Neurocomputing 138 (2014) 283–296

Contents lists available at ScienceDirect

Neurocomputing

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

Dimensionality reduction using graph-embedded probability-based semi-supervised discriminant analysis

Wei Li [a,b], Qiuqi Ruan [a,b], Jun Wan [a,b]

[a] Institute of Information Science, Beijing Jiaotong University, Beijing, PR China
[b] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, PR China

ARTICLE INFO

Article history:
Received 13 June 2013
Received in revised form
30 January 2014
Accepted 1 February 2014
Communicated by Shiguang Shan
Available online 25 February 2014

Keywords:
Semi-supervised dimensionality reduction
Feature extraction
Discriminant analysis
Graph embedding
Class membership probability
Face recognition
Facial expression recognition

ABSTRACT

Probabilistic semi-supervised discriminant analysis (PSDA) is a recently proposed semi-supervised dimensionality reduction approach. It quantifies class membership probability to each unlabeled sample by using a well-designed soft assignment technique. Then discriminant analysis is performed over both labeled and unlabeled data which bears an analogy to the Fisher criterion. However, PSDA mainly focuses on discriminative information hidden in unlabeled data and ignores the local geometric information which is critical to reveal the intrinsic distribution of data points, especially for face image data. In this paper, we develop a graph-based semi-supervised learning method based on PSDA, termed as graph-embedded probability-based semi-supervised discriminant analysis (GPSDA) for dimensionality reduction. By introducing a similarity measurement of fuzzy sets to investigate the inexact class information of unlabeled data, an adjacency graph is modeled based on both neighborhood structure and category information, which is more relevant to classification compared with the unsupervised graph constructed in traditional graph-based semi-supervised dimensionality reduction technique. Since more information is learnt from unlabeled data, GPSDA is expected to enhance performance in classification task. We present experimental evidence on face and facial expression recognition suggesting that our algorithm is able to use unlabeled data effectively.

© 2014 Published by Elsevier B.V.

1. Introduction

With the fast growing amounts of high-dimensional data which may cause the "curse of dimensionality" as well as computational complexity, dimensionality reduction becomes a crucial step in many practical machine learning tasks. The objective of dimensionality reduction is to find a low-dimensional feature representation for data compression and enhance discrimination for subspace classification. Commonly used dimensionality reduction methods include Principal Component Analysis (PCA) [1] and Linear Discriminant Analysis (LDA) [2,3]. PCA provides an optimal linear transformation from the original data space to an orthogonal eigenspace with reduced dimensionality in the sense of least mean squared reconstruction error. Thus, PCA has little to do with discriminative features optimal for classification. Unlike PCA which is unsupervised, LDA is a supervised technique that prior class information is used to guide the learning procedure. The purpose of LDA is to find the projection axes on which the data points have maximum between-class separation and minimum within-class dispersion. It is widely believed that supervised learning algorithms generally outperform unsupervised learning techniques because of the utilization of label information. However, it is usually difficult to collect sufficient labeled data because labeling often requires expensive human labor and much time. In contrast, unlabeled data are relatively much easier to obtain. So learning with both labeled and unlabeled data, called semi-supervised learning is becoming an important area in the machine learning community.

There is an increasing interest in graph-based semi-supervised learning methods [4–12] recently, which not only consider the label information, but also utilize a consistency assumption, namely, nearby points are likely to have the same label in classification tasks. In general, existing graph-based semi-supervised learning methods can be divided into transductive learning (e.g. [4–6]) and inductive learning (e.g. [7–13]). With the purpose of estimating the labels of the given unlabeled data, transductive learners only work on the labeled and unlabeled training data, and cannot deal with unseen (test) data, whereas inductive learners try to induce a decision function that has a low classification error rate on the whole sample space, thus can naturally handle unseen data. In this paper, we focus on inductive learning in the context of classification. Cai et al. [7] proposed semi-supervised discriminant analysis by introducing a smoothness constraint into the objective function of LDA. This constraint based on graph Laplacian regularization [14]

E-mail addresses: vv706@sina.com (W. Li), qqruan@center.njtu.edu.cn (Q. Ruan), 09112088@bjtu.edu.cn (J. Wan).

http://dx.doi.org/10.1016/j.neucom.2014.02.005
0925-2312 © 2014 Published by Elsevier B.V.

aims to keep nearby points to have similar representations in the low-dimensional feature space. Nie et al. [8] imposed an orthogonal constraint and formulated a semi-supervised orthogonal discriminant analysis method for dimensionality reduction. Zhang et al. [9] employed domain knowledge which specifies whether a pair of instances belongs to the same class (must-link constraints) or different classes (cannot-link constraints) to discover the underlying structure of data set. Cevikalp et al. [10] utilized the information provided by must-link and cannot-link to modify locality preserving projections (LPP) [15] scheme such that it can better discover groups within data. Xu et al. [11] proposed an adaptive regularization method to better depict the interplay between the labeled and unlabeled data. Zhang et al. [12] first estimated the class labeled of all the unlabeled data points by solving a constrained optimization problem. Then the unlabeled data with sufficiently high confidence are selected to maximize an optimality criterion of LDA. In addition to these methods, a general semi-supervised dimensionality reduction framework was proposed in [13], which takes advantage of labeled data to characterize class structure while unlabeled data are utilized to capture the intrinsic local geometry. Many classical dimensionality reduction methods can be unified into this framework as claimed in [13]. Semi-supervised LDA (SSLDA) and semi-supervised maximum margin criterion (SSMMC) [13] are two outcomes by generalizing LDA and maximum margin criterion (MMC) [16] to this framework respectively.

Most of the aforementioned algorithms cannot promise good performance when the size of labeled data set is small. This is resulted from inaccurate class matrix variance estimated by insufficient labeled training data. Moreover, the class information of large amount of unlabeled data is ignored in decision making, which may be helpful to uncover class distribution. Probabilistic semi-supervised discriminant analysis (PSDA) algorithm is a newly proposed method presented by Li et al. [17], which utilizes unlabeled samples to approximate class structure instead of local geometry. Based on the basic assumption that points lying on the same structure are likely to have the same label, each unlabeled sample is quantified class membership degree according to corresponding class reconstruction error. Then unlabeled data points can be combined with labeled ones to approximate class distribution. This property can compensate for the drawbacks induced by limited size of labeled data set and further improves the discriminant ability. However, PSDA is suboptimal for classification due to that it mainly focuses on the discriminative information hidden in unlabeled data points and ignores the geometric information which is critical to reveal the intrinsic distribution of data points.

In this paper, following the intuition that naturally occurring face may be sampled from the data with a probability distribution on a submanifold of ambient space, a graph-based extension of PSDA, referred to as graph-embedded probability-based semi-supervised discriminant analysis (GPSDA), is developed for face and facial expression recognition. The basic idea of GPSDA is not only explore the category information of unlabeled data but also preserve the intra-class and inter-class local information in the processing of dimensionality reduction. To this end, we first introduce a similarity measurement to investigate the inexact class information of unlabeled data. With this characterization, a graph is built with the guide of category information, which is more powerful to discover the intrinsic structure of data set compared with the graph modeled based merely on local structure in traditional semi-supervised dimensionality reduction technique. Finally, we use manifold regularizer for constraining the consistency similar to [7]. With the complementary geometric information learnt from unlabeled data, the decision boundary constructed by GPSDA precisely match what we expect, bring superior classification results when compared with prior work [10–13,17].

The remainder of the paper is organized as follows. Section 2 outlines PSDA. Section 3 details our proposed graph-embedded probability-based semi-supervised discriminant analysis method. The experimental results and discussions are presented in Section 4. We conclude the paper in Section 5.

2. Probabilistic semi-supervised discriminant analysis

Different from traditional semi-supervised learning methods [7,13] which use unlabeled samples to depict manifold structure, PSDA utilizes unlabeled data points to characterize class distribution. This property contributes to make PSDA more powerful and more robust when the amount of labeled data set is insufficient. Based on the key assumption that points lying on the same structure are likely to have the same label, PSDA first quantifies class membership probability to each unlabeled sample by using a well-designed soft assignment technique. Then discriminant analysis is carried out over both labeled and unlabeled data which bears an analogy to the Fisher criterion.

Assume a partially labeled data set $X=\{x_i\}_{i=1}^{N}$ with N samples, where $x_i \in R^D$ represents a vectorized image. The first l points $X_L=[x_1,\ldots,x_l]$ have labels $\{y_1,\ldots,y_l\} \in \{1,\ldots,C\}$ and the remaining u points $X_U=[x_{l+1},\ldots,x_N]$ are unlabeled. The kth class has n_k samples, $\sum_{k=1}^{C} n_k = l$. The global centroid of the whole data is defined by: $\mu=(1/N)\sum_{j=1}^{N} x_j$. Suppose the data points in X_L are given as $\{X_1,\ldots,X_C\}$, where $X_i (1 \le i \le C)$ describes the data matrix of the ith class containing observed vectors in its column. Samples in X_i spread as a low-dimensional linear subspace which is represented by a basis matrix $B_i \in R^{D\times d}$ $(d \ll D)$ obtained by retaining the first d columns of the orthogonal basis matrix $B_i^{ori} \in \Re^{D\times D}$ of the ith class, s.t. $(X_i-\mu_i e_{n_i}^T)(X_i-\mu_i e_{n_i}^T)^T = B_i^{ori}\Lambda_i (B_i^{ori})^T$, where μ_i is the centroid of the ith class, $\mu_i=(1/n_i)\sum_{x_j \in C_i} x_j$, $e_{n_i} = [1,1,\ldots,1]^T \in R^{n_i}$.

The crux of PSDA algorithm is to estimate the label information of unlabeled data. Rather than just assigning the binary is or is not label to an unlabeled instance, a soft assignment measured by class reconstruction error is designed. Given an unlabeled point x_k, its d-dimensional projection vector onto the subspace of class i is $z_k = B_i^T(x_k-\mu_i)$. The reconstructed vector of x_k is represented as $\overline{x_k} = B_i B_i^T (x_k-\mu_i)+\mu_i$ in original space R^D. The error between x_k and $\overline{x_k}$ denotes as $\varepsilon_k^{(i)} = x_k - \overline{x_k}$. The reconstruction error is utilized to measure the class membership degree of x_k to the ith class:

$$w_{ki} = g(\varepsilon_k^{(i)}) \tag{1}$$

where $g(x)$ is generally a monotonically decreasing function because the classes that yield smaller reconstruction error of a given unlabeled sample are likely to have the same label with the sample and should be given a greater value. In [17], Gaussian kernel function is applied $g(x)=\exp(-(x^2/2\sigma^2))$. Thus, a size of $u \times C$ matrix W is obtained whose rows specify category information of unlabeled data. The normalized class membership degree $w_{ki} = (w_{ki}/\sum_{i=1}^{C} w_{ki})$can serve as the probability of x_k belonging to class i.

After obtaining the probabilistic distribution matrix W, the weight matrix $S \in \Re^{N\times C}$ which summarizes the class information of the whole input data is built:

$$S(i,j) = \begin{cases} 1 & \text{if } x_i \in X_L \text{ and } x_i \in C_j \\ w_{ij} & \text{if } x_i \in X_U \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Obviously, each row sum of S equals to 1, that is $\sum_{i=1}^{C} s_{ji} = 1, j = 1,2,\ldots,N$.

Considering both labeled and unlabeled data, the global centroid and class centroid are updated by the following formulas:

$$\overline{\mu} = \frac{\sum_{i=1}^{C}\sum_{j=1}^{N} s_{ji}x_j}{\sum_{i=1}^{C}\sum_{j=1}^{N} s_{ji}} = \frac{\sum_{j=1}^{N}\sum_{i=1}^{C} s_{ji}x_j}{\sum_{j=1}^{N}\sum_{i=1}^{C} s_{ji}} = \frac{\sum_{j=1}^{N} x_j}{N} = \mu \tag{3}$$

Download English Version:

https://daneshyari.com/en/article/406533

Download Persian Version:

https://daneshyari.com/article/406533

Daneshyari.com