Knowledge-Based Systems 99 (2016) 112–122

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Optimal feature extraction methods for classification methods and their applications to biometric recognition

Jun Yin*, Weiming Zeng, Lai Wei

College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

ARTICLE INFO

Article history:
Received 24 September 2015
Revised 6 January 2016
Accepted 30 January 2016
Available online 8 February 2016

Keywords:
Nearest feature line
Sparse representation
Linear discriminant analysis
Discriminative projection

ABSTRACT

Classification is often performed after feature extraction. To improve the recognition performance, we could develop the optimal feature extraction method for a classification method. In this paper, we propose three feature extraction methods Discriminative Projection for Nearest Neighbor (DP-NN), Discriminative Projection for Nearest Mean (DP-NM) and Discriminative Projection for Nearest Feature Line (DP-NFL), which are optimal for classification methods Nearest Neighbor (NN), Nearest Mean (NM) and Nearest Feature Line (NFL), respectively. We also prove that DP-NN and DP-NM are equivalent to Linear Discriminant Analysis (LDA) under a certain condition. In the experiments, LDA, DP-NFL and SRC steered Discriminative Projection (SRC-DP) are used for feature extraction and then the extracted features are classified by NN, NM, NFL, Sparse Representation based Classification (SRC) and Collaborative Representation Classifier (CRC). Experimental results of biometric recognition show that the proposed DP-NFL performs well, and that combining an effective classification method with the optimal feature extraction method for it can perform best.

1. Introduction

Feature extraction and classification are two essential procedures of pattern recognition. The original data usually have high dimensionality. Due to time-consuming and curse of dimensionality [1], it is not suitable to recognize these data directly. Feature extraction could reduce dimensionality and get the most important features simultaneously. In the past few decades, numerous feature extraction techniques were proposed. Principal Component Analysis (PCA) [2] and Linear Discriminant Analysis (LDA) [3] are two of the most classical methods. PCA is an unsupervised method which projects the data along the direction of maximal variance. LDA is a supervised method, which seeks the projection that maximizes the between-class scatter and minimizes the within-class scatter simultaneously.

PCA and LDA are linear feature extraction methods, and they cannot discover the underlying manifold structure of the data. After that, a number of manifold learning methods were proposed to analyze the high dimensional data which lie on or near a low dimensional manifold, such as Locally Linear Embedding (LLE) [4], Isometric Mapping (Isomap) [5], Laplacian Eigenmaps (LE) [6] and Local Tangent Space Alignment (LTSA) [7]. These manifold learning methods are effective in representing the nonlinear data, but they are only defined in the training samples and cannot find the low dimensional representation of new test samples. Therefore, they cannot be directly applied to the pattern recognition problem. Neighborhood Preserving Embedding (NPE) [8], Isometric Projection (IsoProjection) [9], Locality Preserving Projections (LPP) [10] and Linear Local Tangent Space Alignment (LLTSA) [11] solve this problem by acquiring the explicit projections from original high dimensional space to low dimensional embedding space, and they can be seen as the linear version of LLE, Isomap, LE and LTSA. However, these linear feature extraction methods based on manifold learning are unsupervised, and they are designed to preserve the locality of samples in the low dimensional space rather than good discriminating ability. To increase discriminating ability, some supervised feature extraction methods based on manifold learning were proposed. The representative methods include Local Discriminant Embedding (LDE) [12], Marginal Fisher Analysis (MFA) [13], Discriminant Simplex Analysis (DSA) [14], Neighborhood Discriminant Projection (NDP) [15], Local Sensitive Discriminant Analysis (LSDA) [16] and multi-manifold discriminant analysis (MMDA) [17]. These methods thoroughly consider the within-class information and the between-class information.

For the feature extraction methods based on manifold learning, it is unclear to select the neighborhood size and define the adjacent weight matrix which are key problems for these methods. Recently, with the development of sparse representation theory [18–22], some feature extraction methods used sparse representation coefficients to construct the adjacent weight matrix.

* Corresponding author. Tel.: +86 02138282823.
E-mail address: yinjun8429@163.com, junyin@shmtu.edu.cn (J. Yin).

http://dx.doi.org/10.1016/j.knosys.2016.01.043
0950-7051/

Sparsity Preserving Projection (SPP) [23] and Sparse Neighborhood Preserving Embedding (SNPE) [24] are the representative methods. Both SPP and SNPE use sparse representation coefficients to construct the adjacent weight matrix. Unlike manifold learning based feature extraction methods, these sparse representation based methods do not need to select the model parameters during construction of the adjacent weight matrix. In order to improve the recognition ability, some research works introduced class label information to sparse representation, such as Discriminant Sparsity Preserving Embedding (DSPE) [25], Discriminant Sparse Neighborhood Preserving Embedding (DSNPE) [26, 27] and Weighted Discriminative Sparsity Preserving Embedding (WDSPE) [28].

Classification is an indispensable procedure for recognizing the data. After feature extraction, a classification method should be used to recognize these extracted features. Several classification approaches have been proposed over the past few decades [29, 30]. Among them, the classification methods Nearest Neighbor (NN) and Nearest Mean (NM) are widely used because of their simpleness and availability. NN and NM are based on the nearest distance between two data points, while the classification method Nearest Feature Line (NFL) [31] is based on the nearest distance from the data point to the feature line passing through two data points. Motivated by sparse representation, Wright et al. proposed Sparse Representation based Classification (SRC) [18] for classification. SRC assigns a test sample to the class which has the minimal sparse reconstruction residual. L1-norm was adopted in sparse representation. Zhang et al. pointed out that collaborative representation strategy is more important than the L1-norm-based sparsity constraint and proposed Collaborative Representation Classifier (CRC) [32]. Yin et al. proposed Kernel Sparse Representation based Classification (KSRC) [33]. They considered that the data in a higher dimensional space probably have better linear separability and performed SRC in the transformed higher dimensional space by the kernel trick.

Recently, based on SRC, Yang et al. proposed SRC steered Discriminative Projection (SRC-DP) [34, 35]. They tried to make the feature extraction method has a direct connection to SRC. The connection between SRC and SRC-DP is constructed by utilizing sparse representation coefficients in the transformed low dimensional space. We know classification is performed after feature extraction. Therefore, to improve the recognition performance, we could strengthen the connection between classification and feature extraction and make the feature extraction method fit the classification method as much as possible. This idea is not only for SRC. In this paper, we start from classification methods NN, NM and NFL, and develop feature extraction methods Discriminative Projection for NN (DP-NN), Discriminative Projection for NM (DP-NM) and Discriminative Projection for NFL (DP-NFL) which are optimal for NN, NM and NFL, respectively. For DP-NN and DP-NM, we also prove that they are equivalent to LDA under a certain condition. For a pattern recognition system, it's a good scheme to combine an effective classification method with the optimal feature extraction method for this classification method. Experimental results of biometric recognition prove this viewpoint.

The remainder of this paper is organized as follows: Section 2 briefly reviews the related classification and feature extraction methods. Section 3 describes DP-NN, DP-NM algorithms and gives the relationship among DP-NN, DP-NM and LDA. DP-NFL algorithm is also presented in this section. Section 4 presents experiments and discussions of the experimental results. The conclusions are summarized in Section 5.

The important notations used throughout the rest of the paper are listed in Table 1.

Table 1
Summary of the notations.

Notations	Description
n	Number of training samples
n_i	Number of training samples in class i
c	Number of classes
d	Dimension of samples
m	Center of all the training samples
m_i	Center of training samples in class i
$X = [X_1, X_2, \ldots, X_c] \in R^{d\times n}$	Training sample matrix
$X_i = [x_{i1}, x_{i2}, \ldots, x_{in_i}] \in R^{d\times n_i}$	Training sample matrix of class i

2. Related classification and feature extraction methods

In this section, we first review three classification methods: NM, NN and NFL and then outline the popular feature extraction method LDA.

2.1. Classification methods

NM:

For a test sample y, NM calculates the distance between y and the center of each class $m_i (i = 1, 2, \ldots, c)$, where $m_i = (\sum_{i=1}^{n_i} x_{in_i})/n_i$. y is assigned to the class with the minimal distance:

$$identity(y) = \arg\min_i \|y - m_i\|_2 \tag{1}$$

NN:

For a test sample y, NN calculates the distance between y and each training sample $x_{ij} (i = 1, 2, \ldots, c;\ j = 1, 2, \ldots, n_i)$. y is assigned to the class that the training sample with the minimal distance belongs to :

$$identity(y) = \arg\ \min_i \left\| y - x_{ij} \right\|_2 \tag{2}$$

NFL:

For a test sample y, NFL calculates the distance from y to each feature line. The feature line is obtained by connecting every two training samples from the same class. Suppose x_{ij} and x_{ik} are two training samples from class i, the feature line passing through x_{ij} and x_{ik} can be denoted as $\overline{x_{ij}x_{ik}}$ and the distance from y to $\overline{x_{ij}x_{ik}}$ is denoted as $d(y, \overline{x_{ij}x_{ik}})$. To compute $d(y, \overline{x_{ij}x_{ik}})$, y is projected onto $\overline{x_{ij}x_{ik}}$ as point $\tilde{y}$ and we have $d(y, \overline{x_{ij}x_{ik}}) = \|y - \tilde{y}\|_2$. $\tilde{y}$can be computed as:$\tilde{y} = x_{ij} + \mu(x_{ik} - x_{ij})$, where $\mu = (y - x_{ij})^T (x_{ik} - x_{ij})/(x_{ik} - x_{ij})^T (x_{ik} - x_{ij})$. The test sample y is assigned to the class which has the minimal $d(y, \overline{x_{ij}x_{ik}})$:

$$identity(y) = \arg\ \min_i d(y, \overline{x_{ij}x_{ik}}) \tag{3}$$

where $1 \le i \le c$ and $1 \le j < k \le n_i$.

2.2. LDA

LDA tries to find the projection by maximizing the between-class scatter and minimizing the within-class scatter simultaneously. The between-class scatter matrix S_b, the within-class scatter matrix S_w and the global scatter matrix S_t can be defined by

$$S_b = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^T \tag{4}$$

$$S_w = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_{ij} - m_i)(x_{ij} - m_i)^T \tag{5}$$