Neurocomputing 129 (2014) 33–40

Contents lists available at ScienceDirect

Neurocomputing

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

Graphical lasso quadratic discriminant function and its application to character recognition

Bo Xu [a], Kaizhu Huang [e,*], Irwin King [c], Cheng-Lin Liu [b], Jun Sun [d], Naoi Satoshi [d]

[a] Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, Beijing 100190, PR China
[b] National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, Beijing 100190, PR China
[c] Department of CSE, The Chinese University of Hong Kong, Hong Kong
[d] Fujitsu Research and Development Center, Beijing, China
[e] Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, 111 Ren Ai Road, Dushu Lake Higher Education Town, Suzhou, Jiangsu 215123, China

ARTICLE INFO

Article history:
Received 31 January 2012
Received in revised form
16 June 2012
Accepted 11 August 2012
Available online 7 November 2013

Keywords:
Graphical lasso
Quadratic discriminant function
Character recognition

ABSTRACT

Multivariate Gaussian distribution is a popular assumption in many pattern recognition tasks. The quadratic discriminant function (QDF) is an effective classification approach based on this assumption. An improved algorithm, called modified QDF (or MQDF in short) has achieved great success and is widely recognized as the state-of-the-art method in character recognition. However, because both of the two approaches estimate the mean and covariance by the maximum-likelihood estimation (MLE), they often lead to the loss of the classification accuracy when the number of the training samples is small. To attack this problem, in this paper, we engage the graphical lasso method to estimate the covariance and propose a new classification method called the graphical lasso quadratic discriminant function (GLQDF). By exploiting a coordinate descent procedure for the lasso, GLQDF can estimate the covariance matrix (and its inverse) more precisely. Experimental results demonstrate that the proposed method can perform better than the competitive methods on two artificial and nine real datasets (including both benchmark digit and Chinese character data).

1. Introduction

Statistical techniques are widely used for classification in various pattern recognition problems [14]. Statistical classifiers include linear discriminant function (LDF), quadratic discriminant function (QDF), Parzen window classifier, nearest-neighbor (1-NN), k-NN rules and margin classifiers [13,12]. QDF is derived under the assumption of multivariate Gaussian distribution for each class. Despite its simplicity, QDF and its variants have achieved great success in many fields. In a performance evaluation study of classifiers in handwritten character recognition, QDF and its variants were shown to be superior in the resistance to noncharacters even though they were not trained with noncharacter data. The parameters involved in QDF, e.g., the mean and the covariance, are often obtained via the principle of the maximum-likelihood estimation (MLE) [10]. MLE has a number of attractive features. First, it usually has good convergence properties as the number of training samples increases. Furthermore, it can often be simpler than alternative methods, such as Bayesian techniques. However, when the number of training samples is small (especially when compared to dimensionality), the estimated covariance based on MLE could be often ill-posed, making the covariance matrix singular; this further leads its inverse matrix to not be computed reliably.

To solve this problem, there have been a number of approaches in the literature. Modified quadratic discriminant function (MQDF) [15] is proposed to replace the minor eigenvalues of covariance matrix of each class with a constant parameter. This small change proves very effective and has made MQDF a state-of-the-art classifier in character recognition. However, the substitution of minor eigenvalues with a constant inevitably loses some class information. Meanwhile, the cutoff threshold of minor eigenvalues and the constant selection are critical for the final performance. Liu et al. [19] proposed a discriminative learning algorithm called discriminative learning QDF (DLQDF). It optimizes the parameters of MQDF with the aim to improve the classification accuracy based on the criterion of minimum classification error (MCE). Similar to MQDF, DLQDF has the same problem in parameter selection. Alternatively, the regularized discriminant analysis (RDA) [6] improves the performance of QDF by covariance matrix interpolation. Hoffbeck and Landgrebe further extended RDA by optimizing the interpolation coefficients [11]. Empirical results showed that these two algorithms can usually improve the classification performance of QDF. However, the improvements are also dependent on two critical parameters β and γ. In short, all of the above-mentioned

* Corresponding author.

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.08.073

methods need empirical settings of parameters to achieve the best results, which are however both time-consuming and task-dependent in real applications.

Different from the above approaches, in this paper, we present a novel method, called the graphical lasso quadratic discriminant function (GLQDF). By engaging the graphical lasso, the covariance estimation of the ordinal QDF can be successfully conducted even when the number of training samples is very small. Moreover, we can estimate the inverse of the covariance directly and hence avoid singular problems involved in QDF. One appealing feature is that the whole process is parameter-insensitive. This presents one big advantage over the other methods.

The rest of the paper is organized as follows. In the next section, we make an overview of QDF and MQDF. In Section 3, we introduce our novel GLQDF in detail. In Section 4, we conduct a series of experiments to verify our method. Finally, we set out concluding remarks in Section 5.

2. Review of QDF and MQDF

In this section, we review the QDF and the MQDF and also present some basic notations used throughout the paper.

2.1. Quadratic discriminant function

In this section we briefly review the algorithm of QDF. Let $x=(x_1,\ldots,x_d)^T$ represent a feature of a pattern, the posteriori probability can be computed by the Bayes rule:

$$P(\omega_i|x)=\frac{P(\omega_i)p(x|\omega_i)}{p(x)},\quad i=1,\ldots,M \tag{1}$$

where $P(\omega_i)$ is the prior probability of class ω_i, $p(x|\omega_i)$ is the class probability density function (pdf) and $p(x)$ is the mixture density function. Since $p(x)$ is independent of class label, the nominator of Eq. (1) can be used as the discriminant function for classification:

$$g(x|\omega_i)=p(\omega_i)p(x|\omega_i). \tag{2}$$

Assume the pdf of each class is multivariate Gaussian:

$$p(x)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-u)^t\Sigma^{-1}(x-u)\right\}, \tag{3}$$

where x is a d-component vector, μ is the mean vector, and Σ is the $d\times d$ covariance matrix. The quadratic discriminant function is derived from Eq. (3) as follows:

$$g(x|\omega_i)=(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i)+\log|\Sigma_i|. \tag{4}$$

The QDF is actually a distance metric in the sense that the class of minimum distance is assigned to the input pattern.

By K–L transform, the covariance matrix can be diagonalized as

$$\Sigma=\Phi\Lambda\Phi^T \tag{5}$$

where $\Lambda=diag[\lambda_1,\ldots,\lambda_d]$ with $\lambda_i,\ i=1,\ldots,d$, being the eigenvalues (in decreasing order) of Λ, and $\Phi=[\phi_1,\ldots,\phi_d]$ with $\phi_i,\ i=1,\ldots,d$, being the ordered eigenvectors.

Thus the QDF can be rewritten in the form of eigenvectors and eigenvalues:

$$\begin{aligned}g(x|\omega_i)&=[\Phi_i^T(x-\mu_i)]^T\Lambda_i^{-1}\Phi_i^T(x-\mu_i)+\log|\Lambda_i|\\&=\sum_{j=1}^{d}\frac{((x-\mu_i)^t\varphi_{ij})^2}{\lambda_{ij}}+\sum_{j=1}^{d}\log\lambda_{ij}.\end{aligned} \tag{6}$$

This function will lead to the optimal classifier, provided that (1) the actual distribution is normal, (2) the prior probabilities of all categories are equal and (3) the parameters μ and Σ can be reliably provided. However, since the parameters are usually unknown, the sample mean vector $\hat{\mu}$ and sample covariance matrix $\hat{\Sigma}$ are used

$$\begin{aligned}\hat{g}(x|\omega_i)&=[\hat{\Phi}_i^T(x-\hat{\mu}_i)]^T\hat{\Lambda}_i^{-1}\hat{\Phi}_i^T(x-\hat{\mu}_i)+\log|\hat{\Lambda}_i|\\&=\sum_{j=1}^{d}\frac{((x-\hat{\mu}_i)^t\hat{\varphi}_{ij})^2}{\hat{\lambda}_{ij}}+\sum_{j=1}^{d}\log\hat{\lambda}_{ij},\end{aligned} \tag{7}$$

here λ_{ij} is the i-th eigenvalue of $\hat{\Sigma}_i$ and $\hat{\varphi}_i$ is the eigenvector. It is well-known that small eigenvalues in Eq. (7) are usually inaccurate; this hence causes the reduction of recognition accuracy. Moreover, the computational cost of Eq. (7) is $O(d^3)$ for d-dimensional vectors, which may be computationally difficult when the dimension is high.

2.2. Modified quadratic discriminant function

MQDF is a modified version of the ordinary QDF. QDF suffers from the quadratic number of parameters, which cannot be estimated reliably when the number of samples per class is smaller than the feature dimensionality. MQDF reduces the complexity of QDF by replacing the small eigenvalues of covariance matrix of each class with a constant. Consequently, the small eigenvectors will disappear in the discriminant function. This reduces both the space and the computational complexity. More importantly, this small change proves to improve the classification performance significantly. Denote the input sample by a d-dimensional feature vector $x=(x_1,x_2,x_3,\ldots,x_d)^T$. For classification, each class c_i is assumed to have a Gaussian density $p(x|c_i)=N(u_i,\sigma_i)$, where μ_i and σ_i are the class mean and covariance matrix, respectively. Assuming equal a priori class probabilities, the discriminant function is given by the log-likelihood

$$-2\log p(x|\omega_i)=(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)+\log|\Sigma_i|+CI \tag{8}$$

where CI is a class-independent term, and is usually omitted. We take the minus log-likelihood to make the discriminant function a distance measure. The covariance matrix Σ_i can be diagonalized as Λ_i, where $\Lambda_i=diag[\lambda_{i1},\ldots,\lambda_{ik},\ldots,\lambda_{id}]$ has the eigenvalues of λ_{ik} (in descending order) as diagonal elements, φ_{ik} is an ortho-normal matrix comprising as columns the eigenvectors of λ_{ik}. Replacing the minor eigenvalues with a constant, i.e., replacing Λ_i with $diag\,[\lambda_{i1},\ldots,\lambda_{ik},\delta_i,\ldots,\delta_i]$ (k is the number of principal eigenvectors to be retained), the discriminant function of Eq. (7) becomes what we call MQDF:

$$\begin{aligned}g(x|\omega_i)=&\sum_{j=1}^{k}\frac{((x-\mu_i)^t\varphi_{ij})^2}{\lambda_{ij}}+\sum_{j=1}^{k}\log\lambda_{ij}\\&+\frac{1}{\delta_i}\left(\|x-\mu_i\|^2-\sum_{j=1}^{k}|(x-\mu_i)^T\varphi_{ij}|^2\right)+(d-k)\log\delta_i\end{aligned} \tag{9}$$

where $i,j=1,\ldots,k$ are the principal eigenvectors of the covariance matrix of class ω_i.

By defining

$$r_i(x)=\|x-\mu_i\|^2-\sum_{j=1}^{k}|(x-\mu_i)^T\phi_{ij}|^2 \tag{10}$$

where $r_i(x)$ is the residual of subspace projection, Eq. (9) can be rewritten as

$$g(x|\omega_i)=\sum_{j=1}^{k}\frac{((x-\mu_i)^t\varphi_{ij})^2}{\lambda_{ij}}+\sum_{j=1}^{k}\log\lambda_{ij}+\frac{1}{\delta_i}r_i(x)+(d-k)\log\delta_i \tag{11}$$

The parameters of MQDF are estimated as follows. The mean vector and covariance matrix of a class are estimated from the sample data of this class. The class-dependent δ_i is calculated by