Pattern Recognition 48 (2015) 2676–2684

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Generalized quadratic discriminant analysis

Smarajit Bose [a], Amita Pal [a,*], Rita SahaRay [a], Jitadeepa Nayak [b]

[a] Interdisciplinary Statistical Research Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700 108, India
[b] Indian Institute of Management, Kolkata, India

ARTICLE INFO

Article history:
Received 6 February 2014
Received in revised form
4 August 2014
Accepted 18 February 2015
Available online 6 March 2015

Keywords:
Linear discriminant analysis
Mahalanobis distance
Elliptically symmetric distributions

ABSTRACT

In linear discriminant analysis, the assumption of equality of the dispersion matrices of different classes leads to a classification rule based on minimum Mahalanobis distance from the class centres. However, without this assumption, the resulting quadratic discriminant classifier involves, in addition to the Mahalanobis distances, the ratio of the determinants of the dispersion matrices as a factor. In fact, it has been observed that, for discriminating between populations with underlying elliptically symmetric distributions, such classifiers also incorporate similar factors, apart from the Mahalanobis distances.

In this paper, a nonparametric classification technique which generalizes discriminant analysis has been proposed. The method of cross-validation is used to make the technique adaptive to a given dataset. An extensive simulation study is presented to illustrate the potential of the method. Finally, through implementation on a number of real-life data sets, it has been demonstrated that the proposed generalized quadratic discriminant analysis (GQDA) compares very favourably with other nonparametric methods, and is computationally cost-effective.

1. Introduction

In traditional linear discriminant analysis, the class-conditional probability densities are assumed to be multivariate normal distributions having different mean vectors for different classes. The assumption that the dispersion matrices are identical for all classes leads to linear discriminant analysis (LDA). The resulting classification rule assigns an observation to the class for which the Mahalanobis distance between the observation and the class mean is minimum. This classification rule will henceforth be referred to as the minimum Mahalanobis distance (MMD) classification rule. When it is not possible to assume the class dispersion matrices to be equal, the result is quadratic discriminant analysis (QDA). The classification rule under QDA is quite different from the MMD rule as it also involves a factor based on the ratio of the determinants of the dispersion matrices. Anderson [1] provides an excellent introduction to discriminant analysis.

Despite being a popular choice for classification, QDA does not perform very well when the class-conditional probability densities are very different from the normal distribution. In several such situations, the MMD classification rule, which is purely a nonparametric method, produces better results.

In spite of this, discriminant analysis, linear as well as quadratic, has generally proved to be highly effective in providing solutions to a variety of classification problems. As such, it has not only been very popular with researchers, it has also been a significant area of research. A lot of recent research in the area has been directed towards making these methods more effective in situations where they fail or are less effective. One such situation arises when the number of random variables is much larger than the number of observations available on them, as is very often the case with many modern-day real-life problems. In such situations, sample estimates of the class dispersion matrices may be unstable and even singular. Numerous methodologies based on discriminant analysis which are tailor-made for dealing with such situations have been published. These include vertex discriminant analysis (VDA) of Wu and Lange [24]; sparse discriminant analysis based on the optimal scoring interpretation of linear discriminant analysis, proposed by Clemmensen et al. [6]; discriminant analysis constructed via lasso penalized least squares, by Mai et al. [15]; the regularized optimal affine discriminant (ROAD) of Fan et al. [8]; shrinkage-based and regularization diagonal discriminant methods proposed by Pang et al. [17]; a thresholding-based approach reported by Shao et al. [20]; penalized methods proposed, for example, by Hastie et al. [12] as well as Witten and Tibshirani [23], and a kernel QDA method proposed by Wang et al. [22]. Other modifications include the approach of Suzuki and Itoh [21] which focuses on reducing the processing time involved, and the linear boundary discriminant analysis proposed by Na et al. [16], which aims to improve accuracy

[*] Corresponding author.
E-mail addresses: smarajit@isical.ac.in (S. Bose), pamita@isical.ac.in (A. Pal), rita@isical.ac.in (R. SahaRay), jitadeepa@gmail.com (J. Nayak).

http://dx.doi.org/10.1016/j.patcog.2015.02.016
0031-3203/© 2015 Elsevier Ltd. All rights reserved.

through increase in class separability by reflecting the different significances of non-boundary and boundary patterns. For the same class of problems, Zhu and Martinez [26] proposed a variation of discriminant analysis, called subclass discriminant analysis (SDA), by approximating the underlying distribution of each class with a mixture of Gaussians. Gkalelis et al. [11] established a theoretical link between mixture subclass discriminant analysis (MSDA) [10] and a restricted Gaussian model and hence proposed two generalizations, namely, fractional step MSDA (FSMSDA) and kernel MSDA (KMSDA).

Other approaches that have been proposed for tackling high-dimensional, small-sample classification problems, include null space LDA (NLDA) by Chen et al. [5], orthogonal LDA (OLDA) and uncorrelated LDA (ULDA) by Ye [25], subspace LDA [3], regularized LDA [9], and pseudo-inverse LDA [19]. Null space LDA computes the discriminant vectors in the null space of the within-class scatter matrix, while uncorrelated LDA and orthogonal LDA belong to a family of algorithms for generalized discriminant analysis proposed by Ye [25]. The features in ULDA are uncorrelated, while the discriminant vectors in OLDA are orthogonal to each other. Subspace LDA (or PCA+LDA) incorporates an intermediate dimensionality reduction stage such as PCA to reduce the dimensionality of the original data before classical LDA is applied. Regularized LDA uses a scaled multiple of the identity matrix to make the scatter matrix nonsingular. Pseudo-inverse LDA employs the pseudo-inverse to overcome the singularity problem. More details on these methods, as well as their relationship, can be found in [25].

Most of the above approaches deal with the limitations of discriminant analysis in high-dimensional, small-sample problems (also referred to as small-n, large-p problems, where n is the sample size and p is the dimension of the random vector involved). Unlike these, this paper proposes a novel generalization of the QDA in the conventional framework (large n, small p) to make it more appropriate for the situations when the class-conditional probability densities may not be normal but may have an elliptically symmetric nature. However, this generalization is totally nonparametric and QDA and MMD are its two special cases. Further, the method can be made adaptive to a given data which makes it very flexible. The effectiveness of the proposed method has been illustrated through extensive experimentation, and its performance has been found to compare very favourably with some well-known powerful nonparametric classifiers in a variety of examples.

The organization of this paper is as follows: In Section 2 the proposed classification scheme, which generalizes the QDA and MMD classifiers, is presented for two-class problems. Generalization of the method to the multi-class (more than two classes) case is discussed in Section 3. Simulation studies for two class as well as multi-class problems are presented in Section 4. The proposed classification scheme is illustrated with some experimental data sets for both two-class and multi-class cases in Section 5, which also contains a comparison of performance of the proposed method with that of MMD, QDA and some standard multivariate classifiers over a variety of experimental datasets. Finally, concluding remarks and some suggested directions for further research are given in Section 6.

2. Two-class classification

In classification problems, the objective is to classify a given observation $\boldsymbol{x}$ on a p-variate random vector $\boldsymbol{X}$ into one of m competing populations by a decision rule $d(\boldsymbol{x}) : \mathbb{R}^d \to \{1, 2, \ldots, m\}$. This rule is constructed with the help of a set of n observations on $\boldsymbol{X}$, called the training set, in which all the m populations are represented. The optimal classification rule, namely, the Bayes rule (see [1,7], for example) assumes that, for $j = 1, 2, \ldots, m$, the random vector $\boldsymbol{X}$ has a probability distribution function $f_j(\boldsymbol{x})$ in population j, and that the *a priori* probability for an observation to arise out of population j is π_j, where $\sum_{i=1}^{m} \pi_i = 1$. Based on these, it assigns an observation $\boldsymbol{x}$ to the class with the largest *a posteriori* probability $\pi(j|\boldsymbol{x}) = \pi_j f_j(\boldsymbol{x})/K$, where $K = \sum_{i=1}^{m} \pi_i f_i(\boldsymbol{x})$.

In particular, when $m=2$, that is, in a two-class classification problem, an observation $\boldsymbol{x} = (x_1, \ldots, x_p)'$ on a random measurement vector $\boldsymbol{X}$ is taken for a single individual (or object) and, on the basis of $\boldsymbol{x}$, the individual (or object) is classified into one of the two classes, say, C_1 and C_2, in which $\boldsymbol{X}$ is assumed to have probability density functions $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ respectively. The classification rule essentially partitions the measurement space $\mathcal{X}$ into R_1 and R_2 such that if $\boldsymbol{x} \in R_1$ the individual (or object) is classified into C_1, and is classified into C_2 otherwise. If the prior probabilities π_1 and $\pi_2 = 1 - \pi_1$ are assumed to be equal for the two classes then, for the optimal Bayes rule, we have

$$R_1 = \left\{ \boldsymbol{x} : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \right\} = \left\{ \boldsymbol{x} : \log \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 0 \right\},$$

$$R_2 = \left\{ \boldsymbol{x} : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1 \right\} = \left\{ \boldsymbol{x} : \log \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 0 \right\}.$$

If $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ are multivariate normal densities with mean vectors $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and dispersion matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ respectively, we have

$$\log \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = \frac{1}{2} \log \left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} \right) + \frac{1}{2} [(\boldsymbol{x} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2) - (\boldsymbol{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1)]$$

$$= \frac{1}{2} \log \left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} \right) + \frac{1}{2} \Delta_d^2 \quad \text{say,} \qquad (2.1)$$

where

$$\Delta_d^2 = (\boldsymbol{x} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2) - (\boldsymbol{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1),$$

which is nothing but the difference of the squared Mahalanobis distances of $\boldsymbol{x}$ from the two classes C_1 and C_2. This leads to the QDA rule, for which

$$R_1 = \left\{ \boldsymbol{x} : \frac{1}{2} \log \left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} \right) + \frac{1}{2} \Delta_d^2 > 0 \right\}$$

$$= \left\{ \boldsymbol{x} : \Delta_d^2 > \log \left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} \right) \right\}. \qquad (2.2)$$

Note that when $\boldsymbol{\Sigma}_1$ is assumed to be equal to $\boldsymbol{\Sigma}_2$, this reduces simply to $R_1 = \{ \boldsymbol{x} : \Delta_d^2 > 0 \}$, which is identical to the MMD. However, MMD fails when this assumption is not true and QDA will be optimal in that case. On the other hand, if the class-conditional probability densities are not normal, it is possible that MMD might be a better choice compared to QDA. For example, simulation with the Cauchy distribution (with appropriate location and scale parameters) will show that QDA fails miserably while MMD performs relatively better. These results are presented in Table 3, in which the columns corresponding to $c=0$ and $c=1$ respectively contain results obtained with MMD and QDA.

Now, if $f_i(\boldsymbol{x})$ is assumed to be p-variate t-distribution having q degrees of freedom (d.f.), for i=1,2, that is, if

$$f_i(\boldsymbol{x}) = A |\boldsymbol{\Sigma}_i|^{-1/2} \left[1 + \frac{1}{q} (\boldsymbol{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) \right]^{-(p+q)/2}, \quad i = 1, 2,$$

where

$$A = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{q}{2}\right)} q^{p/2} \pi^{p/2},$$