Neurocomputing 152 (2015) 69–76

Contents lists available at ScienceDirect

Neurocomputing

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

A novel semi-supervised learning for face recognition

Quanxue Gao [a,*], Yunfang Huang [a], Xinbo Gao [a,b], Weiguo Shen [a], Hailin Zhang [a]

[a] State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, Shaanxi, 710071 China
[b] School of Electronic Engineering, Xidian University, Xi'an, Shaanxi, China 710071 China

ARTICLE INFO

Article history:
Received 20 November 2013
Received in revised form
9 July 2014
Accepted 3 November 2014
Communicated by D. Tao
Available online 15 November 2014

Keywords:
Semi-supervised learning
Discriminant analysis
Dimension reduction
Diversity

ABSTRACT

Laplacian embedding (LE) has been widely used to learn the intrinsic structure of data. However, LE ignores the diversity and may impair the local topology of data, resulting in unstable and inexact intrinsic structure representation. In this article, we build an objective function to learn the intrinsic structure that well characterizes both the similarity and diversity of data, and then incorporate this structure representation into linear discriminant analysis to build a semi-supervised approach, called stable semi-supervised discriminant learning (SSDL). Experimental results on two databases demonstrate the effectiveness of our approach.

1. Introduction

Dimensionality reduction has been an active research topic in face recognition, human action recognition, image classification, and machine learning [1–4]. The goal of dimensionality reduction is to reveal the meaningful structures and unexpected relationships embedded in multivariate data for different tasks such as data representation and classification. For classification, many previous researches have demonstrated that the performance can be significantly improved by discriminant approaches [5–8]. Although their motivations are different, all these approaches can be unified in the graph embedding framework [9] or patch alignment framework [10], and the performance of these approaches will seriously deteriorate when the number of the labeled training samples is small or very small. In real-world applications, the labeled data are hard or expensive to obtain and meanwhile it is easy to acquire abundant unlabeled data at very low cost. Thus, how to efficiently exploit the large unlabeled data is becoming an interesting research topic for improving the performance of discriminant approaches [11–14].

Unsupervised learning especially unsupervised manifold learning has demonstrated that some intrinsic geometric structures embedded in the unlabeled data are of great importance for data or image classification [15–17]. Motivated by this, many semi-supervised learning approaches were developed by combining supervised and unsupervised learning [11,13,14,18–20]. These algorithms 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 [12,13,18]. For semi-supervised dimensionality reduction, graph-based semi-supervised manifold learning techniques are successful and effective in many applications such as face recognition, action recognition, and image retrieval [19–21]. These semi-supervised approaches can be unified within the graph-based semi-supervised framework [22, 23].

Semi-supervised discriminant analysis (SDA) [18] and maximum margin projection (MMP) [24] are two of the most representative graph-based semi-supervised manifold learning approaches. SDA imposes a smoothness constraint into the objective function of linear discriminant analysis (LDA). This constraint based on graph Laplacian regularization [16] aims to enforce nearby points to have similar representations in the low-dimensional feature space. Different from SDA, MMP adds the same constraint into the local discriminant objective function. Motivated by SDA and MMP, some graph embedding based semi-supervised approaches were developed by different similarity metrics [25] or label propagation for the unlabeled data [26–28]. Although the motivations of the above- mentioned semi-supervised approaches are different, all these semi-supervised approaches employ Laplacian embedding (LE) to learn the intrinsic structure of data or the smoothness constraint.

It is generally considered that LE has the local topology preserving property. However, LE emphasizes the large distance pairs and may enforce that the larger the distance between two nearby points is, the closer they are embedded in the reduced space, resulting in the impairment of the local topology of data [29–31]. This impairs the intrinsic structure embedded in data. Moreover, LE maps nearby data in the observed data space to nearby points with the low- dimensional representation, and

* Corresponding author. Tel.: +86 29 88204753; fax: +86 29 88204753.
E-mail address: gaoqxnpu@163.com (Q. Gao).

http://dx.doi.org/10.1016/j.neucom.2014.11.018
0925-2312/

mainly characterizes the similarity of data. In the ideal case, all nearby points may be mapped to a single point in the reduced space. This leads to over- fitting and impairs the stableness of the algorithms.

Recently, some works have showed that the diversity of data, which can be obtained by maximizing the variation of data, also reflects the intrinsic structure of data and is of great importance for data representation and classification [32–35]. Moreover, in many real-world applications, the intrinsic structure of data is unknown and complex, and testing data is usually different from the training data due to many factors. Thus, only similarity or diversity may not be sufficient to characterize the intrinsic structure of data [31]. Motivated by the analysis of great insight, we propose a novel semi-supervised dimensionality reduction algorithm called stable semi- supervised discriminant learning (SSDL), which explicitly takes into account both the diversity and similarity among nearby data points in this article. To be specific, we construct an adjacency graph, which characterizes the diversity of data, to overcome the shortcoming of LE, and combine it with LE to learn the intrinsic structure, and then incorporate the intrinsic structure into the objective function of LDA. Experiments on two databases demonstrate the effectiveness of our approach.

2. Semi-supervised discriminant analysis (SDA)

SDA respects the intrinsic geometrical structure inferred from both labeled and unlabeled data points and integrates the intrinsic geometrical structure of data into the objective function of LDA. Given a labeled set $\{(\mathbf{x}_i,\tau_i)\}i=1^l$ belonging to $\boldsymbol{c}$ classes and an unlabeled set $\{\mathbf{x}_i\}i=l+1^n$, where $\mathbf{x}_i\in R^d$, τ_i denotes the label of training data $\mathbf{x}_i$. Suppose that the k-th class has l_k samples, then $\sum_{k=1}^{c} l_k = l$. The objective function of SDA is as follows [18]:

$$\underset{\boldsymbol{\alpha}^T\boldsymbol{\alpha}=1}{\arg\max}\ \frac{\boldsymbol{\alpha}^T\mathbf{S}_b\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\left(\mathbf{S}_w+\beta\sum_{i,j=1}^{n}\left((\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^T S_{ij}\right)\right)\boldsymbol{\alpha}} \tag{1}$$

$$\mathbf{S}_b=\frac{1}{n}\sum_{k=1}^{c} l_k\left(\mathbf{m}^{(k)}-\mathbf{m}\right)\left(\mathbf{m}^{(k)}-\mathbf{m}\right)^T \tag{2}$$

$$\mathbf{S}_w=\frac{1}{n}\sum_{k=1}^{c}\left(\sum_{i=1}^{l_k}\left(\mathbf{x}_i^{(k)}-\mathbf{m}^{(k)}\right)\left(\mathbf{x}_i^{(k)}-\mathbf{m}^{(k)}\right)^T\right) \tag{3}$$

where $\boldsymbol{m}$ is the total sample mean vector, $\mathbf{m}^{(k)}$ is the average vector of the k-th class. $\mathbf{S}_w$ and $\mathbf{S}_b$ are the within-class scatter matrix and between-class scatter matrix, respectively. $\mathbf{x}_i^{(k)}$ is the i-th sample in the k-th class. The elements S_{ij} in weight matrix $\boldsymbol{S}$ are defined as follows:

$$S_{ij}=\begin{cases}1, & \text{if } \mathbf{x}_i\in N_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j\in N_k(\mathbf{x}_i)\\ 0, & \text{Otherwise}\end{cases}$$

where $N_k(\mathbf{x}_i)$ denotes the set of k nearest neighbors of $\boldsymbol{x}_i$.

The denominator of the objective function (1) can be decomposed into the following two objective functions:

$$\underset{\boldsymbol{\alpha}^T\boldsymbol{\alpha}=1}{\arg\min}\ \sum_{k=1}^{C}\left(\sum_{i=1}^{l_k}\boldsymbol{\alpha}^T\left(x_i^{(k)}-\mathbf{m}^{(k)}\right)\left(x_i^{(k)}-\mathbf{m}^{(k)}\right)^T\boldsymbol{\alpha}\right) \tag{4}$$

$$\underset{\alpha^T\alpha=1}{\arg\min}\ \sum_{i,j=1}^{n}\left(\alpha^T(x_i-x_j)(x_i-x_j)^T S_{ij}\alpha\right) \tag{5}$$

The objective function (4) represents the within-class compactness and characterizes the similarity of data points having the same class label. The objective function (5) aims to map nearby data points in the observed data space to nearby data points with the low-dimensional representation. However, the objective functions (4) and (5) result in the following problems:

- They may lead to unstable intra-class compact representation. Eq. (4) aims to map the within-class data points to be close to each other in the embedding space. In the ideal case, the data points having the same class label are mapped to a single point in the reduced space. It means that Eq. (4) mainly characterizes the common geometrical properties, i.e. similarity of data, and ignores the different geometrical properties, i.e. diversity of data. Moreover, Eq. (5), i.e. LE still characterizes the similarity of data and does not overcome the shortcoming caused by the objective function (4). As previously discussed, only similarity of data may not be sufficient to characterize the intrinsic structure of data. Thus, Eqs. (5) and (4) do not sufficient to represent the within-class compactness of data.
- The objective function (5) may impair the local topology. Roughly speaking, local topology preserving means that: for any pair of points x_i and x_j, the smaller the distance between them is, the closer they should be embedded together in the embedding space [28]. In real-world applications, the data distribution is uneven, thus some nearby data points may lie on a sparse region, and the distance among these data points is large. In this case, these data points with large distance will dominate the objective function (5). Thus, Eq. (5) does not guarantee that the smaller the distance between nearby data points is, the closer they should be embedded in the reduced space [29–31,33]. Taking the unlabeled data points, which are randomly selected, in Fig. 1 (a) as an example, we show the projection directions of Eq. (5) in Fig. 1 (a) and the corresponding one-dimensional embedded results in Fig. 1 (b). It is easy to see that Eq. (5) leads to the violation of local topology preserving at small distance data pairs in circle.

3. Stable semi-supervised discriminant learning

3.1. Intrinsic structure representation

We construct two adjacency graphs $G_{g-s}=\{\mathbf{X},\mathbf{B}\}$ and $G_{g-v}=\{\mathbf{X},\mathbf{D}\}$ with a vertex set $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n\}$ and two weight matrices $\mathbf{B}$ and $\mathbf{D}$ to model the intrinsic structure of data. Where matrix $\boldsymbol{B}$ characterizes the similarity relationship among nearby points, and matrix $\boldsymbol{D}$ characterizes the diversity relationship among nearby points. Motivated by [16, 30, 31], we define the elements B_{ij} in $\boldsymbol{B}$ and D_{ij} in $\boldsymbol{D}$ as follows:

$$B_{ij}=\begin{cases}\exp\left(-d^2(\mathbf{x}_i,\mathbf{x}_j)/t^2\right), & \text{if } \mathbf{x}_i\in N_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j\in N_k(\mathbf{x}_i)\\ 0, & \text{Otherwise}\end{cases}$$

$$D_{ij}=\begin{cases}1-\exp\left(-d^2(\mathbf{x}_i,\mathbf{x}_j)/t^2\right), & \text{if } \mathbf{x}_i\in N_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j\in N_k(\mathbf{x}_i)\\ 0, & \text{otherwise}\end{cases}$$

where $d(\mathbf{x}_i,\mathbf{x}_j)=\|\mathbf{x}_i-\mathbf{x}_j\|$ denotes the Euclidean distance between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$.

Now consider the problem of mapping the data points to a line so that the following two conditions can be satisfied. First, nearby points in the observed data space are mapped to be close to each other. Second, nearby points, which lie on a sparse region in the high-dimensional space, are not mapped to be very close in the reduced space. Thus, the local topology, which characterizes both the similarity and diversity of data, can be well preserved. As it happens, a good map is to optimize the following two objective functions.

$$\underset{y}{\arg\min}\ \sum_{ij}\left(y_i-y_j\right)^2 B_{ij} \tag{6}$$