Pattern Recognition Letters 34 (2013) 545–551

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Super-class Discriminant Analysis: A novel solution for heteroscedasticity

Xinzhong Zhu

College of Mathematics, Physics, and Information Engineering, Zhejiang Normal University, Jin Hua, Zhejiang 321004, China

ARTICLE INFO

Article history:
Received 24 April 2012
Available online 29 November 2012

Communicated by Sarkar

Keywords:
Heteroscedasticity problem
Super-class
Super-class Discriminant Analysis
Divide and conquer

ABSTRACT

The *heteroscedasticity problem* is a great challenge in pattern recognition, particularly in statistics-based methods. The traditional method that is mainly used to solve this problem is heteroscedastic Discriminant Analysis. In this study, we propose a novel solution to the problem, called *Super-class Discriminant Analysis* (*SCDA*). Our method uses the "divide and conquer" methodology to partition the heteroscedastic dataset into super-classes with reduced heteroscedasticity and models them separately. Theoretically, a super-class should contain a set of classes having the same within-class variation. In practice, a heteroscedastic dataset can be coarsely divided into several super-classes based on certain semantic criteria such as gender or race. We evaluate our method with toy data and three real-world datasets, which can be divided into super-classes according to gender and race. Experimental results indicate that the proposed method can effectively resolve the problem of heteroscedasticity.

1. Introduction

Face recognition has been a topic of active research in computer vision and pattern recognition owing to the great challenges that it presents and its widespread applications in public security and private life. The main challenge of face recognition is the large variation in appearance caused by pose, illumination, expression, age, gender, race, partial occlusion, etc. Over the past several decades, researchers have proposed various methods to tackle these problems (as reviewed by Zhao et al. (2003)). These methods can be approximately divided into two categories: geometric-feature based methods and appearance-based methods. The former method exploits geometric features such as the relative position of facial components, contour, shape and size (Wiskott et al., 1997; Harmon et al., 1981). The latter method represents a face by descriptions that contain shape and texture information, such as Eigenface (Turk and Pentland, 1991), Fisherface (Brunelli and Poggio, 1993), and Independent Component Analysis (ICA) (Bartlett and Sejnowski, 1997). Appearance-based methods have attracted considerable attention (Turk and Pentland, 1991), and are dominant technologies in the field of face recognition (Brunelli and Poggio, 1993; Bartlett and Sejnowski, 1997).

Discriminant Analysis, one of the most popular appearance-based methods in face recognition, aims to find a low-dimensional subspace that can best distinguish classes. Discriminant Analysis methods can be categorized into Linear Discriminant Analysis and Non-linear Discriminant Analysis. Linear Discriminant Analysis (LDA) (Belhumeur et al., 1997) is widely used owing to its numerous advantages such as simplicity, good generalization, and effectiveness. LDA has been extended to different variants for various scenarios (Kan et al., 2011; Zhu and Martinez, 2006), and can also be easily extended to non-linear methods by Kernel tricks (Yang et al., 2005; Yang, 2002; Liu et al., 2002).

However, as a statistical learning method, LDA depends on the extent to which the model assumption matches the actual distribution. If the model assumption is a close approximation of the actual data distribution, the trained model will be robust and generalized; else, the model will deteriorate. In LDA, the within-class scatter of each class is assumed to follow the same distribution, but in practice, the within-class scatter of each class often differs owing to the varied appearances of a person. Thus, different classes follow complex heteroscedastic distributions. This poses one of the great challenges called *heteroscedasticity problem* in LDA. Heteroscedasticity indicates a biased estimation of the within-class scatter in LDA, which prevents the success of such methods in many real-world applications.

In recent years, many approaches have been proposed to tackle the heteroscedasticity problem in LDA, e.g., Mixture Discriminant Analysis (Hastie and Tibshirani, 1996). Some of these methods consider heteroscedasticity explicitly by embedding heteroscedastic within-class scatter in probabilistic distance measures such as Bhattacharyya distance (Bhattacharyya, 1943), Chernoff distance (Chernoff, 1952), and K–L divergence (Chernoff, 1952; Loog and Duin, 2004), which are designed for binary classification. In order to tackle heteroscedasticity in multi-classification, μ-measure distance (Nenadic, 2007) models each class by class-conditional Probability Density Function (PDF) and analytically estimates a mixture PDF for multi-classification. However, many samples are required for each class in order to accurately estimate each

E-mail address: znuzxz@gmail.com

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.11.006

class-conditional PDF, and the mixture PDF is estimated only by a mixture of multiple PDFs. Another effective method is Subclass Discriminant Analysis (SDA) (Zhu and Martinez, 2006). SDA first clusters each class into an equal number of subclasses, and then, replaces the class centers with the subclass centers to calculate the between-class scatter. This method obtains a more discriminative between-class scatter. However, its performance is greatly dependent on the clustering methods. Some nonparametric methods can also be exploited to diminish the impact of heteroscedasticity (Qiu and Wu, 2005; Li et al., 2005; Su et al., 2011, 2008; Yan et al., 2007).

It should be noted that, in some studies (Zhu and Martinez, 2006; Hastie and Tibshirani, 1996), heteroscedasticity implies that each class has a complex distribution, i.e., a mixture of multiple Gaussian distributions is assumed. This differs from the assumption of a simple, common, single Gaussian distribution in LDA. Further, in several other studies (Chernoff, 1952; Loog and Duin, 2004), heteroscedasticity implies that each class has a different distribution, thus differing from the assumption of the same distribution for all classes in LDA. In this study, we focus on applying LDA when each class has a different distribution.

All the above methods for the multi-class case try to reduce the heteroscedasticity in a unified model, such as computing the sum of the within-class scatter of all classes in LDA, increasing fineness of the between-class scatter (Zhu and Martinez, 2006), or using a mixture of multiple PDF (Nenadic, 2007).

This study proposes a different method, Super-class Discriminant Analysis (SCDA), which tackles the problem of heteroscedasticity using limited samples for each class, and can be employed in the small sample size scenario. The principle of this method is that the *heteroscedasticity problem* is caused by the difference in within-class scatter of different classes; therefore, the classes can be divided into homoscedastic groups and LDA can be applied separately to each group. Using this "divide and conquer" approach, in our study, a heteroscedastic data set is divided into several groups, called *super-classes*, each containing classes with the same within-class scatter. An LDA is applied to each homoscedastic super-class.

All existing methods directly model the heteroscedastic distribution of the whole data, which is difficult especially in case that only limited number of samples are available in the real applications, while our SCDA method divides this problem into several easier ones which can still work well in case of small number of samples scenario. Besides, most of the methods only consider the heteroscedasticity of the within-class variations, while our method considers both the heteroscedastic of the between-class variations and the within-class variations. Although SDA (Zhu and Martinez, 2006) exploited the sub-class to deal with the heteroscedasticity, however the total between-class scatter matrix is still a compromise of all heteroscedastic between-class variations, while our methods model each homoscedastic between-class variations independently, not a compromise.

The proposed SCDA has two advantages: (1) The classes in each super-class have the same or similar within-class scatter, and hence, the within-class variations can be estimated more accurately with more consistent samples (see Section 4.1 for more details). (2) Although classes in each super-class are harder to distinguish owing to their greater similarity, the LDA specific to each super-class implies a more elaborate discriminative model for these classes, and thus, a greater potential for correct classification.

The remainder of this paper is organized as follows: In Section 2, SCDA is formulated. Section 3 consists of the evaluation of the degree of heteroscedasticity of a real-world dataset. The results of applying SCDA to AR, FERET and CAS-PEAL face databases are presented in Section 4. Section 5 contains the conclusion.

2. Super-class Discriminant Analysis

As mentioned above, heteroscedasticity is a great challenge in LDA. In this section, we will first illustrate LDA in detail. Next, we will analyze heteroscedasticity in the context of super-class. Finally, SCDA will be described in detail.

2.1. Linear Discriminant Analysis in case of heteroscedasticity

2.1.1. Linear Discriminant Analysis

LDA (Belhumeur et al., 1997) was proposed with the objective of determining a projection matrix to project the samples to a low-dimensional space in which the samples can be better distinguished by maximizing the Fisher discrepancy. It is illustrated as follows:

$$W_{opt} = \arg\max_{W} trace\left(\frac{WS_bW^t}{WS_wW^t}\right), \tag{1}$$

$$S_w = \sum_{i=1}^{c} p(\omega_i)S_i = \sum_{i=1}^{c} p(\omega_i)\sum_{j=1}^{N_i}(x_{ij} - m_i)(x_{ij} - m_i)^t, \tag{2}$$

$$S_b = \sum_{i=1}^{c} p(\omega_i)(m_i - m)(m_i - m)^t. \tag{3}$$

Eq. (3) can also be expressed in an equivalent form:

$$S_b = \sum_{i=1}^{c}\sum_{j=1}^{c} p(\omega_i)p(\omega_j)(m_i - m_j)(m_i - m_j)^t, \tag{4}$$

where c is the number of classes, $p(\omega_i)$ is the prior probability of class i, S_i is the within-class scatter of class i, x_{ij} is the jth sample in class i, m_i is the mean of class i, N_i is the number of samples in class i, and m is the mean of all samples.

LDA assumes that each class has the same within-class scatter. Based on this assumption, S_i in Eq. (2) can be considered as a sample from the distribution of the actual within-class scatter. Then, Eq. (2) represents the sample mean, i.e., a sample-based estimation of the actual within-class scatter, which is usually an unbiased estimation. The within-class scatter is assumed to be the same, and hence, the differences between classes lie mainly at the centers of each class. Therefore, Eq. (3) can reflect between-class discrepancy well. Thus, the LDA is successful in applications that meet the assumption.

2.1.2. Linear Discriminant Analysis in case of heteroscedasticity

However, in general, real-world data does not meet this assumption, i.e., large variations cause the within-class scatter of each class to be unequal. Therefore, the above formulation may lead to inaccurate modeling of the distribution of the whole data.

For heteroscedastic data, let us assume that it can be divided into n groups, and each group, called a super-class, contains classes having the same within-class scatter. Then, we can rewrite Eq. (2) as follows:

$$\begin{aligned} S_w &= \sum_{k=1}^{n} S_w(k) = \sum_{k=1}^{n}\sum_{i=1}^{c_k} p(\omega_{ki})S_{ki} \\ &= \sum_{k=1}^{n}\sum_{i=1}^{c_k} p(\omega_{ki})\sum_{j=1}^{N_{ki}}(x_{kij} - m_{ki})(x_{kij} - m_{ki})^t, \end{aligned} \tag{5}$$

where $S_w(k)$ is the within-class scatter of the kth super-class and c_k is the number of classes in the super-class k. It can be observed that all $S_w(k)$ ($i = 1, \ldots, n$) is heteroscedastic, and Eq. (5) computes the mean within-class scatter of all super-classes, irrespective of the heteroscedasticity. This is an approximation of the n heteroscedastic