Diagonal Discriminant Analysis for Gene-Expression Based Tumor Classification

Gokmen Zararsiz, Selcuk Korkmaz, and Dincer Goksuluk
Department of Biostatistics, Hacettepe University, Ankara, Turkey
Email: gokmenzararsiz@hotmail.com, {selcuk.korkmaz, dincer.goksuluk}@hacettepe.edu.tr

Vahap Eldem
Department of Biology, Istanbul University, Istanbul, Turkey
Email: eldemvahap@gmail.com

Ahmet Ozturk
Department of Biostatistics, Erciyes University, Kayseri, Turkey
Email: ahmets67@hotmail.com

***Abstract*— A reliable and accurate tumor classification is crucial for successful diagnosis and treatment of cancer diseases. With the recent advances in molecular genetics, it is possible to measure the expression levels of thousands of genes simultaneously. Thus, it is feasible to have a complete understanding the molecular markers among tumors and make a more successful and accurate diagnosis. A common approach in statistics for classification is linear and quadratic discriminant analysis. However, the number of genes (p) is much more than the number of tissue samples (n) in gene expression datasets. This leads to data having singular covariance matrices and limits the use of these methods. Diagonal linear and diagonal quadratic discriminant analyses are more recent approaches that ignore the correlation among genes and allow high-dimensional classification. Nearest shrunken centroids algorithm is an updated version of diagonal discriminant analysis, which also selects the genes that mostly contributed in class prediction. In this study we will discuss these algorithms and demonstrate their use both in microarray and RNA sequencing datasets.**

***Index Terms*—classification, discriminant analysis, gene expression, RNA sequencing, tumor classification**

I. Introduction

Gene expression is a measure that used for a gene in the synthesis of a functional gene product. It is very important to understand the function of many biological systems. Currently, microarrays and RNA-Sequencing (RNA-Seq) are the most capable technologies to extract gene expression. Due to some advantages as producing less noisy data and detecting novel transcripts and isoforms, RNA-Seq is accepted as a more efficient technique and more widely used at this moment. Using any of these two technologies, it is possible to obtain the expression values of thousands of genes simultaneously. After some bioinformatics data pre-processing, we obtain a dimensional gene expression matrix (*p*: number of genes, *n*: number of tissue samples) from both microarray and RNA-Seq data [1]-[4].

One major task using gene expression data is tumor classification. Conventional methods are subjective and classify tumors with examining morphological images under microscope. The success of tumor classification is directly correlated with the experience of pathologists. Microarray and RNA-Seq technologies make this process objective and produce reliable and accurate results based on the used statistical methods for the successful diagnosis and treatment of cancer diseases. Gene-expression data have become a standard tool for biomedical studies and currently, it is widely collected from patients in clinical trials. Successful classifications can assist physicians for accurate diagnosis and identify the right treatment for patients [5], [6].

The problem here is the high-dimension of gene expression data. If we were interested with only one gene, we could simply apply ROC analysis and define a cut-off value to classify tumors. If we had a low-dimensional data, we could apply discriminant analysis, logistic regression or other statistical algorithms. The data become more and more complex when the number of dimension increases, and we cannot directly use these statistical methods in order to classify our data. Number of genes is much more than the number of tissue samples in gene-expression data and this leads to the 'curse of dimensionality' problem. Here, we meet with singular matrices and cannot calculate the inverses of matrices. Thus, we cannot assign samples to their correct classes [7].

In this study we will discuss diagonal discriminant analysis and nearest shrunken centroids algorithms which are extensions of Fisher's discriminant analysis and developed for microarray based gene-expression classification. We will also demonstrate the use of these methods for RNA-Seq data for the purpose of tumor

Manuscript received November 6, 2014; revised May 5, 2015.

doi: 10.12720/jait.6.2.59-62

classification. We will demonstrate the use of these methods on publicly available real datasets.

II. Diagonal Discriminant Analysis

In statistical decision theory, we need the posterior tumor class probabilities $P(C|X)$ for optimal classification. Let $f_k(x)$ the class conditional density of X and π_k the prior probability (mostly $\hat{\pi}_k = n_k/n$) for class $C = k$, where $\sum_{k=1}^{K} \pi_k = 1$. Using Bayes theorem, we can obtain $P(C|X)$ as follows:

$$P(C = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l} \quad (1)$$

Various solutions are available to model $f_k(x)$:

- Gaussian densities for linear and quadratic discriminant analysis,
- Flexible mixture Gaussian densities for nonlinear classification,
- Nonparametric density estimates for each class for the most flexibility,
- Naïve Bayes models where the class densities are the products of marginal densities.

Due to the curse of dimensionality problem in high dimensional gene expression data, Naïve Bayes models can be used to estimate conditional densities of tumor classes and covariates can be assumed as independent. This can be provided by using diagonal covariance matrices $\hat{\Sigma}_{C=k} = diag(\sigma_{1k}^2, \sigma_{2k}^2, \ldots, \sigma_{pk}^2)$, where all off-diagonal elements are set to be zero [6,8]. This is also called as 'independence rule' and Bicken *et al.* [9] showed theoretically that diagonal discriminant analysis performs better than traditional discriminant analysis in high-dimensional classification analysis. Using diagonal covariance matrices, we obtain the following discriminant rule for class *k*:

$$\delta_k^Q(x^*) = -\sum_{j=1}^{p} \frac{\left(x_j^* - \bar{x}_{kj}\right)^2}{s_{kj}^2} - \sum_{j=1}^{p} \log(s_{kj}^2) + 2\log(\pi_k) \quad (2)$$

here, this rule is called as diagonal quadratic discriminant analysis (DQDA). x_j^* refers to a vector of test observations values. $\bar{x}_{kj}$ and s_{kj}^2 are the mean and variance statistics of j^{th} feature (gene or transcript in gene expression data) in class *k*, respectively. A new test object will be assigned to the class which maximizes the $\delta_k(x^*)$ discriminating function.

The second rule, diagonal linear discriminant analysis (DLDA) assumes that covariance matrices are equal across groups. Again, a test observation is assigned to a class which maximizes the following discriminating function:

$$\delta_k^L(x^*) = -\sum_{j=1}^{p} \frac{\left(x_j^* - \bar{x}_{kj}\right)^2}{s_j^2} + 2\log(\pi_k) \quad (3)$$

here, s_j^2 within-class variances are used instead of s_{kj}^2.

III. Nearest Shrunken Centroids

Both DLDA and DQDA algorithms are capable for classification in $n<p$ setting. However, one problem is to obtain very complex models for high-dimensional data. It is also crucial to determine the genes which contribute most to class prediction. For this purpose, Tibshirani *et al.* [5] proposed nearest shrunken centroids (NSC) sparse classification algorithm. NSC basically selects the most significant gene subsets for more simple and interpretable results and uses them for class prediction.

NSC approximates the standardized class gene expression means to the standardized overall gene expression means, then eliminates the approximated genes and builds a classification model with the remaining genes. Let d_{kj} the difference scores, which can also be considered as the *t* statistic as a difference between a classes mean expression and overall mean expression:

$$d_{kj} = \frac{\bar{x}_{kj} - \bar{x}_j}{m_k(s_j + s_0)} \quad (4)$$

We can simply call mean gene expressions as centroids. Here, s_0 is a positive constant, mostly the median value of s_j over the set of genes. m_k is a standard error correction term set as $\sqrt{1/n_k + 1/n}$ .We can rewrite (4) as:

$$\bar{x}_{kj} = \bar{x}_j + m_k(s_j + s_0)d_{kj} \quad (5)$$

Next, each d_{kj} is shrunk to zero and shrunken centroids can be written as follows:

$$\bar{x}'_{kj} = \bar{x}_j + m_k\left(s_j + s_0\right)d'_{kj} \quad (6)$$

Due to its more reliable mean estimates, the commonly used shrinkage here is soft-thresholding. An alternative method here is the hard-thresholding, however it is less widely used due to its less reliable mean estimates [10,11]. Each d_{kj} is shrank by an amount λ(shrinkage parameter) and set to zero if its absolute value is negative:

$$d'_{kj} = sign(d_{kj})max(\left|d_{kj}\right| - \lambda, 0) \quad (7)$$

Genes with zero shrunken differences for all classes *k* are eliminated and the classification is made with the remaining genes. Cross-validation is used to identify the optimal shrinkage parameter λ. For a range of λ values, optimal λ is the one that gives the minimum classification error. For each λ, DLDA is used as a classification algorithm and the active genes that mostly contributed to the class prediction can be identified based on the optimal classification model. A very important point here is the usage of shrunken centroids rather than the simple centroids in classification.

A test observation is assigned to the class that maximizes the following NSC discriminating function:

$$\delta_k^N(x^*) = -\frac{1}{2}\sum_{j=1}^{p} \frac{(x_j^* - \bar{x}_{kj})^2}{(s_j + s_0)^2} + \log(\pi_k) \quad (8)$$

Posterior tumor class probabilities $P(C|X)$ for both diagonal discriminant analysis and nearest shrunken centroids can be obtained as follows:

$$\hat{p}_k(x^*) = \frac{e^{\delta_k(x^*)/2}}{\sum_{l=1}^{K} e^{\delta_l(x^*)/2}} \quad (9)$$

IV. EXPERIMENTS

Experimental Datasets: We used two real datasets to demonstrate the use of these algorithms in microarray and RNA-Seq data classification. The first data is the small round blue cell tumor of childhood (SRBCT) [12]. Gene expressions were obtained with cDNA microarray experiment. Probe labeling, hybridization and image acquisition were conducted based on the National Human Genome Research Institute protocol. The data consists of the expression values of 2,308 genes belonging to 83 tissue samples.Khan *et al.* [12] provided 63 training and 25 test samples, in which 5 of the samples are non-SRBCT and not considered here. Training samples contain 23 Ewing family of tumors (EWS), 20 rhabdomyosarcoma (RMS), 12 neuroblastoma (NB) and 8Burkitt lymphomas (BL).Test samples contain 6 EWS, 5 RMS, 6 NB and 3BL.

Second data is the cervical cancer miRNAdata [13]. This is an RNA-Seq dataset that contains the expression values of miRNAs in tumor and non-tumor human cervical tissue samples. Cervical data includes 714 mapped miRNA read counts to human reference genome. It contains 58 samples, where 29 of them are tumor and the remaining 29 are non-tumor. In tumor samples, 6 of them are adenocarcinomas (ADC), 21 are squamus cell carcinomas (SCC) and 2 are unclassified. We considered the data as a two-class problem and 20 samples in each class were randomly defined as training samples, where the remaining 9 samples were defined as test samples.

Model Building: We applied some pre-processing analysis to cervical data to get a continuous gene expression data required by the used classifiers. Firstly, we filtered the miRNA's, which 10% or fewer training observations have non-zero counts. We also selected 500 miRNA's with highest variance to get more reliable results. After, we applied deseq normalization [14] to adjust the read counts to sample specific differences. Then, we applied regularized logarithmic transformation (rlog) to estimate the mean and variance relationship of the data and transform it based on this relationship to obtain expression data that is hierarchically closer to microarrays [15]. For SRBCT data, we selected the 2,000 genes that have the highest variances. Finally, we obtained 2,000x63 training, 2,000x20 test SRBCT gene expression data, and 500x40 training, 500x18 test cervical cancer gene expression data.

In order to avoid over-fitting, we made a grid search for the tuning parameters and used 5-fold cross validation to identify the optimal parameters of each model. All classifiers were fit with these optimal parameters. As an evaluation criterion, we considered the model accuracy. For this purpose, we calculated misclassification errors of test datasets for each model. Since, both data have very small sample size and model accuracies may vary based on the selected training and test sets; we fit the process ten times and averaged the misclassification errors. For comparison, support vector machines (SVM) and random forests (RF) were also considered and the same model building process was also applied for them. Number of trees in RF algorithm was set as 500. Radial-based kernel function was used in SVM modeling. Parameter optimization for SVM and RF was conducted in caret package [16] of R.

V. RESULTS

Results are given in Table I. NSC identified 43 genes and 12 miRNA's with optimal parameters 3.859 and 3.416 for SRBCT and cervical datasets, respectively (Fig.1).For RF algorithm, number of genes sampled at each split was 67 and 206; for SVM algorithm complexity parameter was 1, sigma parameter was 0.00023 and 0.0016, for SRBCT and cervical datasets, respectively.

TABLE I. PERFORMANCE OF CLASSIFIERS IN REAL GENE EXPRESSION DATASETS

Classifier	Misclassification error
SRBCT dataset	
DLDA	4.87
DQDA	8.36
NSC	1.54
SVM	5.15
RF	1.67
Cervical dataset	
DLDA	10.30
DQDA	8.64
NSC	6.97
SVM	13.79
RF	7.12

DLDA: Diagonal linear discriminant analysis, DQDA: Diagonal quadratic discriminant analysis, NSC: Nearest shrunken centroids, SVM: Support vector machines, RF: Random forests.

NSC algorithm outperformed other algorithms in both datasets. Performances of diagonal discriminant analysis were compatible with SVM algorithm, but less than RF algorithm. When compared to each other, DLDA performed better in SRBCT dataset, while DQDA performed better in cervical dataset.

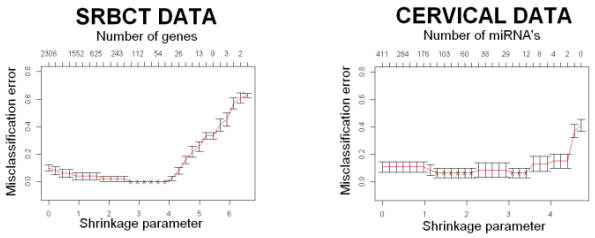


Figure 1. Identification of shrinkage parameters for NSC algorithm

VI. CONCLUSIONS

Diagonal discriminant analyses performed remarkably well in two gene expression applications. Their results were compatible with the more sophisticated classifiers SVM and RF. In fact, NSC gave the highest accuracy in both datasets. These algorithms are very easy to implement, give accurate results and unlock the use of discriminant analysis classifiers in high-dimensional settings. Thus, they should be considered as a method of choice in high-dimensional tumor classification problems.

As an advantage, NSC algorithm selects the most significant gene subset on class prediction. Identifying the most informative genes may provide potential molecular markers for tumor classification. Instead of classification, this algorithm can also be used in biomarker discovery problems. Another advantage of this algorithm is to use for clustering purpose as shown in [5]. After gene selection, one can use the shrunken differences d'_{kj} and detect the co-regulated gene clusters that are significant on class prediction.

We implemented these algorithms for RNA-Seq based gene-expression classification for the first time. With some data pre-processing, one can transform the data and make it hierarchically closer to microarrays. In this setting, normalization and transformation may have significant effect on classification of RNA-Seq data. Since this study aimed to demonstrate the application of these algorithms rather than aiming to compare their performances, we simply applied deseq normalization and rlog transformation. A comprehensive study is required to compare the performances of these algorithms under different scenarios.

Users can easily implement DLDA and DQDA in **sfsmisc** package [17], NSC in **pamr** package [18] of R software (www.r-project.org).

ACKNOWLEDGMENT

This work was supported by the Research Fund of Erciyes University [TDK-2015-5468].

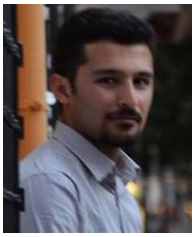
REFERENCES

[1] Z. Wang, *et al.* "RNA-Seq: A revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, pp. 57-63, 2009.
[2] J. C. Marioni, *et al.* "RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays," *Genome Res.*, vol. 18, no. 9, pp. 1509-1517, 2008.
[3] G. Zararsiz, *et al.* "Classification of RNA-Seq data via bagging support vector machines," *bioRxiv*, 2014.
[4] G. Zararsiz, *et al.* "MLSeq: Machine learning interface for RNA-Seq data," *R Package Version 1.1.6.*, 2014.
[5] R. Tibshirani, *et al.* "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *PNAS*, pp. 6567-6572, 2002.
[6] H. Pang, *et al.* "Recent advances in discriminant analysis for high-dimensional data classification," *J. Biomet. Biostat.*, vol. 3, no. 1, 2012.
[7] R. Bellman, *Dynamic Programming*, Princeton Univ. Press, 1957.
[8] S. Dudoit, *et al.* "Comparison of discrimination methods for the classification of tumors using gene expression data," *JASA*, vol. 97, no. 457, pp. 77-87, 2002.
[9] P. J. Bickel, *et al.* "Some theory of Fisher's linear discriminant function,` naive Bayes', and some alternatives when there are many more variables than observations," *Bernoulli*, vol. 10, pp. 989-1010, 2004.
[10] T. Hastie, *et al. The Elements of Statistical Learning*; *Data Mining, Inference and Prediction*, Springer, New York, 2001.
[11] R. Tibshirani, *et al.* "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc. B.*, vol. 58, no. 1, pp. 267-288, 1996.
[12] J. Khan, *et al.* "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nat. Med.*, vol. 7, pp. 673–679, 2001.
[13] D. M. Witten, *et al.* "Ultra-high throughput sequencing-based small RNA discovery an discrete statistical biomarker analysis in a collection of cervical tumors and matched controls," *BMC Biology*, vol. 8, no. 58, 2010.
[14] S. Anders, *et al.* "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, R106, 2010.
[15] M. I. Love, *et al.* "Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2," *bioRxiv*, 2014.
[16] M. Kuhn "Building predictive models in R using the caret package," *Journal of Statistical Software*, vol. 28, no. 5, 2008.
[17] M. Maechler *et al.* sfsmisc: Utilities from seminar fuer statistik ETH Zurich. *R package version 1.0-26.* [Online]. Available: http://CRAN.R-project.org/package=sfsmisc
[18] T. Hastie, *et al.* (2014). pamr: Pam: prediction analysis for microarrays. R package version 1.54.1. [Online]. Available: http://CRAN.R-project.org/package=pamr

Gökmen Zararsız is a research assistant in Biostatistics department of Hacettepe University. His main working areas include next-generation sequencing data analysis, machine learning, genomics and transcriptomics. With the other co-authors of this study, S.K. and D.G., they started the 'BIOSOFT Project' to develop algorithms and free, user-friendly, easy-to-use, up-to-date and comprehensive tools in the area of biostatistics and bioinformatics. Some of these tools can be reached at http://www.biosoft.hacettepe.edu.tr/ . G.Z. currently works on developing novel applications for RNA-Seq classification.

Vahap Eldem is a research assistant in Biology department of Istanbul University. For more than 5 years, V.E. has been working for computational biology and bioinformatics. His other research interests include marine genomics, small RNAs and phylogenetics. V.E. currently works on de novo transcriptome assembly analysis of European anchovy (Engraulis encrasicolus).

Selcuk Korkmaz is a research assistant in Biostatistics department of Hacettepe University. His working area and expertise is mostly related with proteomics, drug discovery, virtual screening and machine learning. S.K. currently works in University of California, San Diego Supercomputing Center with the project on developing new approaches for the prediction of biological assemblies and protein-protein interfaces.



Dincer Goksuluk is a research assistant in Biostatistics department of Hacettepe University. His research mostly focuses on software engineering, machine learning, transcriptomics, proteomics and metabolomics. D.G. currently works for developing novel applications for various problems in transcriptomics field.

Ahmet Ozturk is an associate professor in Biostatistics department of Erciyes University. A.O. has great expertise and more than a hundred papers in the field of biostatistics. His expertise area includes biostatistics, pediatrics, epidemiology, machine learning and clinical trials.