Pattern Recognition 47 (2014) 789–805

Contents lists available at ScienceDirect

Pattern Recognition

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

Integrated Fisher linear discriminants: An empirical study

Gao Daqi*, Ding Jun, Zhu Changming

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

ARTICLE INFO

Article history:
Received 6 December 2012
Received in revised form
16 July 2013
Accepted 26 July 2013
Available online 9 August 2013

Keywords:
Fisher linear discriminants
Imbalanced datasets
Empirical thresholds
Neighborhood-preserving transformations
Iterative learning

ABSTRACT

This paper studies Fisher linear discriminants (FLDs) based on classification accuracies for imbalanced datasets. An optimal threshold is found out from a series of empirical formulas developed, which is related not only to sample sizes but also to distribution regions. A mixed binary–decimal coding system is suggested to make the very dense datasets sparse and enlarge the class margins on condition that the neighborhood relationships of samples are nearly preserved. The within-class scatter matrices being or approximately singular should be moderately reduced in dimensionality but not added with tiny perturbations. The weight vectors can be further updated by a kind of epoch-limited (three at most) iterative learning strategy provided that the current training error rates come down accordingly. Putting the above ideas together, this paper proposes a type of integrated FLDs. The extensive experimental results over real-world datasets have demonstrated that the integrated FLDs have obvious advantages over the conventional FLDs in the aspects of learning and generalization performances for the imbalanced datasets.

1. Introduction

Linear classifiers are the basic pattern recognition models, mainly including classical Fisher linear discriminants (FLDs) [1–5], single-layer perceptrons (SLPs) [6–8] and linear support vector machines (SVMs) [9–11]. Because of simple structures and low computational costs, FLDs are considered as the most popular linear classifiers in many applications [12–14]. The key of success of FLDs seems to lie in the fact that the linear decision hyperplanes obtained usually offer reasonable partitions. However, a major issue existing in FLDs is that the parameters, namely the mean vectors and within-class scatter matrices, have to be estimated by limited available training samples in order to determine the weight vectors and thresholds [5,15,16]. Though the parameter estimations associated with Gaussian distributions reveal some good statistical properties, the Gaussian assumptions are not able to suit all distribution cases [17,18].

The criterion function $J(.)$ that maximizes the ratio of between-class scatters to within-class scatters in order to find the proper weight vectors is well known and a great creation, and thus called the Fisher's criterion [15]. Up to date, many criterion functions have been presented [11,16,17,19]. In a sense, the diverse quadratic optimization functions used in SVMs are a natural generalization of the Fisher's criterion [20,21]. As a matter of fact, FLDs are consistent with SVMs at the aim to maximize the class margins by constructing the optimal separating hyperplanes or projected directions [22].

Imbalance is absolute while balance is relative. Currently, more attention has been paid to imbalanced datasets [23–28], the most of which is centered on the imbalance of sample sizes for the sake of intuitiveness and simplicity. A classifier tends to categorize the present samples to the majority class when learning an imbalanced dataset [29]. In other words, it is prone to generate a classifier that has a strong bias toward the majority class, resulting in a large number of false negatives.

Re-sampling techniques are popular in solving the imbalanced problems, i.e., either the minority classes are over-sampled or the majority classes are under-sampled or some combinations of the two are employed [25,28,30–32]. In the meantime, boosting and bagging algorithms have been developed for successively training component classifiers, named the cost-sensitive classifiers [10,33]. They work by assigning larger weights to the mislabeled samples, otherwise smaller. Adaboost algorithm, a variation of boosting, is the popular one [34]. These algorithms can handle the imbalanced cases, to some extent.

Is an FLD able to solve a linearly separable problem? The answer is “Yes” for a “balanced” dataset. However, while a two-class dataset is seriously imbalanced, such an FLD may fail. The more serious the imbalance is, the poorer the resulting FLD performs. The imbalance of distribution regions usually has a far more important influence on the performances of classifiers than the imbalance of sample sizes does [35]. Therefore, it is relatively consistent with the actual situations to consider the imbalance of distribution regions, e.g., variances [5,10,16,18]. Indeed, scatters,

* Corresponding author. Tel.: +86 21 6425 3780; fax: +86 21 6425 2984.
E-mail address: gaodaqi@ecust.edu.cn (G. Daqi)

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.07.021

manifestations of variances, are already included in $J(.)$. However, the sum of two within-class scatter matrices is only used as the denominator of $J(.)$. In other words, $J(.)$ is not related to the difference of within-class scatters or the imbalance of distribution regions.

It is the threshold in an FLD that finally determines the location of a separating hyperplane. Based on the above observations, we firstly have to answer the question: "Can the classification accuracy of an FLD be improved by selecting a proper threshold?"

The weight vector $\boldsymbol{w}$ in an FLD for solving a two-class problem $\{\omega_1, \omega_2\}$ is only determined by the within-class scatter matrix $\boldsymbol{S}_W$ as well as two mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. If $\boldsymbol{S}_W$ is or approximately singular, the FLD will no longer work. In order to address this issue, $\boldsymbol{S}_W$ is often added with a tiny perturbation matrix [1,12,36,34]. However, it is well known that $(\boldsymbol{A}+\boldsymbol{B})^{-1} \neq (\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1})$. Therefore, we need to answer the second question: "Is it feasible for a nearly singular matrix $\boldsymbol{S}_W$ to be added with a tiny perturbation?"

Large attribute elements usually have a larger influence on the parameters of classifiers than those small do during learning courses; on the contrary, small class margins usually have a larger influence on the generalization performances of classifiers than those large do while making decisions [17,22,35]. The unduly large elements in a small part of attributes may make $\boldsymbol{S}_W$ close to singular, and the unduly small margins will increase the difficulty to seek the optimal separating hyperplanes. The above two cases will make the generalization performances of classifiers designed become poor. In a sense, feature representation is a crucial step for designing classifiers, regardless of whether they are linear or nonlinear.

Decimal (DEC) and binary (BIN) codes are two common feature representation systems. Normalized, proportional, logarithmic and sigmoid transformations are several popular equal-dimensional ones. SVMs enlarge the class margins by making the original data sparse in the higher-dimensional feature spaces through nonlinear transformations chosen in prior, e.g., polynomial and radial basis function (RBF) kernels [20,21]. Quoting the thought, we can transform a data from a lower-dimensional input space to a higher-dimensional feature space directly by coding in advance. The premise to do like this is that the original information must be preserved as much as possible [19]. Therefore, the third question to be answered is: "How to develop an effective feature representation system so as to enlarge the class margins as much as possible, lessen the within-class scatters and ease the unfavorably large components, on condition that the neighborhood relationships are approximately preserved?"

The solution of weight vectors in FLDs is in essence an analytic learning process, which is often faster than the iterative ones used in neural networks and SVMs [15,20]. In spite of an accustomed practice, the process of one-time analytic solution of weights is not certainly optimal. Therefore, the fourth question to be answered is: "How to develop an iterative learning algorithm in order to alleviate the imbalance and accordingly update the weights and thresholds by means of properly selecting a portion of the training samples?"

This paper aims to noticeably improve the classification accuracies of FLDs around answering the above-mentioned questions. In other words, this paper motivates to empirically optimize the weights and thresholds of FLDs to make the minimum-error-rate classification on the basis of Bayesian decision theory, from the heuristic point of view. The rest of this paper is organized as follows. Section 2 introduces the related work of FLDs. In Section 3, a series of empirical threshold formulas is proposed to alleviate the imbalance. Section 4 illustrates some mixed feature representation approaches and the condition for carrying out feature extraction by principal component analysis (PCA). Section 5 goes into details on the epoch-limited iterative learning strategy for further alleviating the imbalance. Section 6 presents numerous experimental results. Finally, Section 7 comes to our conclusions.

2. Related work

First of all, let us consider a two-class classification problem $\{\omega_1, \omega_2\}$ as well as the linear discriminant function. For a pattern $\boldsymbol{x}=(x_1, x_2, \ldots, x_m)^T \in R^m$ in the m-dimensional input space, the decision hyperplane π can be written as

$$\pi : f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0 = \boldsymbol{w}^T\boldsymbol{x} - \theta = 0 \tag{1}$$

where $\boldsymbol{w}=(w_1, w_2, \ldots, w_m)^T \in R^m$ is the weight vector, often called the normal or projected direction of π, and $\theta = -w_0$ is the threshold or bias.

The decision rule is

$$\begin{cases} \boldsymbol{x} \in \omega_1, & \text{if} \quad \boldsymbol{w}^T\boldsymbol{x} > \theta \\ \boldsymbol{x} \in \omega_2, & \text{if} \quad \boldsymbol{w}^T\boldsymbol{x} < \theta \\ \text{Indefinite}, & \text{if} \quad \boldsymbol{w}^T\boldsymbol{x} = \theta \end{cases} \tag{2}$$

Suppose the two input data matrices are marked as $\boldsymbol{X}^{(1)} \in R^{N1 \times m}$ with N_1 patterns in class ω_1 and $\boldsymbol{X}^{(2)} \in R^{N2 \times m}$ with N_2 in class ω_2, respectively, two mean vectors $\boldsymbol{\mu}_j \in R^m$ $(j=1, 2)$ are estimated by the training subset $\boldsymbol{X}_{12}=\{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\}$. The within-class scatter matrix $\boldsymbol{S}_W \in R^{m \times m}$ and the between-class scatter matrix $\boldsymbol{S}_B \in R^{m \times m}$ are defined by

$$\boldsymbol{S}_W = \boldsymbol{S}_1 + \boldsymbol{S}_2 = \sum_{j=1}^{2} \sum_{\boldsymbol{x}_p \in \omega_j} (\boldsymbol{x}_p - \boldsymbol{\mu}_j)(\boldsymbol{x}_p - \boldsymbol{\mu}_j)^T \tag{3}$$

$$\boldsymbol{S}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \tag{4}$$

An FLD seeks the projected direction $\boldsymbol{w}$ by maximizing the Fisher criterion function $J(\boldsymbol{w})$

$$\max_{\boldsymbol{w}} J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T\boldsymbol{S}_B\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w}} \tag{5}$$

which is also called the Rayleigh quotient [15,20].

As a matter of fact, the final vector $\boldsymbol{w}$ is only related to $\boldsymbol{S}_W$ and the difference $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, and is given by

$$\boldsymbol{w} = \boldsymbol{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{6}$$

If $\boldsymbol{S}_W$ is singular, the FLD will be no longer in force; and if approximately singular, the $\boldsymbol{w}$ obtained will not deserve trustworthy. Under these cases, in order to make the FLD work, $\boldsymbol{S}_W$ is often added with a tiny perturbation matrix, either a normal diagonal $diag(\xi_1, \xi_2, \ldots, \xi_m)\delta\boldsymbol{I}$ or directly a constant $\delta\boldsymbol{I}$, where $\boldsymbol{I} \in R^{m \times m}$ is an identity matrix, δ a regularization parameter [1,12,34,36,37], and $\xi_i \sim N(0, 1)$. The weight vector $\boldsymbol{w}$ thus becomes

$$\boldsymbol{w} = (\boldsymbol{S}_W + diag(\xi_i)\delta\boldsymbol{I})^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{7}$$

or often directly

$$\boldsymbol{w} = (\boldsymbol{S}_W + \delta\boldsymbol{I})^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{8}$$

Given the expected output for the pth training pattern $\boldsymbol{x}_p$ is d_p, the sum of squared errors between the expected and the real outputs for all the training patterns in $\{\omega_1, \omega_2\}$ is

$$E(\boldsymbol{w}, \theta) = \sum_{p=1}^{N_1+N_2} (d_p - f(\boldsymbol{x}_p))^2 = \sum_{p=1}^{N_1+N_2} (d_p - (\boldsymbol{w}^T\boldsymbol{x}_p - \theta))^2 \tag{9}$$

Let $\partial E(\boldsymbol{w}, \theta)/\partial\theta = 0$, the threshold θ is calculated by

$$\theta = \frac{1}{N_1+N_2} \sum_{p=1}^{N_1+N_2} (\boldsymbol{w}^T\boldsymbol{x}_p - d_p) = \frac{N_1\mu_{\boldsymbol{w}}^{(1)} + N_2\mu_{\boldsymbol{w}}^{(2)}}{N_1+N_2} - \frac{1}{N_1+N_2} \sum_{p=1}^{N_1+N_2} d_p \tag{10}$$