MICROARRAY CLASSIFICATION USING BLOCK DIAGONAL LINEAR DISCRIMINANT ANALYSIS WITH EMBEDDED FEATURE SELECTION

Lingyan Sheng[1], Roger Pique-Regi[1], Shahab Asgharzadeh[2], Antonio Ortega[1]

[1] Signal and Image Processing Institue, Department of Electrical Engineering
Viterbi School of Engineering, University of Southern California
[2] Department of Pediatrics, Childrens Hospital Los Angeles

ABSTRACT

In this paper, block diagonal linear discriminant analysis (BDLDA) is improved and applied to gene expression data. BDLDA is a classification tool with embedded feature selection, that has demonstrated good performance on simulated data. However, by using cross validation in training, BDLDA is time consuming, thus not an appropriate algorithm for gene expression data, which has a large number of features and relatively small number of samples. In our algorithm, estimated error rate is used as a measure to choose the best model. The algorithm is optimized by repeating the model construction procedure with previously selected features removed, which leads to increased classification robustness. Our algorithm is tested using 10 fold cross validation. In most simulated and real data, our method outperforms the state-of-the-art techniques, showing promise for its use in microarray classification problems. The resulting block structure allows to identify discriminating correlated genes, which is potentially useful in cancer research.

Index Terms— Microarray, LDA, Block Diagonal, Feature Selection

1. INTRODUCTION

RNA microarray technology allows researchers to analyze patterns of gene expression simultaneously recorded in a single experiment. Different gene expression patterns among patients or different tissues can be used for diagnosis or prognosis in cancer research. These datasets have a large number of gene expression values per experiment (several thousands to tens of thousands, even millions), and a relatively small number of experiments (a few dozen).

Traditional linear discriminant analysis (LDA) [1, 2] cannot be applied to gene expression data, because of the singularity of the within-class scatter matrix due to the small sample size. Thus, for these data sets some form of feature selection will always be needed. A number of solutions based on LDA have been proposed to tackle this challenge. One solution is to assume a diagonal covariance matrix, which essentially ignores potential correlation between different features. Examples include diagonal linear discriminant analysis (DLDA) [3] or nearest shrunken centroid (NSC) [4], as well as sequential DLDA (SeqDLDA) [5], a modified DLDA technique that incorporates embedded feature selection. Alternative solutions use regularization methods to impose a structure on the covariance matrix, e.g., shrunken centroid regularized discriminant analysis (SCRDA) [6], where a diagonal regularization matrix is employed. But SCRDA has the same problem as DLDA in that it does not perform well in data with correlations (as will be illustrated by our experiments). While it would be possible to consider more complex classification tools (e.g., support vector machines [7], neural networks [8] and random forests [9]), these tend to not perform as well as simpler LDA-based approaches, e.g., SCRDA, when applied to gene expression data. One likely reason is that these more complex models cannot be accurately learned by limited data. Thus we have a bias-variance trade-off problem: lower variance seen in simple classification methods compensates for the additional bias they introduce [2].

In order to improve performance in the presence of feature correlation (while staying within the general LDA framework), in this paper we focus on block diagonal linear discriminant analysis (BDLDA), first proposed in [10]. Cancer research tends to assume that only a few genes are associated with the disease, and thus BDLDA restricts the maximum number of features to be selected in the model. However, even with limited number of features, reliably estimating all correlations is difficult with small sample size. To reduce the parameters estimated while keeping important correlations between features, BDLDA imposes a block diagonal structure on the covariance matrix. A greedy algorithm is applied to find features to add into candidate models with different block diagonal structures. Cross validation is used to select the best model among all candidate models. Unlike DLDA or NSC, BDLDA performs classification with embedded feature selection, while considering correlations between features. In [10], BDLDA was shown to outperform DLDA on simulated data with sparse covariance structure (e.g., Toeplitz or block diagonal). While these results were promising, feature selection using cross-validation made it impractical for large datasets, e.g., gene expression data.

In this paper, we improve feature selection in BDLDA by using estimated an error rate to select the best model among all candidate models. The estimated error rate is derived from LDA and can be obtained for each candidate block diagonal covariance structure. Within BDLDA direct computation of these error rates is possible even when using a very small number of training samples, because the block diagonal structure is limited to use only small blocks. This error rate metric allows us to avoid cross validation for model selection, and enables BDLDA to be computationally practical even when working on large datasets. In this paper we apply BDLDA to real gene expression data for the first time, with very competitive results. Other improvements with respect to the original BDLDA approach include a repeated feature subset selection (RFSS) technique and a prescreening procedure. With RFSS, that is repeating model construction with previously selected features removed, the algorithm chooses more discriminating features that are independent from previous models. This is useful for gene expression data, because genes belonging to the same pathway tend to be have sparse correlations. The prescreening procedure eliminates features that are not significantly different between two classes, which accelerates model search

and improves performance by removing noise. In Section 3, test results are presented, that show our improved version of BDLDA works particularly well on simulated data with correlated features and outperforms the other three algorithms in real data.

The remainder of this paper is organized as follows. Section 2 presents the design and improvement of our algorithm. Section 3 gives the experimental results to compare against those in [5], [4] and [6]. Section 4 concludes the paper.

2. ALGORITHM DESCRIPTION

2.1. Model Selection Metric

We start by deriving the estimated error rate of LDA, which will be used as a model selection metric in Section 2.2. LDA assumes that both class A and class B have multivariate Gaussian distribution with means $\boldsymbol{m_A}$ and $\boldsymbol{m_B}$ and a common covariance matrix $\boldsymbol{K}$, $f_A(\boldsymbol{x}) \sim N(\boldsymbol{m_A}, \boldsymbol{K})$, $f_B(\boldsymbol{x}) \sim N(\boldsymbol{m_B}, \boldsymbol{K})$. The discriminant function is

$$g(\boldsymbol{x}) = \boldsymbol{w^t x} - \boldsymbol{b} \begin{cases} \geq 0 \Rightarrow \text{class A} \\ < 0 \Rightarrow \text{class B} \end{cases} \quad (1)$$

where $\boldsymbol{x}$ is the feature vector of the sample to classify, $\boldsymbol{w}$ is a vector orthogonal to the hyperplane, and $\boldsymbol{b}$ defines the decision boundary $g(\boldsymbol{x}) = 0$.

For each model in BDLDA, the mean vectors $\boldsymbol{m_A}$, $\boldsymbol{m_B}$, the covariance matrix $\boldsymbol{K}$, and the prior class probabilities π_A, π_B are replaced by the maximum likelihood estimators $\hat{\boldsymbol{m}}_{\boldsymbol{A}}$, $\hat{\boldsymbol{m}}_{\boldsymbol{B}}$, $\hat{\boldsymbol{K}}$, $\hat{\pi}_A$ and $\hat{\pi}_B$. $\hat{\boldsymbol{m}}_{\boldsymbol{A}}$, $\hat{\boldsymbol{m}}_{\boldsymbol{B}}$ and $\hat{\boldsymbol{K}}$ are computed corresponding to the features in the model. $\hat{\boldsymbol{w}}$ is the direction that maximizes variance between/within ratio:

$$J_{\hat{\boldsymbol{K}}}(\hat{\boldsymbol{w}}) = \frac{(\hat{\boldsymbol{d}}^{\boldsymbol{t}}\hat{\boldsymbol{w}})^2}{\hat{\boldsymbol{w}}^{\boldsymbol{t}}\hat{\boldsymbol{K}}\hat{\boldsymbol{w}}} \quad (2)$$

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} J_{\hat{\boldsymbol{K}}}(\boldsymbol{w}) = \hat{\boldsymbol{K}}^{-1}\hat{\boldsymbol{d}} \qquad \hat{\boldsymbol{d}} = \hat{\boldsymbol{m}}_{\boldsymbol{A}} - \hat{\boldsymbol{m}}_{\boldsymbol{B}} \quad (3)$$

In the general LDA case, if the dataset has more features than samples, $\hat{\boldsymbol{K}}$ is not invertible. In BDLDA, we restrict the block sizes to be smaller than sample size, which makes $\hat{\boldsymbol{K}}$ invertible.

Given the training data, the estimated probability of error of a model in BDLDA is

$$\begin{aligned} \hat{\boldsymbol{P}}_{\boldsymbol{e}}|T \quad = \quad & \pi_A \phi\left(-\frac{\frac{1}{2}\hat{\boldsymbol{d}}^{\boldsymbol{t}}\hat{\boldsymbol{K}}^{-\boldsymbol{1}}\hat{\boldsymbol{d}} + \ln(\frac{\pi_A}{\pi_B})}{\sqrt{\hat{\boldsymbol{d}}^{\boldsymbol{t}}\hat{\boldsymbol{K}}^{-\boldsymbol{1}}\hat{\boldsymbol{d}}}}\right) \\ & + \pi_B \phi\left(-\frac{\frac{1}{2}\hat{\boldsymbol{d}}^{\boldsymbol{t}}\hat{\boldsymbol{K}}^{-\boldsymbol{1}}\hat{\boldsymbol{d}} - \ln(\frac{\pi_A}{\pi_B})}{\sqrt{\hat{\boldsymbol{d}}^{\boldsymbol{t}}\hat{\boldsymbol{K}}^{-\boldsymbol{1}}\hat{\boldsymbol{d}}}}\right) \end{aligned} \quad (4)$$

where ϕ is the cdf of standard normal distribution. $\hat{\boldsymbol{d}}$ and $\hat{\boldsymbol{K}}$ are the corresponding mean difference and covariance matrix of the model in BDLDA. T denotes the training dataset.

2.2. Model Construction and Feature Selection

Enumerating models with all possible features and structures of $\hat{\boldsymbol{K}}$ and $\hat{\boldsymbol{d}}$ is obviously impractical due to computation and memory limitation. In [10], a block diagonal structure is imposed on the covariance matrix, with the dimensions of both subblocks and the resulting covariance matrix kept small. An example of resulting candidate models is shown in Figure 1, where each arrow denotes adding a feature to the model and forming a new model. The feature can start a new subblock (solid line in Figure 1) or be combined with the current subblock (dashed line in Figure 1). The feature that generates the largest J in (2) among all candidate features is selected. For simplicity, it is assumed the sizes of subblocks are nonincreasing, because through exhaustive search for features, the features added first are considered of better classification power than those selected later. Thus their correlations are considered more important.

J in (2) is the maximized projected class mean divided by projected variance in the feature space. J increases with the number of features and has no upper limit. Such increase is sometimes due to increasing number of features and does not necessarily improve performance, thus using J by itself for model selection would make the selection undesirable. In order to compare all models with different number of features, [10] uses cross validation. Cross validation is an unbiased method, which does not make any assumptions on the data. Despite its advantages, it is time consuming, making it impractical to select model in large data. Instead in our paper, we propose to use the estimated error rate in (4) for each covariance metric $\hat{K}$ as a way to compare different candidate structures. The covariance structure with smallest $\hat{P}_e$ will be selected at each step in the sequential search shown in Figure 1. Unlike J (2), $\hat{P}_e$ lies in the range of 0 to 1 for all models. If the data has a Gaussian distribution and means and covariance matrix can be estimated, $\hat{P}_e$ is a good measure of each model's performance. In Section 3, experiments on simulated data has demonstrated this advantage. Moreover, in experiments on real gene expression data, which is not strictly Gaussian distributed, the measure still generates better results than algorithms tested in this work.

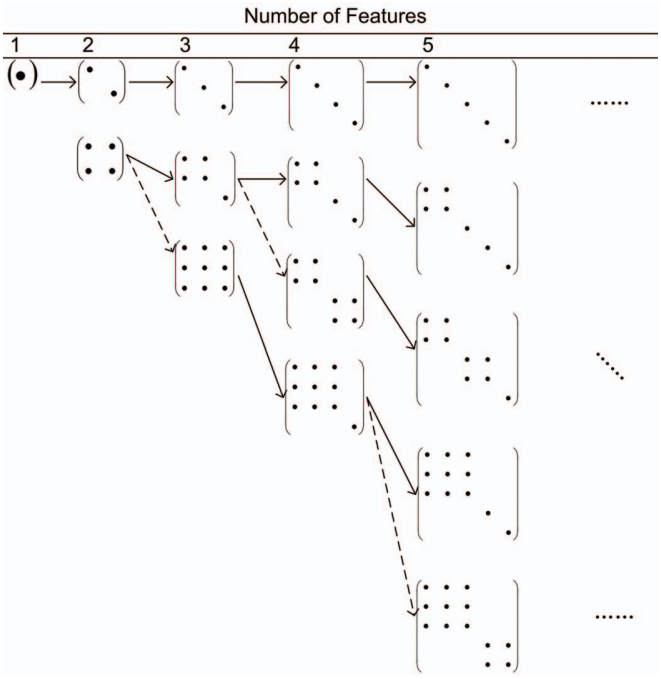

Fig. 1. Sequential generation of candidate covariance matrix models for BDLDA. Starting with an empty list, we add one feature at a time (namely, the one that maximizes J). The best of all these models is selected using $\hat{P}_e$.

2.3. Algorithm Improvement

Repeated feature subset selection (RFSS) is applied to reduce the

impact of previously selected features. RFSS repeats the model construction and feature selection N times. At the start, a model with a predefined maximum number of features ($MaxFeature$ in Table 1) is selected. Then model construction and feature selection is performed again, with the features selected during the first iteration removed from the set of candidate features. This procedure is repeated N times, each time a feature selection iteration does not consider features already selected in previous iterations. Then the N models are combined by vector concatenating N means and block diagonally concatenating N covariance matrices. The feature sets in all N models are different and uncorrelated. The model construction is performed N times or stops when there are not enough candidate features. This improvement enables the algorithm to find more discriminating features without being influenced by previously selected models. The complete algorithm is described in Table 1.

Algorithm: Model Construction with RFSS
$\mathcal{S} = \emptyset, \mathcal{T} =$ all candidate features, $\mathcal{M} = \emptyset$,
$F = 0, L = 0$
1. Construct the first model by adding feature i,
$i = \arg\max_{i \in T} \frac{d_i}{\sigma_i}$
$\mathcal{S} = \mathcal{S} + \{i\}, \mathcal{T} = \mathcal{T} - \{i\}, \mathcal{M} = \mathcal{M} + \{Model\,1\}$,
$F = 1, L = 1$
2. For models with feature size F
(1) Add a feature as an independent subblock.
The new feature is selected by $i = \arg\max_{i \in T} J$ given in Eq. (2)
$\mathcal{S} = \mathcal{S} + \{i\}, \mathcal{T} = \mathcal{T} - \{i\}, \mathcal{M} = \mathcal{M} + \{Model\,j\}$,
$F = F + 1, L = 1$
(2) Add a feature to the last subblock if
$F < MaxGrow$
and $F + 1$ does not exceed any previous subblocks
The new feature is selected by $i = \arg\max_{i \in T} J$ given in Eq. (2)
$\mathcal{S} = \mathcal{S} + \{i\}, \mathcal{T} = \mathcal{T} - \{i\}, \mathcal{M} = \mathcal{M} + \{Model\,j\}$,
$F = F + 1, L = L + 1$
3. Repeat Step 2 until the $MaxFeature$ is reached.
4. Select among $\mathcal{M}$ the model with minimum P_e given in Eq. (4)
5. Remove S and repeat steps 1-4 N times
6. Combine N selected models
$\mathcal{S}$ is the set of selected features. $\mathcal{T}$ is the set of candidate features. $\mathcal{M}$ is the set of candidate models. F is the number of features in the the model. L is the number of features in the last subblock. $MaxGrow$ is the largest size of a subblock. $MaxFeature$ is the largest number of features in the models.

Table 1. Model construction and feature selection

Prescreening is based on the observation that features with the same means and variances are not discriminating in BDLDA. Some of them may be noise and interfere with classification. Removing these features can improve performance and reduce computation time. A prescreening of all the features is applied before the model construction in Table 1. The separation of two classes on feature i is represented by $|\frac{d_i}{\sigma_i}|$. $d_i = m_{Ai} - m_{Bi}$ and $\sigma_i^2 = \frac{1}{K_s}(\sum_{k \in ClassA}(x_{ki} - m_{Ai})^2 + \sum_{k \in ClassB}(x_{ki} - m_{Bi})^2) + c$, where x_{ki} is the ith feature of sample k, K_s is the total number of samples, m_{Ai} is the mean of feature i that belongs to class A and similarly for m_{Bi}, and c is a regularization value.

Only features with $|\frac{d_i}{\sigma_i}|$ above a threshold will be taken into the algorithm. In simulated data, we use $\frac{1}{3}max_i(|\frac{d_i}{\sigma_i}|)$ as the threshold. To avoid the impact of outliers in real data, instead of using the top ranking $|\frac{d_i}{\sigma_i}|$, the average of 10 largest $|\frac{d_i}{\sigma_i}|$ is used, that is, the threshold is one third of the average of 10 largest $|\frac{d_i}{\sigma_i}|$. The prescreening procedure can also be applied to other classification tools.

3. EXPERIMENTAL RESULTS

3.1. Simulated Data

Our algorithm is tested on both simulated data and real data. The results are compared with SeqDLDA [5], NSC [4] and SCRDA [6]. The test results shown in Table 2 and Table 3 are obtained by doing 10 fold cross validation 50 times. The first two datasets are also used in [6]. The error rates of SCRDA in [6] are presented as a matrix according to two tuning parameters. The smallest error rate among all parameter pairs is shown in Table 2 and Table 3.

3.1.1. Block diagonal covariance matrix

The distributions of two classes are $N(\boldsymbol{\mu_A}, \boldsymbol{K})$ and $N(\boldsymbol{\mu_B}, \boldsymbol{K})$ with total number of features $P = 10000$, $\boldsymbol{\mu_A} = (\underbrace{000\cdots00}_{10000})$, and $\boldsymbol{\mu_B} = (\underbrace{0.5\cdots0.5}_{200}\underbrace{00\cdots00}_{9800})$. The block diagonal structure of $\boldsymbol{K}$ is shown in (5). Each subblock has an autoregressive structure, which is a symmetric Toeplitz matrix with the first row $(1\ \rho\ \cdots\ \rho^{98}\ \rho^{99})$. The subblock size is 100×100 and there are a total of 100 subblocks. It is assumed the autocorrelation within each subblock is $|\rho| = 0.9$ and we set alternating signs for each subblock. 220 samples are generated. The average error rates and standard deviations are shown in Table 2.

$$\mathbf{K} = \begin{pmatrix} \boldsymbol{K_\rho} & 0 & 0 & \ddots & \ddots & \ddots \\ 0 & \boldsymbol{K_{-\rho}} & 0 & 0 & \ddots & \ddots \\ 0 & 0 & \boldsymbol{K_\rho} & 0 & \ddots & \ddots \\ \ddots & 0 & 0 & \boldsymbol{K_{-\rho}} & 0 & \ddots \\ \ddots & \ddots & \ddots & 0 & \ddots & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}_{10000 \times 10000} \quad (5)$$

3.1.2. Diagonal covariance matrix

The distributions of two classes are $N(\boldsymbol{\mu_A}, \boldsymbol{K})$ and $N(\boldsymbol{\mu_B}, \boldsymbol{K})$ with total number of features $P = 10000$, $\boldsymbol{\mu_A} = (\underbrace{000\cdots00}_{10000})$, and $\boldsymbol{\mu_B} = (\underbrace{0.5\cdots0.5}_{100}\underbrace{00\cdots00}_{9900})$. We assume that features are independent so that the covariance can be written, $\boldsymbol{K} = \boldsymbol{I_P}$, where $\boldsymbol{I_P}$ is the $P \times P$ identity matrix. 220 samples are generated. The results are shown in Table 2.

3.1.3. Toeplitz covariance matrix

The distributions of two classes are $N(\boldsymbol{\mu_A}, \boldsymbol{K})$ and $N(\boldsymbol{\mu_B}, \boldsymbol{K})$ with total number of features $P = 1000$. The difference of means are assumed to be fading exponentially. $\boldsymbol{\mu_A} = (\underbrace{000\cdots00}_{1000})$. $\boldsymbol{\mu_{Bj}} = e^{-\gamma j}, (j = 1, 2\cdots1000)$. $\gamma = 0.05$. It is assumed $\boldsymbol{K}$ is the following Toeplitz matrix with the first row $(1\ \frac{-1}{2}\ \frac{2}{5}\ 0\ \cdots\ 0)$. 120 samples are generated. The results are shown in Table 2.

Table 2. Average Error Rate (Standard Deviation) for Simulated Data

	Block Diagonal covariance	Diagonal covariance	Toeplitz covariance
BDLDA	**0.36%** (0.19%)	3.57% (1.09%)	**4.51%** (1.02%)
SeqDLDA	19.64% (1.5%)	2.57% 0.83%	8.97% (1.71%)
NSC	18.15% (1.34%)	6.89% (1.12%)	10.82% (1.74%)
SCRDA	9.45% (1.23%)	**1.97%** (0.62%)	10.2% (1.37%)
N=5, MaxGrow=3, MaxFeature=20			

3.2. Real Data

We test our algorithm on two-class cancer data publicly available online: colon cancer (62 samples, 2000 features) [11] and prostate cancer(102 samples, 6033 features) [12]. The neuroblastoma dataset (102 samples, 44298 features) consists of samples from Neuroblastoma stage 4 with MYCN not amplified obtained at diagnosis. 10 fold cross validation is done 50 times on each dataset. The average error rates and standard deviations are shown in Table 3.

Table 3. Average Error Rate (Standard Deviation) for Real Data

	Colon Cancer	Prostate Cancer	Neuro-blastoma
BDLDA	**10.06%** (1.15%)	**5.21%** (0.85%)	**10.61%** (1.29%)
SeqDLDA	12.06% (1.87%)	5.53% (0.9%)	13.87% (2.41%)
NSC	10.31% (1.02%)	7.65% (0.42%)	17.98% (1.67%)
SCRDA	11.41% (1.69%)	5.41% (0.89%)	14.22% (1.39%)
N=5, MaxGrow=3, MaxFeature=20			

3.3. Discussion

In simulated data with block diagonal covariance matrix and Toeplitz covariance matrix, our algorithm (BDLDA) performs much better than all other three methods. Data with diagonal covariance matrix is the only case that BDLDA does not show much advantage. SeqDLDA and SCRDA have slightly lower error rate. But the margin is much smaller than in other two simulated datasets. This result is consistent with the block diagonal assumption BDLDA makes for the covariance matrix. Its ability to find the best discriminating covariance structure makes it promising for datasets with correlations. It is also able to achieve reasonably good classification when the covariance is diagonal. Due to this advantage, BDLDA can be a competitive algorithm for gene expression data, because genes belonging to the same pathway are likely to be co-expressed. This advantage of BDLDA is demonstrated in the real data.

In all three real datasets, BDLDA has the lowest error rates. Among them, Neuroblastoma, with more than 40,000 features, is considered the most challenging. Our algorithm reduced the error rate by more than 2%, compared to the second best algorithm, SeqDLDA.

4. CONCLUSIONS

This paper proposes a classification algorithm applied to microarray expression data. The subset of features and their covariance structure are selected during classification. Block diagonal structure is imposed on the covariance matrix with predefined sizes of matrices and subblocks. Estimated error rate is used to select the best model. RFSS and prescreening are used to improve the algorithm. Our method outperforms SeqDLDA [5], NSC [4] and SCRDA [6] in most simulated data and all real data used in our tests. BDLDA with the feature selection strategy proposed in this paper is very promising to handle datasets with small number of training samples, a very large number of features and an unknown sparse correlation structure. The method is especially useful for microarray data, where sparse correlations will occur among correlated genes that belong to the same biological pathway.

5. REFERENCES

[1] DG Stork RO Duda, PE Hart, *Pattern Classification*, Wiley-Interscience, 2 edition, 2000.

[2] T Hastie, R Tibshirani, and JH Friedman, *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, Springer, 2001.

[3] S Dudoit, J Fridlyand, and TP Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, 2002.

[4] R Tibshirani, T Hastie, B Narasimhan, and G Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression.," *Proc Natl Acad Sci U S A*, vol. 99, no. 10, pp. 6567–6572, May 2002.

[5] R Pique-Regi, A Ortega, and S Asgharzadeh, "Sequential diagonal linear discriminant analysis (seqdlda) for microarray classification and gene identification," in *CSBW*, 2005, pp. 112–116.

[6] Y Guo, T Hastie, and R Tibshirani, "Regularized linear discriminant analysis and its application in microarrays," *Biostatistics*, pp. 86–100, 1 2007.

[7] Vladimir Vapnik, *The Nature of Statistical Learning Theory*, Springer, 2 edition, 2000.

[8] B.D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, 1996.

[9] Ramon Diaz-Uriarte and Sara Alvarez de Andres, "Gene selection and classification of microarray data using random forest.," *BMC Bioinformatics*, vol. 7, pp. 3, 2006.

[10] R Pique-Regi and A Ortega, "Block diagonal linear discriminant analysis with sequential embedded feature selection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2006*, 2006, vol. 5, pp. V–V.

[11] U. Alon et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.," *Proc Natl Acad Sci U S A*, vol. 96, no. 12, pp. 6745–6750, Jun 1999.

[12] Dinesh Singh et al., "Gene expression correlates of clinical prostate cancer behavior.," *Cancer Cell*, vol. I, no. 2, pp. 203–209, Mar 2002.