Sequential Diagonal Linear Discriminant Analysis (SeqDLDA) for Microarray Classification and Gene Identification

Roger Pique-Regi[1], Antonio Ortega[2], Shahab Asgharzadeh[3]
[1,2]University of Southern California, [3]Children's Hospital of Los Angeles
rpique@ieee.org, ortega@sipi.usc.edu, shahab@chla.usc.edu

Abstract

In microarray classification we are faced with a very large number of features and very few training samples. This is a challenge for classical Linear Discriminant Analysis (LDA), since reliable estimates of the covariance matrix cannot be obtained. Alternative techniques based on Diagonal LDA (DLDA) combined with an independent gene selection (filtering) have been proposed.

In this paper we propose a novel sequential DLDA (SeqDLDA) technique that combines gene selection and classification. At each iteration, one gene is sequentially added and the linear dicriminant (LD) recomputed using the DLDA model (i.e., a diagonal covariance matrix). Classical DLDA will add the gene with highest t-test score without checking the resulting model. In contrast, SeqDLDA will find the one gene that better improves class separation after recomputing the model measured using a robustified t-test score.

We evaluate the new method in several 2-class datasets (Neuroblastoma, Prostate, Leukemia, Colon) using 10-fold cross-validation. For example, for the Neuroblastoma data set, the average misclassification rate of DLDA (16.91%) is significantly reduced to 13.87% using SeqDLDA.

1. Introduction

Linear Discriminant Analysis is a well-known and widely used classification technique [1], [2] that finds a hyperplane that partitions the feature space into two decision regions:

$$g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} - b > 0 \Rightarrow \text{Class A ('<' Class B)} \qquad (1)$$

Where $\mathbf{x}$ represents the sample we want to classify, and $\mathbf{w}$ is the vector normal to the hyperplane (2):

$$\mathbf{w} = \mathbf{K}^{-1}(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B); b = -\tfrac{1}{2}\mathbf{w}^t(\boldsymbol{\mu}_A + \boldsymbol{\mu}_B) \qquad (2)$$

This decision rule is optimal [1] if the samples come from a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{K}$. Even if the model is correct, the parameters are usually unknown so they are typically replaced by ML estimates.

The problem in microarrays is that p, the number of features (genes), is very large compared to n, the number of samples. This makes the covariance estimation unreliable and the LDA procedure unfeasible. One solution to this problem is to assume a diagonal covariance matrix, i.e. the DLDA model [5]. Under this model the discriminant function (1) can be rewritten as:

$$g(\mathbf{x}) = \sum_{i=0}^{p}\left(\frac{\hat{\mu}_A(x_i) - \hat{\mu}_B(x_i)}{\hat{\sigma}(x_i)}\right)\left(\frac{x_i - \hat{\mu}(x_i)}{\hat{\sigma}(x_i)}\right) \qquad (3)$$

This simplification is by itself not enough since many genes are irrelevant for the classification and add noise to (3). This motivates Feature Subset Selection (FSS), which uses only a small fraction of the initial set of p genes. This FSS can be done with two basic approaches, namely using a filter or a wrapper [6].

Nearly all DLDA based techniques [4][5][8] use the filter approach for FSS. That is, genes are first ranked using a statistical score, and then the discriminant function is built by selecting the highest ranking genes.

2. Sequential DLDA approach

The SeqDLDA approach can be seen as a wrapper FSS approach [6]. In contrast to filter DLDA [5], the discriminant function (4) is built by searching the subset of genes $\mathcal{S}_l$ that better improves class separation:

$$g_l(\mathbf{x}) = \log\left(\frac{\hat{p}_A}{\hat{p}_B}\right) + \sum_{i\in\mathcal{S}_l} H(x_i)\left(\frac{x_i - \hat{\mu}(x_i)}{\hat{\sigma}(x_i) + \hat{\sigma}_0}\right) \qquad (4)$$

$$H(x) = \frac{\hat{\mu}_A(x) - \hat{\mu}_B(x)}{\hat{\sigma}(x) + \hat{\sigma}_0} \qquad \hat{\sigma}_0 = \underset{i=1..p}{\text{median}}\left(\hat{\sigma}(x_i)\right) \qquad (5)$$

The class separation is measured by $H(g(\mathbf{x}))$ (5) after computing the discriminant. This score is a robust modification of the t-test and is also used in NSC [4]. The additional term in the denominator protects against an unusually low $\hat{\sigma}(x)$ produced by chance.

Instead of measuring $H(g(\mathbf{x}))$ for all possible combinations of features, we use a greedy search described in [6] as Forward-Selection/Hill-Climbing. Starting from an empty set of features, at every iteration, we add the one gene that most increases $H(g(\mathbf{x}))$.

The difference of our procedure with [6] is that the evaluation of the classifier is done with a statistical test from the same training data instead of using crossvalidation. This was also done in [3] but using a regular t-test which makes the model and the exploratory search not robust resulting in a much lower performance.

3. Results

The proposed algorithm has been evaluated using 100 runs of 10-fold Cross-Validation on several 2-class datasets Table 1). The **leukemia** [8] (n=72, p=7129), **colon** [7] (n=22, p=2000), **prostate** [9] (n=102, p=6033) datasets are publicly available and widely used in other studies. The **neuroblastoma** dataset (n=102, p=44298) consists of samples from Neuroblastoma stage 4 with MYCN not amplified obtained at diagnosis (manuscript in preparation). In all cases, the gene expression has been normalized by clipping values lower than 1 and taking a log-transform.

Using the same evaluation methods, the proposed SeqDLDA approach has been compared to DLDA [5], NSC [4], GP-DLDA [3], ULDA [10] and Linear SVM.

Table 1 Average Cross-validation Error, number of selected genes, and standard deviation (SD).

	Leukemia	Colon	Prostate	Neuro-blastoma
Seq-DLDA	4.11%,180 (1.32%)	12.06%,50 (1.87%)	**5.53%, 26 (0.90%)**	**13.87%,70 (2.41%)**
GP-DLDA	3.82%, 18 (0.77%)	13.08%,16 (1.76%)	6.44%, 20 (0.70%)	15.77%,35 (1.61%)
DLDA	3.38%, 7 (1.30%)	12.40%, 3 (1.44%)	6.99%, 2 (0.33%)	16.91%,55 (1.54%)
NSC	4.18%, 70 (0.80%)	**10.31, 20 (1.02%)**	7.65%, 6 (0.42%)	17.98%,70 (1.67%)
ULDA	3.39%, p (0.747%)	15.19%, p (2.72%)	8.53%, p (1.10%)	**13.42%, p (1.55%)**
Lin-SVM	**2.61%, p (0.57%)**	15.39%, p (2.17%)	8.01%, p (1.14%)	14.13%, p (1.45%)

4. Discussion and conclusions

In the studied datasets SeqDLDA obtains results very close to the best approach, and the best results for the **prostate** and **neuroblastoma** datasets. Additionally Seq-DLDA performs gene selection which is not the case of ULDA and SVM whose classifier uses the whole set of genes. Gene selection is crucial in order to identify genomic targets that may explain the disease.

Classical DLDA filtering approaches ([5], [4]) provide similar results in the absence of gene correlations or inter-pair correlations in GP-DLDA [3]. However correlation among genes is generally present and the SeqDLDA method will allow us to choose genes that may have a lower score (under a diagonal correlation assumption) but can be shown to provide better classification performance when combined with the already selected genes. Additionally, we have also noticed that improvement in performance over DLDA is more noticeable when a larger number of training samples is available.

Finally, Nearest Shrunken Centroid (NSC) obtains the best results in the **colon** dataset. This probably comes from the denoising effect of shrinking which could also be incorporated in our procedure.

5. Acknowledgments

We acknowledge the Children's Oncology Group (COG) for providing the **neuroblastoma** dataset.

6. References

[1] RO Duda, PE Hart, DG Stork, *Pattern Classification* (2nd Edition), Wiley-Interscience, 2000

[2] T Hastie, R Tibshirani, JH Friedman, *The elements of statistical learning: data mining, inference, and prediction*, New York: Springer, 2001

[3] TH Bø, I Jonassen, "New feature subset selection procedures for classification of expression profiles", *Genome Biology*, 2002, 3:research17.1-17.11

[4] R Tibshirani, T Hastie, B Narasimhan, G Chu, "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression", *Proc. Nat'l Academy of Science PNAS*, 2002, vol. 99, no. 10, pp. 6567-6572

[5] S Dudoit, J Fridlyand, TP Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data", *Journal of the American Statistical Association*, 2002, Vol. 97, No.457, pp. 77-87

[6] R Kohavi, GH Jhon, "Wrappers for Feature Subset Selection", *Artificial Intelligence Journal*, 1997, 97:273-324

[7] U Alon, et al., "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proc. Nat'l Academy of Science PNAS*, 1999, vol. 96, pp. 6745-6750.

[8] TR Golub, et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 1999, vol. 286, pp. 531-537

[9] Singh D, et al, "Gene Expression Correlates of Clinical Prostate Cancer Behavior",*Cancer Cell*,2002 Mar;1(2):203-9

[10] J Ye, T Li, T Xiong, R Janardan, "Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene Expression Data", *IEEE/ACM TCBB,* 2004, Vol. 1, No. 4, pp. 181-190.