Neurocomputing 316 (2018) 345–356

Contents lists available at ScienceDirect

Neurocomputing

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

Linear discriminants described by disjoint tangent configurations

José-Luis Sancho-Gómez[a], Juan-Antonio Martínez-García[a,*], Stanley C. Ahalt[b], Aníbal R. Figueiras-Vidal[c]

[a] Departamento de Tecnologías de la Información y las Comunicaciones, Universidad Politécnica de Cartagena, Murcia, Spain
[b] Department of Computer Science, University of North Carolina at Chapel Hill, North Carolina, USA
[c] Departamento de Teoría de la Señal y Comunicaciones, Universidad Carlos III de Madrid, Leganés-Madrid, Spain

ARTICLE INFO

Article history:
Received 11 December 2017
Revised 20 July 2018
Accepted 2 August 2018
Available online 10 August 2018

Communicated by Dr Li Sheng

MSC:
62C10
62C20
62H30
68T05

Keywords:
Bayes theory
Binary classification
Parametric linear discriminant
Accuracy criterion
Minimax probability machine

ABSTRACT

In this paper, a new interpretation of parametric linear discriminants for binary classification problems is presented. Linear discriminants are described in terms of Disjoint Tangent Configurations (DTC) established between the ellipsoidal level surfaces resulting from the means and covariance matrices of the distributions. This is a new framework that allows, first, a new interpretation and analysis of several well-known linear discriminants and, second, the design of new discriminants with very interesting properties. In particular, it is shown that the analytical expression of the Bayes Linear Discriminant —whose explicit expression is still unknown— can be derived from a particular DTC. Besides the Bayes discriminant, other classical linear discriminants are also described according to the DTC analysis, in particular, the Fisher and the Scatter-based Linear Discriminants. On the other hand, two new linear discriminants for the minimax and the Bayesian solutions are obtained from the DTC analysis. Both have a direct analytical expression in contrast to the existing iterative solutions, with which they are compared. The first DTC discriminant, which is called MPDH-DTC, is the solution of the Minimax Probabilistic Decision Hyperplane (MPDH) problem, the same solution that the Minimax Probability Machine (MPM) method approximates by an iterative convex optimization. The second discriminant, called Quasi-Bayes-DTC Linear Discriminant, is designed to be an approximation to the Bayes Linear Discriminant, which requires a search procedure to find the solution.

Considering both the accuracy over several synthetic and real problems and the computational cost, the Quasi-Bayes-DTC is the preferred discriminant due to its high performance and low computational cost, unless a minimax solution is required, in that case the MPDH-DTC is preferred.

1. Introduction

A Bayes classifier minimizes the probability of error [1]. Nevertheless, the Bayes criterion requires the knowledge of the probability density functions which must be estimated from the data. Although density estimation techniques are available, the estimations are computationally complex, and large amounts of data are needed to provide accurate results. Thus, simpler procedures have been developed to solve classification problems, including parametric techniques that specify the mathematical form of the classifier, followed by parametric estimation.

A very common and efficient technique is to choose a linear model both to solve classification problems and as a feature extraction tool [2–7]. Although it is well known that linear classifiers are suboptimal [1], many attempts have been made to design the best linear discriminant for both normal and non-normal distributions. In many cases, a linear discriminant is preferred for simplicity and robustness, and ease of interpretation.

Given a particular classification problem, it is generally convenient to perform a transformation of the data space to overcome the performance obtained in the original space. Many different techniques have been proposed in the literature to make efficient transformations such as the very well-known Principal Components Analysis (PCA) [8]) –paradigm of dimensionality reduction methods–, the Denoising Autoencoders (DAE) [9]) –nowadays one of the most used technique due to its high representation capability–, and other very recent proposals as the low-rank subspace Learning [10] and dictionary learning [11,12], in which data are represented as a sparse linear combination of independent vectors of an over-complete space. However, this is a research field beyond the scope of this article.

* Corresponding author.
E-mail addresses: josel.sancho@upct.es (J.-L. Sancho-Gómez), juan.antonio.mtnez@gmail.com (J.-A. Martínez-García), ahalt@renci.org (S.C. Ahalt), anibalrfv@tsc.uc3m.es (A.R. Figueiras-Vidal).

https://doi.org/10.1016/j.neucom.2018.08.010
0925-2312/

In [1], a procedure to design parametric linear discriminants for binary problems is introduced. This method is optimal with respect to a separability criterion defined in a one-dimensional projected space, i.e., the linear discriminant defines a direction along which the projected data of one class are maximally separated from the projected data of the another class. Different linear discriminants are obtained when different separability criteria are selected. The most important criteria are the Bayes error for normal distributions, the Fisher criterion, and other criteria based on scatter matrices [13–15].

In this paper, a new interpretation of these linear discriminants is presented. Linear discriminants are described in terms of Disjoint Tangent Configurations (DTC) established between the level surfaces of distributions characterized by a mean vector and a covariance matrix. For such distributions, the level surfaces are ellipsoids. The analytical relation between these DTC discriminants and those obtained by the classical parametric design is established.

The DTC discriminant analysis is also applied to minimax classification problems. The minimax solution is an important issue in pattern recognition, for example, when the number of training data of each class does not reflect the actual prior probabilities. Therefore, minimax is a natural classification criterion in the absence of prior information regarding the true frequency of the two classes. For this reason, many researchers prefer to use classifiers operating at Equal Error Rate (EER), that is, classifiers that minimize the maximum of the false alarm and miss rates [16,17]. Moreover, the minimax problem can also be addressed when the information of the class distributions is unknown or not exact. In particular, the authors of [18] and [19] introduce the Minimax Probabilistic Decision Hyperplane (MPDH) as that which separates two classes of points with maximal probability with respect to all distributions having the same means and covariance matrices. The MPDH problem is addressed there through iterative algorithms that impose bounds to the classification errors and use convex optimization methods, the most common being the Minimax Probability Machine (MPM). Here, a non-iterative and simple expression of the MPDH solution is obtained as a particular DTC, and its performance over several artificial and real problems is analyzed.

Finally, a new linear discriminant is also obtained from a particular DTC. This linear discriminant presents two remarkable properties: It produces an accuracy performance close to that of the Bayes Linear Discriminant (for this reason, it is called Quasi–Bayes-DTC), and it is obtained from a simple expression with a competitive computational cost because it is neither an iterative procedure nor a search method.

The paper is organized as follows. Section 2 shows a brief review of the design of parametric linear discriminants including the MPDH. In Section 3, the linear discriminants presented in the previous section are described in terms of the DTC analysis. A separate section is reserved to the totally new Quasi-Bayes-DTC linear discriminant, it is presented in Section 4. Section 5 presents a comparative study of the performance of all the linear discriminants considered in this work. The accuracy results using both synthetic and real data are analyzed. Conclusions and ideas for future work complete the paper.

The main contributions of this article are:

1. The presentation of a new framework in which several well-known linear discriminants —such as Bayes, Fisher, those based on the dispersion of the classes (Scatter-based) or the Minimax Probability Machine (MPM)— acquire a new common interpretation: they are discriminants described by disjoint tangent configurations (DTC). In this sense, these classic discriminants present a new way of being understood and analyzed.
2. This framework allows obtaining new linear discriminants with very interesting properties. In particular, two new linear discriminants are presented:
   (a) the MPDH-DTC discriminant which is a direct solution of the minimax MPDH problem, unlike the MPM method which is an iterative approach.
   (b) the Quasi-Bayes-DTC discriminant that provides a very close accuracy to that of the Bayes Linear Discriminant, but with the advantage of needing a lower computational cost.

2. A brief review of designs of parametric linear discriminants

One of the most straightforward procedures to design classifiers is to assume a specific analytical form of the discriminant which contains a number of adjustable parameters. The values of these parameters can then be optimized to obtain the best classification.

The simplest choice is the linear form, whose discriminant function can be written as

$$h(\mathbf{x}) = \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^n$ is called the weight vector, n is the dimension of $\mathbf{x}$, and $\mathbf{x}_0 \in \mathbb{R}^n$ the bias vector. The decision rule implemented by a linear discriminant function is normally expressed as

$$h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + \omega_0 \underset{C_2}{\overset{C_1}{\gtrless}} 0 \tag{2}$$

where

$$\omega_0 = -\mathbf{w}^T\mathbf{x}_0 \tag{3}$$

This equation indicates that the n-dimensional vector $\mathbf{x}$ is projected onto the vector $\mathbf{w}$, and it is classified as either class C_1 or class C_2, depending on whether variable $z = \mathbf{w}^T\mathbf{x}$ is greater or less than $-\omega_0$. Hence, ω_0 is called the threshold weight. Thus, the task of designing a linear classifier consists of finding a weight vector $\mathbf{w}$ and a threshold value ω_0 (or a bias vector $\mathbf{x}_0$) that provide the smallest error in the one-dimensional projected space, or h-space. Equation $h(\mathbf{x}) = 0$ describes the decision boundary, which is a hyper-plane. This hyper-plane is determined by the weight vector $\mathbf{w}$, that defines the orientation, and by the bias vector $\mathbf{x}_0$, which is a point of the hyper-plane that fixes its relative position in the data space. Thus, a linear discriminant divides this space into two half-spaces by means of a hyper-plane.

In order to design a linear classifier, it is established a separability criterion whose optimization provides the values of $\mathbf{w}$ and ω_0. A function $f(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ of the means and variances of $h(\mathbf{x})$ can be an appropriate criterion to measure the class separability because, even when samples $\mathbf{x}$ are not normally distributed, $h(\mathbf{x})$ could be close to normal for large n [1]. These means and variances are given by

$$\begin{aligned}\mu_j &= E\{h(\mathbf{x})|C_j\} = \mathbf{w}^T E\{\mathbf{x} \mid C_j\} + \omega_0 \\ &= \mathbf{w}^T\mathbf{m}_j + \omega_0\end{aligned} \tag{4}$$

$$\begin{aligned}\sigma_j^2 &= Var\{h(\mathbf{x})|C_j\} \\ &= \mathbf{w}^T E\{(\mathbf{x} - \mathbf{m}_j)(\mathbf{x} - \mathbf{m}_j)^T \mid C_j\}\mathbf{w} \\ &= \mathbf{w}^T\Sigma_j\mathbf{w}\end{aligned} \tag{5}$$

where $\mathbf{m}_j \in \mathbb{R}^n$ and $\Sigma_j \in \mathbb{R}^{n\times n}$ are the mean vector and covariance matrix of $\mathbf{x} \in C_j$, $j = 1, 2$, respectively. It can be shown [1] that the optimization of function f produces a linear discriminant given by

$$\mathbf{w} = [s\Sigma_1 + (1 - s)\Sigma_2]^{-1}(\mathbf{m}_2 - \mathbf{m}_1) \tag{6}$$

where

$$s = \frac{\partial f/\partial\sigma_1^2}{\partial f/\partial\sigma_1^2 + \partial f/\partial\sigma_2^2} \tag{7}$$