Neural Networks 81 (2016) 59–71

Contents lists available at ScienceDirect

Neural Networks

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

Pseudo-inverse linear discriminants for the improvement of overall classification accuracies

Gao Daqi *, Dastagir Ahmed, Guo Lili, Wang Zejian, Wang Zhe

Department of Computer Science, State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai 200237, China

HIGHLIGHTS

- Relationship between a PILD and an FLD is clarified based on overall accuracy.
- A PILD is not certainly equivalent to an FLD if the desired outputs are fixed.
- A PILD has nothing in common with an FLD when the desired outputs are changeable.
- Accuracies of PILDs are improved by optimal thresholds related to sizes and regions.
- The iterative learning strategy of PILDs is proposed, realized and verified.

ARTICLE INFO

Article history:
Received 22 June 2015
Received in revised form 17 May 2016
Accepted 28 May 2016
Available online 7 June 2016

Keywords:
Pseudo-inverse linear discriminants (PILDs)
Fisher linear discriminants (FLDs)
Threshold optimization
Iterative learning
Overall classification accuracies

ABSTRACT

This paper studies the learning and generalization performances of pseudo-inverse linear discriminant (PILDs) based on the processing minimum sum-of-squared error (MS^2E) and the targeting overall classification accuracy (OCA) criterion functions. There is little practicable significance to prove the equivalency between a PILD with the desired outputs in reverse proportion to the number of class samples and an FLD with the totally projected mean thresholds. When the desired outputs of each class are assigned a fixed value, a PILD is partly equal to an FLD. With the customarily desired outputs {1, -1}, a practicable threshold is acquired, which is only related to sample sizes. If the desired outputs of each sample are changeable, a PILD has nothing in common with an FLD. The optimal threshold may thus be singled out from multiple empirical ones related to sizes and distributed regions. Depending upon the processing MS^2E criteria and the actually algebraic distances, an iterative learning strategy of PILD is proposed, the outstanding advantages of which are with limited epoch, without learning rate and divergent risk. Enormous experimental results for the benchmark datasets have verified that the iterative PILDs with optimal thresholds have good learning and generalization performances, and even reach the top OCAs for some datasets among the existing classifiers.

1. Introduction

Performances of classifiers can be evaluated from two aspects: processing and targeting criteria. The most often-used processing evaluation criterion functions are the minimum sum-of-squared errors (MS^2Es) and the least-mean-squared (LMS) errors (Duda, Hart, & Stork, 2000). And the most conventional targeting evaluation criterion functions are the overall classification accuracies (OCAs), or called the overall recognition rates (ORRs) or the overall error rates (OERs) (Huang & Ling, 2005). There are two main learning procedures: analytical and iterative, to determine the parameters of classifiers (Duda et al., 2000; Suykens, Gestel, Brabanter, Moor, & Vandewalle, 2002). Specified to a linear classifier $\pi_l : \theta + \boldsymbol{w}^T\boldsymbol{x} = 0$, two main types of algorithms to determine the thresholds θ and the weight vectors $\boldsymbol{w}$ are as follows: (A) analytical procedures, e.g., the MS^2E solutions, depending upon the processing MS^2E criterion functions $J(\theta, \boldsymbol{w}) = \|\theta\mathbf{1} + \boldsymbol{X}\boldsymbol{w} - \boldsymbol{d}\|^2$ by one-time calculation of explicit equations in a lump, and (B) iterative procedures, e.g., the gradient descent solutions, mainly depending upon the processing LMS error functions $J(\theta(\tau), \boldsymbol{w}(\tau)) = \sum_p \left(\theta(\tau) + \boldsymbol{w}^T(\tau)\boldsymbol{x}_p - d_p\right)^2$ by numerous repeated iterations (Duda et al., 2000; Koford & Groner, 1966). The analytical procedures have two advantages over the iterative ones: (i) fast computational speeds and (ii) without local minimums.

* Correspondence to: Department of Computer Science, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China. Fax: +86 21 6425 2984.
E-mail address: gaodaqi@ecust.edu.cn (G. Daqi).

http://dx.doi.org/10.1016/j.neunet.2016.05.006
0893-6080/

Fisher linear discriminants (FLDs), also called linear discriminant analyses (LDAs), are very popular (Billings & Lee, 2002; Cawley & Talbot, 2003; Koford & Groner, 1966). Following the force of habit, FLDs are often taken for and in particular equated with linear classifiers (Cooke, 2002; Raudys & Duin, 1998; Rozza, Lombardi, Casiraghi, & Campadelli, 2012). However, it must be clarified that FLDs are only a type of analytical learning algorithms based on the Rayleigh quotients $J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}}$ in the Bayesian decision theory (Duda et al., 2000). Here, $\boldsymbol{S}_B$ and $\boldsymbol{S}_W$ are the between- and within-class scatter matrices, respectively. Of course, FLDs are more used as a kind of feature extraction tools than a type of classifiers, especially in the image processing fields, e.g., the classical Fisherfaces (Belhumeur, Hespanha, & Kriegman, 2007). Single-layer perceptrons are another type of linear classifiers (Elozondo, 2006; Escalera, Tax, Pujol, Radeva, & Duin, 2008; Suykens et al., 2002). They mainly employ the iterative learning procedures, e.g., back-propagation, to determine the thresholds θ and the weights $\boldsymbol{w}$.

Pseudo-inverse linear discriminants (PILDs), another kind of analytical learning algorithm for linear classifiers, obtain the θ and $\boldsymbol{w}$ according to the processing MS^{2}E criterion functions (Duda et al., 2000). Note that the word "pseudo-inverse" here is borrowed to show that the equation for calculating the θ and $\boldsymbol{w}$ formally contains the *Moore–Penrose* inverse item (Hoyle, 2011; Raudys & Duin, 1998; Tapson & van Schaik, 2013), but not to imply the non-invertibility of a rectangular matrix. For the purpose of simplicity, sometimes the abbreviated terms "PILDs" are inexactly used to imply "linear classifiers with pseudo-inverse discriminant algorithm", and so are "FLDs" to "linear classifiers with Fisher discriminant algorithm", henceforth.

A PILD has a bit higher computational complexity than an FLD, because the former processes an $(m+1) \times (m+1)$ square matrix while the latter does an $m \times m$ one. However, a PILD has two main advantages over an FLD: (A) the θ and the $\boldsymbol{w}$ are simultaneously obtained by solving a single analytical equation, and (B) the MS^{2}E criterion function could be further developed to optimize the θ and $\boldsymbol{w}$. It has been proved that a PILD with the specially desired outputs in reverse proportion to the number of samples is equivalent to an FLD with the totally projected mean (TPM) threshold (Duda et al., 2000). FLDs are quite popular (Billings & Lee, 2002; Cooke, 2002; Raudys & Duin, 1998); however, PILDs are rarely applied to practice. The true reason lies in the fact that PILDs are not ideal in OCAs.

Real-world datasets are of diverse sizes and distributed regions (Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2012; Nicolas, Javier, & Aida de, 2012). Class imbalance is an objective being (Cano, Zafra, & Ventura, 2013; Fernandez, Garcia, Jesus, & Herrera, 2008; Huang & Ling, 2005). Taking a binary-class problem $\{\omega_1, \omega_2\}$ for example, there are the following four possible cases: (A) ω_1 is equal to ω_2 both in size and in region; (B) ω_1 is larger in size but smaller than or equal to ω_2 in region; (C) ω_1 is larger in region but smaller than or equal to ω_2 in size; (D) ω_1 is larger both in size and in region than ω_2. It is difficult for a single θ and even a single $\boldsymbol{w}$ obtained by using an analytical equation in one-lump computation to suit all cases.

In order to solve the classification problems of imbalanced datasets, we can proceed from the following approaches: (i) data and (ii) classifiers as well as algorithms. In the data levels are diversely over-, under-, and synthetic-sampling techniques (Barua, Islam, Yao, & Murase, 2014; Galar, Fernández, Barrenechea, & Herrera, 2013; Nicolas et al., 2012). Decision trees (Wang & Yao, 2013), support vector machines (SVMs) (Maldonado & Lopez, 2014; Tang, Zhang, Chawla, & Krasser, 2009), neural networks (Castro & Braga, 2013), FLDs (Gao, Ding, & Zhu, 2014; Rozza et al., 2012) and even k-nearest- neighbor rules (Barua et al., 2014) are the main classification models. And different cost-sensitive learning algorithms, e.g., bagging and boosting (Hand & Till, 2001; Nicolas et al., 2012; Shen & Li, 2010) as well as some criterion functions, especially, the areas under the ROC curves (AUCs) (Adams & Hand, 1999; Bradley, 1997; Huang & Ling, 2005), are paid much attention. Comparatively speaking, PILDs do not attract enough concern.

This paper aims at improving the learning and generalization performances of PILDs based on the processing MS^{2}E and the targeting OCA criterion functions (Adams & Hand, 1999; Bradley, 1997), and will devote to addressing the following problems:

(a) Is a PILD with the designated outputs in reverse proportion to the number of samples indeed equal to an FLD with the TPM threshold? What if the desired outputs are taken the conventional values $d_p \in \{1, -1\}$ (Duda et al., 2000; Koford & Groner, 1966; Rozza et al., 2012)?
(b) Are the weight vectors and the thresholds calculated by the pseudo-inverse solution optimal? If not, how to optimize them in a comprehensive consideration of sample sizes and distributed regions (Gao et al., 2014)?
(c) Is it reasonable to allocate the desired outputs of each class a fixed value? If not, how to re-allocate the rational ones for all training samples, including both the correct and the misclassified ones (Adams & Hand, 1999; Duda et al., 2000)?
(d) How to introduce an iterative learning strategy to the analytical PILDs in order to further optimize the thresholds and weights (Duda et al., 2000; Gao et al., 2014)?

Motivated by the troublesome issues above, this paper concentrates attention on improving the learning and generalization of PILDs from the aspects of thresholds, weight vectors and data in the algorithm level. The contributions of this work are summarized in the following:

(A) A PILD is not exactly equivalent to an FLD even if the desired outputs in the PILD are in reverse proportion to the number of samples. Several examples are given to support the argument.
(B) A PILD is partly equivalent to an FLD when all training samples of each class are allocated a fixedly desired output. And they two have nothing in common when all the training samples are assigned with changeably desired outputs.
(C) The most often-used TPM thresholds usually behave poor in OCAs because of the unsuitable truncation (Gao et al., 2014). A simple practicable MS^{2}E threshold comes into being by using the customarily desired outputs $d_p \in \{1, -1\}$. Furthermore, a rational threshold may be obtained by using the actually algebraic distances as the desired outputs.
(D) Multiple empirical thresholds are developed in a comprehensive consideration of sizes and regions. The optimal thresholds are singled out from among them aiming at the best OCAs.
(E) The iterative learning strategy of PILDs is proposed by means of the processing MS^{2}E criterion and the actually algebraic distances, which is with limited epochs, without learning rate and divergent risk.

We stress that this work is the development of our earlier work (Gao et al., 2014); therefore we will always pay much attention on the difference between PILDs and FLDs. The rest of this paper is organized as follows: Section 2 introduces the related work of PILDs. In Section 3, the relationship between PILDs and FLDs is clarified, and three examples are given to verify their similar and different places. Section 4 develops a series of empirical thresholds related to sizes and regions. Section 5 goes into details on the iterative learning strategy of PILDs based on the processing MS^{2}E criterion functions and the changeably desired outputs. Section 6 presents numerous experimental results for the real-world benchmark datasets to demonstrate the superior learning and generalization performances of PILDs. Finally, we will conclude this work in Section 7.