Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Pang and Tong, J Biomet Biostat 2012, 3:1
DOI: 10.4172/2155-6180.1000e106

Editorial Open Access

Recent Advances in Discriminant Analysis for High-dimensional Data Classification

Herbert Pang[1]* and Tiejun Tong[2]*

[1]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA
[2]Department of Mathematics, Hong Kong Baptist University, Hong Kong, China

Introduction

There are serious challenges posed by high-dimensional data sets. With the arrival of new technologies, high-throughput modeling is becoming a norm in many disciplines such as statistical genetics, epidemiology, astronomy, high energy physics, and ecology. High-dimensional data have emerged from various sources such as digital images, documents, next-gen sequencing, mass spectrometry, metabolomics, microarray, proteomics, online videos and web pages. One area with a growing need for new statistical methods and theory for high-dimensional data is the classification of subgroups. For example, cancer classification has primarily been based on histopathological appearance of tumor. However, patients with similar tumor appearance can have different prognosis and response to treatment. The traditional way to classify cancer by pathological review may cause biased results and misclassify the tumor subtypes for patients. The availability of microarray data allows simultaneous measures of thousands of genes. These high-dimensional data have become a standard tool for biomedical studies and are now commonly collected from patients in clinical trials. The identification of informative genes may result in potential molecular markers for tumor class prediction. Correct classifications can help practitioners identify the right treatment for patients. Due to the cost and/or experimental difficulties in obtaining sufficient biological materials, it is common to see studies with sample size much smaller than the number of dimensions. These problems are referred to as "large p small n" issues, where p is the number of dimensions (or say genes) and n is the sample size. High-dimensional data pose challenges to traditional statistical methods. For instance, owing to small n, there are increased uncertainties in the standard estimations of parameters such as means and variances. As a consequence, statistical analyses based on such parameters estimation are usually unreliable. To have improved parameters estimation, researchers have come up with innovative ways to deal with this.

A common approach for the analysis of high-dimensional data classification is discriminant analysis. The main goal of discriminant analysis is to assign an unknown subject to one of K classes on the basis of observed subjects from each class. Let $X_{k,1},\ldots, X_{k,n_k}$ be independent and identically distributed from p-dimensional multivariate normal distribution with mean vector μ_k and covariance matrix Σ_k for class k=1,..,K. Let $n = n_1+\ldots+n_k$ be the total number of observations. Note that the sample covariance matrices are singular when p is larger than n. Therefore, traditional methods such as Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are not applicable to high-dimensional data classification directly.

Recent Advances

To overcome the singularity problem, Dudoit [1] introduced two simplified discriminant rules by assuming independence between covariates. For each class k, let $\bar{X}_k=(\bar{X}_{1k},\ldots,\bar{X}_{pk})^T=\sum_{j=1}^{n_k} X_{k,j}/n_k$ be the sample mean, and $\hat{\Sigma}_k = diag(\hat{\sigma}^2_{1k},\ldots,\hat{\sigma}^2_{pk})$ be the sample covariance matrix where the off-diagonal elements are all set to be zero. Also let $\hat{\pi}_k = n_k/n$ be the estimated prior probability of observing a class k subject. The first rule developed in Dudoit [1] is called Diagonal Quadratic Discriminant Analysis (DQDA). It classifies a new subject X to class k that minimizes the discriminant score

$$\hat{d}^Q_k(X) = \sum_{i=1}^{p}(x_i-\bar{X}_{ik})^2/\hat{\sigma}^2_{ik} + \sum_{i=1}^{p}\log\hat{\sigma}^2_{ik} - 2\log\hat{\pi}_k$$

The second rule is called Diagonal Linear Discriminant Analysis (DLDA) that classifies the new subject analogously according to the discriminant score $\hat{d}^L_k(X) = \sum_{i=1}^{p}(x_i-\bar{X}_{ik})^2/\hat{\sigma}^2_i - 2\log\hat{\pi}_k$, where $\hat{\sigma}^2_i$ are the pooled variances across the K classes. DQDA and DLDA classifiers are sometimes called "naive Bayes" classifiers because they can arise in a Bayesian setting [2]. Due to the small sample size, DLDA and DQDA, which ignore correlations between genes, perform remarkably well compared to some more sophisticated classifiers in terms of both accuracy and stability. In addition, DQDA and DLDA are easy to implement and have been adopted to analyze high-dimensional data in various fields of science.

Though DQDA and DLDA work for small sample sizes and perform better than some sophisticated classifiers, their performance under the "large p small n" setting is still unreliable due to various reasons. In this section, we review some significant results that have been developed in the literature to improve the diagonal discriminant analysis.

The Nearest Shunken Centroid (NSC) method proposed by Tibshirani [3] is among the first to improve the diagonal discriminant analysis. This method also assumes a diagonal covariance matrix. To improve the classification performance, the mean vector μ_k is estimated by the "shrunken centroid" rather than the sample mean. NSC shrinks each class centroid toward the overall centroid by a certain amount. Specifically, let $d_{ik} = (\bar{x}_{ik}-\bar{x}_i)/(m_k s_i)$ be the standardized distance between each class centroid and the overall centroid, where $m_k = \sqrt{1/n_k - 1/n}$, s_i is the pooled within-class standard deviation for the i^{th} component and s_0 is a positive constant with the same value for all genes. By shrinking d_{ik} toward zero via soft thresholding or hard thresholding, the NSC method uses the achieved shrunken centroids to perform DLDA and then classifies the new subject to the class with nearest shrunken centroid. Note that other variations of NSC are also available in the literature; see for example [4,5].

Uncorrelated discriminant analysis (UDA) is another extension of the diagonal discriminant analysis [6]. Let S_b be the between-class

***Corresponding authors:** Herbert Pang, Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA, E-mail: herbert.pang@duke.edu

Tiejun Tong, Department of Mathematics, Hong Kong Baptist University, Hong Kong, China, E-mail: tongt@hkbu.edu.hk

Received Januray 08, 2011; **Accepted** Januray 08, 2012; **Published** Januray 10, 2012

Citation: Pang H, Tong T (2012) Recent Advances in Discriminant Analysis for High-dimensional Data Classification. J Biomet Biostat 3:e106. doi:10.4172/2155-6180.1000e106

scatter matrix, and S_w be the within-class scatter matrix, and $S_t = S_b + S_w$ be the total scatter matrix. A special property of UDA is that the genes in the transformed space are uncorrelated. The goal of UDA is to find the optimal discriminant vectors that are S_t-orthogonal. Suppose r vectors are obtained, then the (r+1)th vector will be the one that maximizes the Fisher criterion function subject to certain constraints. Ye et al. [6] showed that this can be solved efficiently by solving an optimization problem.

Due to the small sample sizes, another direction to improve the diagonal discriminant analysis is by shrinkage [7,8]. For instance, Pang [7] applied the shrinkage estimates of variances in Tong [9] into the diagonal discriminant scores, and formed two shrinkage-based rules called Shrinkage-based DQDA (SDQDA) and Shrinkage-based DLDA (SDLDA). Pang [7] also applied regularization as in Friedman [10] to further improve the performance of SDQDA and SDLDA. Combining shrinkage-based variances in diagonal discriminant analysis and regularization in a new classification scheme showed improvement over the original DQDA and DLDA, Support Vector Machine, and k-Nearest Neighbors in many scenarios. In addition, Pang H [11] have applied the shrinkage-based discriminant rules to identify genes that help differentiate between estrogen receptor positive and negative samples to investigate genes that are specific to the African American subjects with breast cancer.

Recently, Huang S [11] observed that the diagonal discriminant analysis suffers from serious drawback of having biased discriminant scores. Inspired by this, they proposed bias-corrected diagonal discriminant rules by using unbiased estimates of $\hat{d}_k^Q(X)$ and $\hat{d}_k^L(X)$. Specifically for DQDA, let $\hat{d}_k^Q(X) = \hat{L}_{k1} + \hat{L}_{k2} - 2\log\hat{\pi}_k$, where $\hat{L}_{k1} = \sum_{i=1}^{p}(x_i - \bar{X}_{ik})^2 / \hat{\sigma}_{ik}^2$ and $\hat{L}_{k2} = \sum_{i=1}^{p}\log\hat{\sigma}_{ik}^2$. Huang S [12] observed that $\hat{L}_{k1}$ and $\hat{L}_{k2}$ are biased estimates of the true quantities. Let $\breve{L}_{k1}$ and $\breve{L}_{k2}$ be the bias-corrected estimators. The resulting bias-corrected discriminant score of DQDA is then defined as $\breve{d}_k^Q(X) = \breve{L}_{k1} + \breve{L}_{k2} - 2\log\hat{\pi}_k$. It was shown that the proposed bias-corrected score improves the standard one under the quadratic loss function. Finally, both simulation study and prediction accuracy analysis demonstrated the superiority of bias correction over the original rules, especially when the design is highly unbalanced.

Discussion

Though the progress made thus far is encouraging, we believe that more needs to be done given the increased demand and further improvement are desired. First, note that genes are unlikely to be independent of each other. Therefore, the assumptions made in the diagonal discriminant analysis and its variations may not be realistic. Pang H [13] are studying and extending block-diagonal discriminant analysis methods. In some preliminary study, they have made further improvement possible for class prediction in real data analysis. Second, the performance of the NSC method and its variations may not be satisfactory when the sample size is small due to the large variation in variable selection using cross-validation. In Tong T [14], the authors are proposing a new algorithm that chooses the tuning parameter for variable section by minimizing certain risk functions. Some preliminary simulations indicate that the proposed algorithm performs well compared to the original NSC method by cross-validation when the sample size is small. Third, we can consider the bias-corrected rules for SDQDA and SDLDA. Recall that the shrinkage estimation is to trade off a "small" increase in bias for a possible "significant decrease" in variance. The good performance of SDQDA and SDLDA in Huang S [12] is mainly owing to the largely reduced variance in the shrinkage-based discriminant scores. Instead, the bias term in SDLDA and SDQDA still remains or may be even larger than that in DLDA and DQDA, respectively, as shrinkage may pull in extra bias. To conclude, we reiterate that there is still room for more innovative methdological developments in the area of discriminant analysis for high-dimensional data classification.

References

1. Dudoit S, Fridlyand J, Speed TP (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 97: 77-87.
2. Bickel PJ, Levina E (2004) Some theory of Fisher's linear discriminant function,`naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli 10: 989-1010.
3. Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS 99: 6567-6572.
4. Dabney AR (2005) Classification of microarrays to nearest centroids. Bioinformatics 21: 4148-4154.
5. Wang S, Zhu J (2007) Improved centroids estimation for the nearest shrunken centroid classifier. Bioinformatics 23: 972-979.
6. Ye J, Li T, Xiong T, Janardan R (2004) Using uncorrelated discriminant analysis for tissue classification with gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics 1: 181-190.
7. Pang H, Tong T, Zhao H (2009) Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data. Biometrics 65: 1021-1029.
8. Tong T, Chen L, Zhao H (2011) Improved mean estimation and its application to diagonal discriminant analysis. Bioinformatics [Ahead of Print].
9. Tong T, Wang Y (2007) Optimal shrinkage estimation of variances with applications to microarray data analysis. J Am Stat Assoc 102: 113-122.
10. Friedman JH (1989) Regularized discriminant analysis. J Am Stat Assoc 84: 165-175.
11. Pang H, Ebisu K, Watanabe E, Sue LY, Tong T (2010) Analyzing breast cancer microarrays of African Americans using shrinkage-based discriminant analysis. Human Genomics 5: 5-16.
12. Huang S, Tong T, Zhao H (2010) Bias-corrected diagonal discriminant rules for high-dimensional classification. Biometrics 66: 1096-1106.
13. Pang H, Tong T, Ng MK (2012) Block-diagonal discriminant analysis and its improvement.
14. Tong T, Li G, Peng H, Pang H (2012) Optimal nearest shrunken centroids method for high-dimensional data classification.