Pattern Recognition 47 (2014) 359–367

Contents lists available at ScienceDirect

Pattern Recognition

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

Discriminant analysis and similarity measure

Chengjun Liu *

Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, United States

ARTICLE INFO

Article history:
Received 4 June 2012
Received in revised form
18 June 2013
Accepted 23 June 2013
Available online 29 June 2013

Keywords:
Bayes decision rule
Cosine similarity measure
Discriminant analysis
Face recognition
Face Recognition Grand Challenge (FRGC)
New similarity measure
Normalized correlation
Pattern recognition

ABSTRACT

The cosine similarity measure is often applied after discriminant analysis in pattern recognition. This paper first analyzes why the cosine similarity is preferred by establishing the connection between the cosine similarity based decision rule in the discriminant analysis framework and the Bayes decision rule for minimum error. The paper then investigates the challenges inherent of the cosine similarity and presents a new similarity that overcomes these challenges. The contributions of the paper are thus three-fold. First, the application of the cosine similarity after discriminant analysis is discovered to have its theoretical roots in the Bayes decision rule. Second, some inherent problems of the cosine similarity such as its inadequacy in addressing distance and angular measures are discussed. Finally, a new similarity measure, which overcomes the problems by integrating the absolute value of the angular measure and the l_p norm (the distance measure), is presented to enhance pattern recognition performance. The effectiveness of the proposed new similarity measure in the discriminant analysis framework is evaluated using a large scale, grand challenge problem, namely, the Face Recognition Grand Challenge (FRGC) problem. Experimental results using 36,818 FRGC images on the most challenging FRGC experiment, the FRGC Experiment 4, show that the new similarity measure improves face recognition performance upon other popular similarity measures, such as the cosine similarity measure, the normalized correlation, and the Euclidean distance measure.

1. Introduction

Different feature extraction methods usually apply different similarity measures for achieving good pattern recognition performance [19,29,20,4,18,23,31,21,16,24]. Generally speaking, there are two major types of feature extraction methods: feature extraction for signal representation and feature extraction for classification [6,3]. A popular method for the first type is Principal Component Analysis (PCA), which is the optimal feature extraction method in the sense of mean-square error [11,3]. A broadly applied method for the second type is Discriminant Analysis (DA), which derives the optimal features from criteria defined on scatter matrices [6,3]. The PCA based feature extraction methods, such as the Eigenfaces method [30,12], often apply the whitened cosine similarity measure for achieving good pattern recognition performance [2,24]. The discriminant analysis based methods, such as the Fisherfaces method [1,28,5], however, often utilize the cosine similarity measure for improving pattern recognition performance [18,23].

For the PCA based feature extraction methods, we have investigated in our previous research [15,16] why the whitened cosine similarity measure achieves good pattern recognition performance, and presented new similarity measures to improve upon the whitened cosine similarity for better pattern recognition performance [15]. In particular, we have studied the connection between the popular whitened cosine similarity measure and the Bayes decision rule and derived new similarity measures, such as the PRM Whitened Cosine (PWC) similarity measure and the Within-class Whitened Cosine (WWC) similarity measure, for further enhancing pattern recognition performance for the PCA based feature extraction methods [15].

For the discriminant analysis based feature extraction methods, we will investigate in this paper the following issues. First, we will analyze why the cosine similarity measure helps the discriminant analysis based feature extraction methods improve pattern recognition performance. Second, we will discuss the problems inherent of the cosine similarity measure that reduce its discriminatory power. And finally, we will present a new similarity measure that overcomes the problems of the cosine similarity measure for enhancing pattern recognition performance of the discriminant analysis based feature extraction methods.

Specifically, in order to show why the cosine similarity measure helps the discriminant analysis based feature extraction methods achieve good pattern recognition performance, we analyze discriminant analysis through the simultaneous diagonalization of the within-class scatter matrix and the between-class scatter matrix [6,13]. The first step of the simultaneous diagonalization

* Tel.: +1 9735965280.
E-mail address: chengjun.liu@njit.edu

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.06.023

reveals that applying the cosine similarity measure after whitening the within-class scatter matrix in discriminant analysis is connected to the Bayes decision rule for minimum error under some specific assumptions. The second step of the simultaneous diagonalization diagonalizes the transformed between-class scatter matrix, and the cosine similarity measure is applied after the diagonalization step for pattern classification in discriminant analysis. As a result, the advantage of applying the cosine similarity measure after discriminant analysis comes from its connection to the Bayes decision rule, as the Bayes classifier is the optimal one for minimizing the classification error.

The challenges inherent of the cosine similarity measure come from its inadequacy in addressing both the distance and the angular measures. The inadequacy in addressing the distance measure arises because the cosine similarity measure fails to account for the actual distance between two pattern vectors. The inadequacy in addressing the angular measure occurs when the angle between the pattern vectors is greater than $\pi/2$. Both problems lead to incorrect classification when applying the cosine similarity measure in the discriminant analysis based feature extraction methods. To overcome the angle problem of the cosine similarity measure, researchers propose the normalized correlation measure [27], which in essence takes the absolute value of the cosine similarity. The normalized correlation measure, however, displays its own disadvantages that often deteriorate pattern classification performance.

As both the cosine similarity measure and the normalized correlation display their intrinsic problems, we present in this paper a new similarity measure for improving pattern classification performance. The new similarity measure improves upon the cosine similarity by integrating the absolute value of the angular measure and the l_p norm (the distance measure). As a result, the new similarity measure can classify the pattern vectors that neither the cosine similarity measure nor the normalized correlation is able to classify correctly. The new similarity measure, therefore, improves upon the cosine similarity measure and the normalized correlation for the discriminant analysis based feature extraction methods.

The effectiveness of the proposed new similarity measure is evaluated using a large scale, grand challenge problem, namely, the Face Recognition Grand Challenge (FRGC) problem [25]. Face recognition is a representative pattern recognition problem due to the complexity of the problem itself and its broad applications in the commercial and government sectors [17,22,15,25]. We apply the most challenging FRGC version 2 Experiment 4, which contains 12,776 training images, 16,028 controlled target images, and 8014 uncontrolled query images, to assess our proposed new similarity measure. Experimental results show that our proposed new similarity measure improves face recognition performance upon the cosine similarity measure, the normalized correlation, and the Euclidean distance measure for the discriminant analysis feature extraction method.

2. Within-class scatter matrix induced similarity measure

Different similarity measures are applied after different feature extraction methods in order to achieve good pattern recognition performance. In contrast to the PCA feature extraction method that often uses the whitened similarity measures [25,15], the discriminant analysis feature extraction method often applies the cosine similarity measure for improving pattern recognition performance [14,25]. We now discuss why the cosine similarity measure is preferred to other popular similarity measures after discriminant analysis for pattern recognition.

Discriminant analysis is concerned with extracting features that are most effective for class separability. In theory, the Bayes classifier is the optimal classifier, hence the Bayes error should be the optimal criterion for evaluating feature effectiveness. In practice, however, the accurate estimation of the conditional probability density function of each class is very difficult, if not impossible. As a result, the application of the Bayes classifier or the Bayes error is constrained by the estimation of the conditional probability density functions. Discriminant analysis, which optimizes the criteria of class separability based on scatter matrices, provides an alternative for effective feature extraction. Specifically, let $\mathcal{X}$ be a pattern vector in a d-dimensional space: $\mathcal{X}\in\mathbb{R}^d$, and $\mathcal{X}$ belongs to one of the predefined L classes: $\omega_1, \omega_2, \ldots, \omega_L$. The within-class scatter matrix, $S_w\in\mathbb{R}^{d\times d}$, and the between-class scatter matrix, $S_b\in\mathbb{R}^{d\times d}$, may be defined as follows [6]:

$$S_w = \sum_{i=1}^{L} P(\omega_i)\mathcal{E}\{(\mathcal{X}-\mathbf{M}_i)(\mathcal{X}-\mathbf{M}_i)^t|\omega_i\} \quad (1)$$

$$S_b = \sum_{i=1}^{L} P(\omega_i)(\mathbf{M}_i-\mathbf{M}_0)(\mathbf{M}_i-\mathbf{M}_0)^t \quad (2)$$

where $\mathcal{E}\{(\mathcal{X}-\mathbf{M}_i)(\mathcal{X}-\mathbf{M}_i)^t|\omega_i\} = \int(\mathcal{X}-\mathbf{M}_i)(\mathcal{X}-\mathbf{M}_i)^t p(\mathcal{X}|\omega_i)\, d\mathcal{X}$, $\mathbf{M}_i$ is the mean vector of class ω_i, $\mathbf{M}_0$ is the grand mean vector, and $P(\omega_i)$ and $p(\mathcal{X}|\omega_i)$ are the prior probability and the conditional probability density function of class ω_i, respectively. Discriminant analysis derives the most discriminating features by selecting the eigenvectors corresponding to the largest eigenvalues of $S_w^{-1}S_b$ as the projection vectors [6]. Alternatively, these projection vectors can be obtained by means of the simultaneous diagonalization of the within-class scatter matrix, S_w, and the between-class scatter matrix, S_b: step 1 — whitening the within-class scatter matrix S_w and step 2 — diagonalizing the transformed between-class scatter matrix S [6]. Next, we briefly review these two steps in order to derive the within-class scatter matrix induced similarity measure.

2.1. Whitening the within-class scatter matrix

The first step of the simultaneous diagonalization for discriminant analysis is to whiten the within-class scatter matrix S_w. As S_w is a real symmetric $d\times d$ matrix, there exist d real eigenvalues and d real eigenvectors, such that S_w can be factorized as follows:

$$S_w = \Phi\Lambda\Phi^t \quad (3)$$

where $\Phi\in\mathbb{R}^{d\times d}$ is the orthogonal eigenvector matrix of S_w, and $\Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_d\}\in\mathbb{R}^{d\times d}$ is the diagonal eigenvalue matrix of S_w with diagonal elements in decreasing order: $\lambda_1\geq\lambda_2\cdots\geq\lambda_d$. If we define a whitening transformation matrix W

$$W = \Phi\Lambda^{-1/2} \quad (4)$$

then the within-class scatter matrix S_w is whitened to an identity matrix and the between-class scatter matrix S_b is transformed to a new matrix S as follows:

$$W^tS_wW = \Lambda^{-1/2}\Phi^tS_w\Phi\Lambda^{-1/2} = I \quad (5)$$

$$W^tS_bW = \Lambda^{-1/2}\Phi^tS_b\Phi\Lambda^{-1/2} = S \quad (6)$$

Note that if there are enough samples, the within-class scatter matrix S_w is always full rank, and its eigenvalue matrix Λ is full rank and invertible. But if due to limited number of samples, S_w is not full rank, then its eigenvalue matrix Λ contains at least one eigenvalue whose value is zero. As a result, the eigenvalue matrix Λ is no longer invertible, and the whitening transformation matrix W is not defined. To overcome this difficulty, we can either apply a dimensionality reduction method as the Fisherfaces method does [1], or we can introduce a small positive regularization number to