Pattern Recognition 43 (2010) 1421–1430

Contents lists available at ScienceDirect

Pattern Recognition

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

A novel Bayesian logistic discriminant model: An application to face recognition

R. Ksantini [a,*], B. Boufama [a], Djemel Ziou [b], Bernard Colin [c]

[a] School of Computer Science, University of Windsor, Windsor, ON, Canada N9B 3P4
[b] Département d'informatique, Université de Sherbrooke, 2500 Bl. Université, faculté des sciences, Sherbrooke, Québec, Canada J1K2R1
[c] Département de mathématiques, Université de Sherbrooke, 2500 Bl. Université, faculté des sciences, Sherbrooke, Québec, Canada J1K2R1

ARTICLE INFO

Article history:
Received 18 November 2008
Received in revised form
13 July 2009
Accepted 22 August 2009

Keywords:
Linear discriminant analysis
Logistic regression
Bayesian theory
Variational method
Mixture of Gaussians
Small sample size problem
Face recognition

ABSTRACT

The linear discriminant analysis (LDA) is a linear classifier which has proven to be powerful and competitive compared to the main state-of-the-art classifiers. However, the LDA algorithm assumes the sample vectors of each class are generated from underlying multivariate normal distributions of common covariance matrix with different means (i.e., homoscedastic data). This assumption has restricted the use of LDA considerably. Over the years, authors have defined several extensions to the basic formulation of LDA. One such method is the heteroscedastic LDA (HLDA) which is proposed to address the heteroscedasticity problem. Another method is the nonparametric DA (NDA) where the normality assumption is relaxed. In this paper, we propose a novel Bayesian logistic discriminant (BLD) model which can address both normality and heteroscedasticity problems. The normality assumption is relaxed by approximating the underlying distribution of each class with a mixture of Gaussians. Hence, the proposed BLD provides more flexibility and better classification performances than the LDA, HLDA and NDA. A subclass and multinomial versions of the BLD are proposed. The posterior distribution of the BLD model is elegantly approximated by a tractable Gaussian form using variational transformation and Jensen's inequality, allowing a straightforward computation of the weights. An extensive comparison of the BLD to the LDA, support vector machine (SVM), HLDA, NDA and subclass discriminant analysis (SDA), performed on artificial and real data sets, has shown the advantages and superiority of our proposed method. In particular, the experiments on face recognition have clearly shown a significant improvement of the proposed BLD over the LDA.

1. Introduction

In the context of supervised learning, given a training set of input vectors $\{X_i\}_{i=1}^N$, where $X_i \in \mathbb{R}^k (k \geq 1)$ and $i \in \{1, 2, \ldots, N\}$, along with corresponding tags $\{t_i\}_{i=1}^N$, where $t_i \in \mathbb{N}$ and $i \in \{1, 2, \ldots, N\}$, we wish to learn a model of dependency of the targets on the inputs. The final objective would be to be able to make accurate predictions of t for unseen values of X. In the case of real-world data, the presence of class overlap in classification implies that the principal modelling challenge is to avoid over-fitting of the training set. Typically, we base our predictions upon some function $y(X)$ defined over the input space (or training space) $\mathcal{X}$, and learning is the process of inferring the parameters or weights of this function. We concentrate here on functions of the type corresponding to those implemented by some relevant linear models, such as, the support vector machine (SVM) [23] and the linear discriminant analysis (LDA) [1,16]. The SVM and LDA make predictions based on the function

$$y(X; \mathbf{w}) = \sum_{i=1}^{k} w_i x_i + w_0, \tag{1}$$

where $\{x_i\}_{i=1}^k$ are the components of the vector X and $\{w_i\}_{i=0}^k$ are the unknown weights to compute. In the last decades, a number of powerful linear classifiers, such as SVM [23], LDA analysis [16] and logistic regression (LR) [18], have been proposed in the machine-learning community. However, except for the LDA, none of these classifiers incorporates the probability distributions fitting the transformed classes in order to avoid the noise in the data and optimize the linear separability in the input space. In these methods, it is not necessary to create representations or models for objects as the model of a given object is implicitly defined by the selection of its sample images. Unfortunately, these images are typically represented in spaces that are too large to allow robust and fast object recognition. In particular, the demand for a large number of training samples to construct a 'good' Bayesian classifier is difficult to satisfy due to the lack of training samples. To overcome this problem, the LDA has emerged as a fairly decent alternative to Bayesian classifier

* Corresponding author.
E-mail addresses: ksantini@uwindsor.ca (R. Ksantini), boufama@uwindsor.ca (B. Boufama), djemel.ziou@usherbrooke.ca (D. Ziou), bernard.colin@usherbrooke.ca (B. Colin).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.08.021

[30]. The practical attractiveness of LDA can be explained by its (intrinsically) low model complexity, and its ability to capture the essential characteristics of the data distributions (mean and covariance) from finite training data, and then estimating the decision boundary using these 'global' characteristics of the data. The LDA has proven to be powerful and competitive to several linear classifiers [13]. Its main goal is to find linear projections such that the classes are well separated, i.e., maximizing the distance between means of classes and minimizing their intraclass variances. The LDA was successfully applied in appearance-based methods for object recognition, such as, face recognition [2] and mobile robotics [25]. In fact, the success of LDA is partially due to the fact that only up to second order moments (mean and covariance) of the class distribution are used in LDA. This approach is more robust than estimating the distribution of the data. However, the LDA is incapable of dealing explicitly with heteroscedastic data, i.e., data in which classes do not have equal covariance matrices [16]. Moreover, most of the existing LDA-based methods inherit the parametric nature from the traditional LDA approach: The construction of the scatter matrices relies on the underlying assumption that the samples in each class satisfy the Gaussian distribution. Thus, they suffer from performance degradation in cases of non-normal distribution [5]. In addition, the LDA suffers from the small sample size problem for applications involving high-dimensional data [29]. Many methods have been proposed to address the small sample size problem [2,29,16,1,27,17]. Mika et al. [16] proposed adding a small multiple of the identity matrix to make the within-scatter matrix invertible. In [1,27,17], the authors used QR and generalized singular value decompositions to avoid the singularity of the within-scatter matrix. The proposed method in [2] has overcome the complication of a singular within-scatter matrix by reducing the dimension of the feature space using principal component analysis (PCA). Yu and Yang have proposed in [29] a direct LDA which allows a simultaneous diagonalization of the within-scatter matrix and between-scatter matrix. To overcome the heteroscedasticity problem, Loog and Duin [14] proposed the heteroscedastic LDA (HLDA) which is an heteroscedastic extension of the Fisher criterion and based on the Chernoff distance. To relax the normality assumption, Fukunaga [5] proposed the nonparametric DA (NDA) which is based on a new definition for the between-class scatter matrix, which explicitly emphasizes the samples near boundary. Unfortunately, none of the above methods is capable to overcome all these three problems simultaneously i.e., the small sample size, normality and heteroscedasticity.

In this paper, we propose a novel Bayesian logistic discriminant (BLD) model that avoids the small sample size problem by using a sparsity-promoting Gaussian prior over the unknown parameters or weights. This model is considered as a significant extension and improvement of the model proposed by [12]. In fact, a sparsity-promoting Gaussian is used to avoid the small size problem. Moreover, novel subclass and multinomial versions of the model are proposed to address the problems posed by nonlinearly separable classes and to perform polychotomous classification. Furthermore, each class is represented by its own Gaussian mixture distribution to solve both normality and heteroscedasticity problems targeted by the NDA and HLDA, respectively. In fact, in most real-world problems the form of each class pdf is a priori unknown, and the selection of the DA algorithm that best fits the data is done over trial-and-error. Ideally, one would like to have a single formulation which can be used for most distribution types. This can be achieved by approximating the underlying distribution of each class with a mixture of Gaussians. This can allow more robustness against the noise in the data, optimal linear transformation that maximizes the class separability in the input space and more flexibility and better classification performances than the LDA, HLDA and NDA. The objective or 'likelihood' function of our model has no tractable form. For this reason, variational transformation and Jensen's inequality are used to approximate it with a tractable exponential form which depends only on two variational parameters. Due to the conjugacy, by combining a sparsity-promoting Gaussian prior with the likelihood approximation, we have obtained a close Gaussian form approximation to the posterior distribution of the model. We have particularly targeted the face recognition problem as an application of interest to our proposed model given that it has become one of the most challenging tasks in the pattern recognition area [10]. Furthermore, face recognition is also central to many other applications such as video surveillance and identity retrieval from databases for criminal investigations.

This paper is organized as follows. Section 2 details the derivation of the BLD and defines the procedure for obtaining variational parameters, prior parameter values and the weights. A brief analysis of the complexity and numerical accuracy of the BLD is provided. Furthermore, a subclass and multinomial versions of the BLD are proposed to address the problems posed by nonlinearly separable classes and to perform polychotomous classification, respectively. Section 3 provides a comparative evaluation of the BLD to the LDA, SVM [23], HLDA [14], NDA [5] and SDA [31], carried out on a collection of benchmark synthetic and real data sets. Experiments on face recognition are also provided. The conclusion is presented in Section 4.

2. The Bayesian logistic discriminant model

2.1. Definition of the derivation of the BLD model

The idea of the LDA analysis is to solve the well-known problem of Fisher's linear discriminant in the input space. In the linear case, Fisher's discriminant aims at finding linear projections such that the classes are well separated, i.e., maximizing the distance between means of the classes and minimizing their intraclass variances. Implicitly, the LDA purpose is to find the most discriminative linear projections of the Gaussian distributions modelling the classes in the input space. This can be achieved by maximizing the Rayleigh coefficient (the ratio of the between-scatter matrix against the within-scatter matrix) with respect to the weights [16]. However, according to the form of the Rayleigh coefficient, the classes are assumed to be normally distributed with equal covariance structure, which is not true in many real-world applications. To overcome this problem, instead of using Rayleigh coefficient, a novel objective or 'likelihood' function is proposed that represents each class by its own Gaussian mixture distribution. Although the proposed objective function is theoretically different from the Rayleigh coefficient, it has the same purpose. Let $\mathcal{X}_1=\{X_i\}_{i=1}^{N_1}$ and $\mathcal{X}_2=\{X_i\}_{i=N_1+1}^{N}$ be two different classes constituting an input space of N samples or vectors in $\mathbb{R}^M$. Let us denote by $\underline{x}_1$ and $\underline{x}_2$ two random vectors whose realizations represent the classes $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. We suppose that $\underline{x}_1 \sim Mg_1(\underline{x}_1)$ and $\underline{x}_2 \sim Mg_2(\underline{x}_2)$, where Mg_1 and Mg_2 are two different Gaussian mixture distributions modelling $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. The unknown parameters of Mg_1 and Mg_2 are estimated by the EM algorithm [4] and their component numbers are selected using the minimum message length (MML) validity function, as it has been shown to give good results in [20]. Let $\underline{x}_1$ be associated the tag $t_0=0$ and $\underline{x}_2$ be associated the tag $t_0=1$. The unknown parameters (weights) are considered random variables and are denoted by the random vector $\mathbf{w}=(w_0,w_1,\ldots,w_N)$. The 'likelihood' function is defined as

$$P(t_0=0,t_0=1|\mathbf{w})=\sum_{\underline{x}_1\in\mathcal{X}_1,\underline{x}_2\in\mathcal{X}_2}\left[\prod_{i=1}^{2}P(t_0=i-1|x_i,\mathbf{w})Mg_i(x_i)\right], \quad (2)$$

where, given $F(x)=e^x/(1+e^x)$, the probabilities $P(t_0=i-1|\underline{x}_i,\mathbf{w})=F((2i-3)\mathbf{w}^T\underline{x}_i)$, $i\in\{1,2\}$ represent the logistic modellings of $t_0=0$