Neurocomputing 188 (2016) 225–232

Contents lists available at ScienceDirect

Neurocomputing

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

Feature extraction using maximum nonparametric margin projection

Bo Li [a,b,c,*], Jing Du [a,b], Xiao-Ping Zhang [c,d]

[a] School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, Hubei 430065, China
[b] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan, Hubei 430065, China
[c] Department of Electrical and Computer Engineering, Ryerson University, Toronto, Ontario, Canada M5B 2K3
[d] School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China

ARTICLE INFO

Article history:
Received 29 September 2014
Received in revised form
31 October 2014
Accepted 8 November 2014
Available online 4 December 2015

Keywords:
Dimensionality reduction
Nonparametric margin
Subspace learning
Feature extraction

ABSTRACT

Dimensionality reduction is often recommended to handle high dimensional data before performing the tasks of visualization and classification. So far, large families of dimensionality reduction methods besides the supervised or the unsupervised, the linear or the nonlinear, the global or the local have been developed. In this paper, a maximum nonparametric margin projection (MNMP) method is put forward to extract features from original high dimensional data. In the proposed method, we offer some nonparametric or local definitions to the traditional between-class scatter and within-class scatter, which contributes to remove the disadvantage that linear discriminant analysis (LDA) can not be well-performed in the cases of non-Gaussian distribution data. Based on the predefined between-class scatter and the within-class scatter, a nonparametric margin can be reasoned to avoid the small sample size (SSS) problem. Moreover, the proposed nonparametric margin will be maximized to explore a discriminant subspace. At last, we have conducted experiments on some benchmark data sets such as Palmprint database, AR face database and Yale face database. In addition, performance comparisons have also been made to some related feature extraction methods including LDA, nonparametric discriminant analysis (NDA) and local graph embedding based on maximum margin criterion (LGE/MMC). Experimental results on these data sets have validated that the proposed algorithm is effective and feasible.

1. Introduction

When dealing with pattern classification problems, original data always contain a great deal of information [1]. Among them, some are useful, while some are superfluous and some are even noisy. In addition, the patterns hidden in these data are always represented to high dimensional vectors. For example, a 32-by-32 appearance-based image can be viewed as a 1024-dimensional point in image space [2]. So how to extract discriminant information from original complicated data and how to represent original data with low dimensional vectors are playing more and more important role in data classification. The resulted reasons can be concluded as follows. One is that the recognition accuracy will be greatly degradated because of disturbance of noise; the other lies in that the computational cost will be very expensive if original data are directly involved in classification. It is a typical way to solve the problem using dimensionality reduction techniques. Usually, dimensionality reduction serves as automatic learning to extract features with high efficiency. Due to the property, besides pattern recognition, the topics of dimensionality reduction also appears in many other fields including data mining, computer vision, information retrieval, machine learning and bioinformatics [3,4].

Currently, researchers have developed many classical dimensionality reduction approaches, which can be categorized into two kinds based on whether the class information is considered, i.e., the supervised or the unsupervised. They are also broadly partitioned into linear methods [5–9] and nonlinear models besides artificial neural networks (ANN) [10–13]. Linear dimensionality reduction techniques try to seek a meaningful low dimensional subspace with a linear transformation, where a compact representation of input data can be provided. Among all the linear dimensionality reduction methods, principal component analysis (PCA) [5–7] and linear discriminant analysis (LDA) [6,8,9] are most well-known.

Generally speaking, PCA projects original data into a low dimensional space spanned by the eigenvectors associated to the largest eigenvalues of covariance matrix of all samples, where PCA is the optimal representation of input data in the sense of minimizing mean squared error (MSE) [14,15]. However, PCA is completely unsupervised with regard to data class information, which may result in much discriminant information missing and

* Corresponding author.
E-mail address: liberol@126.com (B. Li).

http://dx.doi.org/10.1016/j.neucom.2014.11.105
0925-2312/

recognition ability weakening, especially encountering a large number of sample points [6].

Unlike PCA, LDA takes full consideration of patterns' class information. It is believed that feature extraction methods under supervised learning will be more discriminative. Thus LDA can enhance data separability more than PCA. LDA maps original data into an optimal subspace by a linear transformation. The linear transformation matrix consists of the corresponding eigenvectors which can maximize the trace ratio of the between-class scatter to the within-class scatter.

To improve the performance of the original LDA, many modified versions based on LDA have been reported, which are validated to be efficient [15–31]. Moreover, when constructing the between-class scatter and the within-class scatter, most of LDA-based methods inherit the parametric formulations in traditional LDA, which heavily depend on the fact that samples in each class should distribute as Gaussian function. However, it is not always true for all the data. In case of non-Gaussian distribution data, these LDA-based methods will lead to performance degradation. In order to overcome the problem, a nonparametric definition on the between-class scatter is presented, where boundary samples are explicitly exploited [8]. Instead of class centers, the new defined between-class scatter is formulated on the whole training set, where their contributions to discrimination are justified by the weights between inter-class sample pairs. Thus the feature extraction algorithm adapts more to those samples without taking into account data distribution model. However, the above nonparametric definition is restricted to data classification for two-class cases rather than multi-class. So Li et al. propose a nonparametric discriminant analysis (NDA) algorithm to deal with multi-class problem [32], where a new between-class scatter is advanced by extending the definition of the original nonparametric between-class scatter matrix. But the proposed NDA just pays attention to local structure information of boundary points, which can not explore all the local structure hidden in the intra-class data.

During last decade, many manifold learning approaches have been developed for nonlinear dimensionality reduction. Besides isometric mapping (ISOMAP) [33], locally linear embedding (LLE) [34], Laplacian eigenmaps (LE) [35], local tangent space alignment (LTSA) [36], maximum variance unfolding (MVU) [37] and Riemannian manifold learning (RML) [38] are their representatives. It has been shown by many examples that these methods have yielded impressive results on artificial and real world data sets. Compared to other dimensionality reduction methods, on the one hand, manifold learning can probe the essential dimensions of manifold embedded in high dimensional space, on the other hand, manifold learning expects to embed original data into a lower dimensional space by locality preserving, where the locality can be approached using k nearest neighbors (KNN) criterion. It is the characteristics of exploring manifold structure locally that manifold learning can be employed to handle data either in the non-Gaussian cases or in the Gaussian cases. Enlightened by the idea of local learning in manifold learning, both the between-class scatter and the within-class scatter can be locally modeled, based on which the proposed feature extraction method will be robustness to data with all kinds of distributions, especially non-Gaussian distribution.

In addition, the above nonparametric versions of LDA still suffer small sample size (SSS) problem, which often occurs to LDA due to case of the limited training set with high dimensionality. Under such circumstance, the within-class scatter is not positive-definite, which leads to serious instability and over-fitting. Until to now, many attempts have been tried, among which maximum margin criterion (MMC) shows its superiority [39]. On the basis of MMC, Qiu et al. present a nonparametric maximum margin criterion (NMMC) for face features extraction [40], where the between-class scatter is defined using the nearest inter-class neighbor pair. But for the within-class scatter, it is not locally constructed because the furthest intra-class points are involved. Meanwhile, the within-class scatter is formulated by the difference between the furthest intra-class data pair, which fails to explore the local structure in the intra-class data. Recently, many manifold learning based maximum margin criterion algorithms have been presented to extract discriminant features from original data, where the local scatter and the non-local scatter instead of the between-class scatter and the within-class scatter are discriminatively advanced [41–45].

In this paper, we propose a maximum nonparametric margin projection (MNMP) method for feature extraction, which tries to overcome the problems in the traditional LDA. In the proposed MNMP, different to the existing nonparametric between-class scatter and within-class scatter, we offer a novel nonparametric or local definition on them, where all the training samples rather than only the boundary points are contained, as a result, data distribution model will not be considered. At the same time, the newly defined nonparametric between-class scatter and within-class scatter are also introduced to reason a margin criterion, which will be maximized to explore a discriminant subspace.

The rest of the paper is organized as follows: Section 2 simply reviews LDA and MMC. In Section 3, the between-class scatter and the within-class scatter is parametric or locally defined, based on which a new margin criterion can be nonparametric deduced, and then the proposed MNMP is addressed in details. Compared to some related feature extraction methods such as LDA, nonparametric discriminant analysis (NDA) and local graph embedding based on maximum margin criterion (LGE/MMC), experimental results on Palmprint data, AR face data and Yale face data are offered in Section 4. At last, the paper is finished with some conclusions in Section 5.

2. Related work

2.1. Notations

The main notations used in the whole paper are summarized as follows:

- The high-dimensional input sample points will be denoted as $X_1, X_2, \ldots, X_n$. Sometimes it will be convenient to work with these sample points as a single matrix $X = [X_1, X_2, \ldots, X_n] \in \Re^{D \times n}$.
- The low-dimensional representations of $X_1, X_2, \ldots, X_n$ are represented as $Y_1, Y_2, \ldots, Y_n$. And the matrix form of these points is $Y = [Y_1, Y_2, \ldots, Y_n] \in \Re^{d \times n}$ $(d < < D)$.
- n is the number of all the sample points.
- D is the dimension of input sample points.
- d is the dimension of output samples.
- k is the number of the nearest neighbors used by a particular algorithm such as k nearest neighbors.
- m_i is the i-th class mean.
- m is the mean of all the sample points.
- c is the number of the class labels.
- n_i denotes the sample number of the i-th class.
- S_B is the between-class scatter.
- S_W denotes the within-class scatter.

2.2. LDA

LDA aims to look for a linear subspace W, within which the projections of the samples from different classes are more apart, as