised methods.

5 Discussion

The advent of array technologies make it possible to collect data on thousands of genes simultaneously while recording both static information (in which of the tissues the gene is expressed) and dynamic information (how the expression pattern of one gene relates to those of others). Typical microarray data contains tens of thousands of genes over relatively few samples. It has been observed in many cases, that among the many genes, only a small fraction are really relevant for providing discriminatory information over the tissue samples or providing other non-discriminatory information about class discovery or cell line analysis. The task of selecting a small subset of genes is particularly challenging from an information theoretic point of view, in light of the few samples. Another challenge in gene selection is the combinatorial explosion introduced when all possible gene subsets are to be scored.

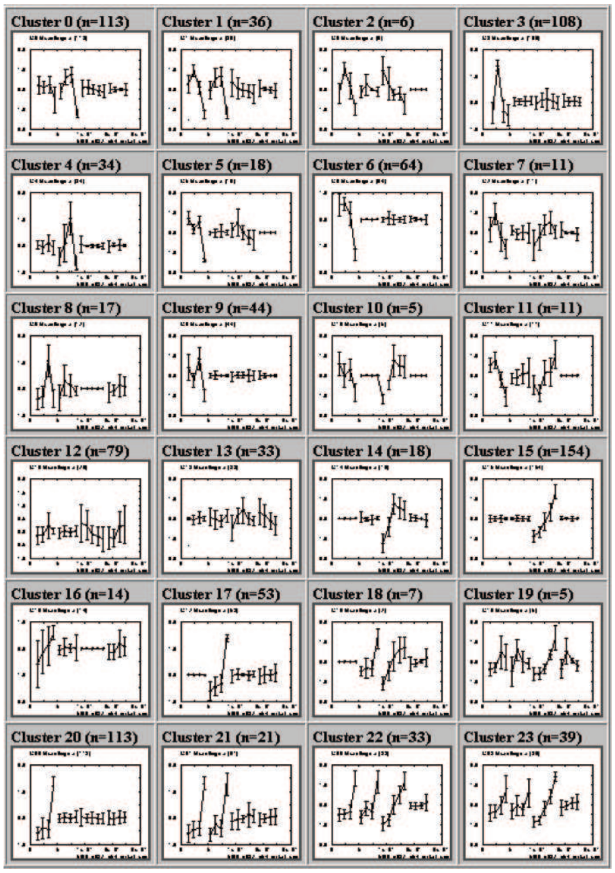

Figure 2: A plot of the 24 SOM clusters from the Hematopoeitic differentiation cell lines. In each of the 24 clusters the time courses of all four cell lines are shown (left to right) HL-60+PMA, U937 + PMA, NB4+ATRA, Jurkat+PMA. This is Figure 4 from [21]

In this work we focused on the unsupervised version of gene selection. The selection is unsupervised when class labels are either absent (as in class discovery) or when the selection is required in the context of cellular process studied in experiments — such as cell cycle or heat shock in yeast experiments, and chemical or genetic perturbations of mammalian cell lines. In these unsupervised settings, our algorithm significantly and consistently outperformed other well studied approaches.

The principle of our method is based on scoring gene subsets by means of measuring the coherence of the cluster arrangements of the sample vectors induced by the selection. The cluster coherence can be indirectly evaluated by the magnitude of the leading eigenvalues of the corresponding affinity matrix. The combinatorial explosion problem is avoided by the special makeup of a key matrix in the algo-

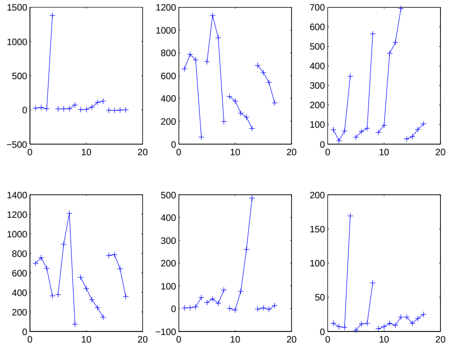

Figure 3: A plot of 6 of the top 40 genes that correspond to clusters 20, 1, $22/23$, 4, 15, 21 in Tamayo et al. In each of the six panels time courses of all four cell lines are shown (left to right) HL-60+PMA, U937 + PMA, NB4+ATRA, Jurkat+PMA.

rithm which makes it possible to use a soft-selection type of approach, yet guarantee sparse solutions.

We have first illustrated the value of our approach on a problem that is inherently unsupervised – that of finding relevant genes in time course data. Instead of directly selecting relevant genes, most algorithms cluster all genes and explain the time courses in terms of gene clusters, and then look for genes in the various clusters to try and understand the underlying biology. We directly find the relevant genes in the time course data. We compared the two approaches on four well studied Hematopoietic cell lines (HL-60, Jurkat, NB4, and U937). Using our approach we were able to find individual genes with time courses very similar to those of gene clusters found using SOMs on this data set. We then applied our algorithm to four (labeled) treatment outcome data sets. Comparisons with other supervised and non-supervised approaches showed a consistent superiority over other unsupervised approaches which we tested in our studies and comparable performance to supervised approaches (despite the fact that our algorithm did not make use of the available class labels). In one case, the performance of our algorithm on the lymph nodes data set for breast tumor study was significantly superior compared to all the methods we compared against — including the supervised methods.

Method	brain	lymph status [1]	breast met.[1]	lymp-. homa
RAW	32	44	34	27
PCA5	22	47	33	40
PCA10	26	47	26	27
PCA20	25	47	25	29
PCA30	31	47	31	33
PCA40	31	47	31	33
PCA50	30	47	30	33
GS5	20	45	32	33
GS10	24	43	31	30
GS20	28	47	32	31
GS30	30	44	33	33
Our Method	15	19	22	15
SNR	16	42	29	18
RFE	14	38	26	14
RMB	13	39	24	14

Table 1: Leave-one-out classification results for the supervised and unsupervised algorithms on the various data sets. In both PCAN and GSN the number N is the number of components used. Parameters for SNR, RFE, and RMB were chosen to minimize the leave-one-out error. Our method considerably outperforms all other unsupervised methods. Furthermore, and somewhat intriguing, is that our algorithm is competitive with the other supervised algorithm (despite the fact that the labels were not taken into account in the course of running the algorithm), and performs significantly better on the lymph status of breast tumors as compared to all other gene selection approaches — including the supervised methods.

[1] Here only the first $7,000$ genes were used.

Acknowledgments

This report describes research done at the Center for Biological & Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain & Cognitive Sciences, and which is affiliated with the Computer Sciences & Artificial Intelligence Laboratory (CSAIL). This research was sponsored by grants from: Office of Naval Research (DARPA) Contract No. MDA972-04-1-0037, Office of Naval Research (DARPA) Contract No. N00014-02-1-0915, National Science Foundation (ITR/SYS) Contract No. IIS-0112991, National Science Foundation (ITR) Contract No. IIS-0209289, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218693, National Science Foundation-NIH (CRCNS) Contract No. EIA- 0218506, and National Institutes of Health (Conte) Contract No. 1 P20 MH66239-01A1. Additional support was provided by: Central Research Institute of Electric Power Industry (CRIEPI), Daimler-Chrysler AG, Compaq/Digital Equipment Corporation, Eastman Kodak Company, Honda R&D Co., Ltd., Industrial Technology Research Institute (ITRI), Komatsu Ltd., Eugene McDermott Foundation, Merrill-Lynch, NEC Fund, Oxygen, Siemens Corporate Research, Inc., Sony, Sumitomo Metal Industries, and Toyota Motor Corporation.

References

[1] O. Atler et al. Singular value decomposition for genome-wide expression data processing and modeling. *P.N.A.S.* 97:18, pp. 10101-10106, (2000).

[2] A. Ben-Dor, N. Friedman, and Z. Yakhini. Class discovery in gene expression data, *RECOMB*, pp. 31-38, (2001).

[3] O. Chapelle et al. Choosing Multiple Parameters for Support Vector Machines *Machine Learning*, 2001.

[4] R. J. Cho et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 95(5), pp. 65-73, (1998).

[5] S. Dudoit, J. Fridlyand and T. P. Speed. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, 2001.

[6] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster Analysis and Display of Genome-Wide Expression Patterns. *P.N.A.S.*, 95, pp. 14863–14868, (1998).

[7] L. E. Gibbons, D. W. Hearn, P. M. Pardalos, and M. V. Ramana. Continuous characterizations of the maximum clique problem. *Math. Oper. Res.*, 22:754–768, 1997.

[8] G. Golub and C.V. Loan. *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.

[9] T. R. Golub et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression. *Science*, 286:531-537, (1999).

[10] I. Guyon, J. Weston, S. Barnhill and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, Vol. 46, pp. 389–422, (2002).

[11] T. Hastie et al. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns *Genome Biology*, 1(2), pp. 1–21, (2000).

[12] T .R. Hughes et al. Functional Discovery via a Compendium of Expression Profiles. *Cell* , 102:109-126, (2000).

[13] L. Miller et al. Optimal gene expression analysis by microarrays *Cancer Cell.*, Vol. 2, pp. 353–361, (2002).

[14] T.S. Motzkin and E.G. Straus. Maxima for graphs and a new proof of a theorem by turan. *Canadian Journal of Math.*, 17:533–540, 1965.

[15] S. L. Pomeroy et al. Gene Expression-Based Classification and Outcome Prediction of Central Nervous System Embryonal Tumors *Nature*, 415:24, pp. 436–442, (2002).

[16] S. Ramaswamy. Personal communication (2004).

[17] S. Raychaudhuri, J. M. Stuart, and R. B. Altman. Principal Components Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series. *Pacific Symposium on Biocomputing*, Honolulu, Hawaii, 452-463, (2000).

[18] R. Schaffer et al. Microarray analysis of diurnal and circadian-regulated genes in Arabidopsis. *Plant Cell*, 13 (1), pp. 113-123, (2001).

[19] A. Shashua and L. Wolf. Kernel Feature Selection with Side Data using a Spectral Approach. European Conference on Computer Vision (ECCV), May 2004.

[20] M. A. Shipp et al. Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene Expression Profiling and Supervised Machine Learning *Nature Medicine*, 8:1, pp. 68–74, (2002).

[21] P. Tamayo et. al. Interpreting patterns of gene expression with self-organizing maps P.N.A.S., 96:6, pp. 2907-2912, (1999).

[22] L. J. van't Veer et al. Expression profiling predicts cliniacl outcome of breast cancer *Nature*, 415, pp. 530–536, (2002).

[23] J. Weston et al. Feature Selection for SVMs *NIPS*, 13, pp. 668–764, (2001).

[24] L. Wolf and A. Shashua. Direct Feature Selection with Implicit Inference. International Conference on Computer Vision (ICCV), 2003.

[25] L. Wolf and S. Bileschi. Combining Unsupervised Variable Selection with Dimensionality Reduction. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), June 2005.