Neurocomputing 188 (2016) 275–283

Contents lists available at ScienceDirect

Neurocomputing

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

A non-negative sparse semi-supervised dimensionality reduction algorithm for hyperspectral data

Xuesong Wang, Yang Gao, Yuhu Cheng *

School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou, Jiangsu 221116, PR China

ARTICLE INFO

Article history:
Received 18 September 2014
Received in revised form
17 November 2014
Accepted 6 December 2014
Available online 4 December 2015

Keywords:
Hyperspectral data
Non-negative sparse representation
Semi-supervised learning
Dimensionality reduction
Discriminant term
Regularization term

ABSTRACT

A non-negative sparse semi-supervised dimensionality reduction algorithm is proposed for hyperspectral data by making adequate use of a few labeled samples and a large number of unlabeled samples. The objective function of the proposed algorithm consists of two terms: (1) a discriminant item is designed to analyze a few labeled samples from the global viewpoint, which can assess the separability between surface objects; (2) a regularization term is used to build a non-negative sparse representation graph based on the unlabeled samples, which can adaptively find an adjacency graph for each sample and then find valuable samples with huge information volume from the original hyperspectral data. Based on the objective function and the maximum margin criterion, a dimensionality reduction algorithm, the non-negative sparse semi-supervised maximum margin algorithm, is proposed. Experimental results on the ROSIS University and AVIRIS 92AV3C hyperspectral data sets show that the proposed algorithm can effectively utilize the unlabeled samples to achieve higher overall classification accuracy and Kappa coefficient when compared with some representative supervised, unsupervised and semi-supervised dimensionality reduction algorithms.

1. Introduction

Despite the fact of being a fast growing research area, hyperspectral data analysis still faces many challenges, including the large number of measured wavelength bands, huge data size, data uncertainty, small samples, curse of dimensionality, and high data redundancy due to strong correlations between bands [1,2]. Such concerns directly affect the accuracy and speed of hyperspectral data classification. As a result, it is difficult to apply traditional classification methods directly for hyperspectral data classification. Therefore, to balance between efficiency, accuracy and adaptability, dimensionality reduction is generally required for classification of hyperspectral data by band selection [3] or band extraction [4,5]. Band selection is to select, directly from the original band space, some bands for subsequent data processing. In recent years, several evolutionary computation methods have been successively applied for dimensionality reduction of hyperspectral data [6]. However, since band selection is notably affected by the specific search method and decision criterion and naturally leads to a huge loss of information regardless of the methods. Band extraction is generally preferred for dimensionality reduction of hyperspectral data.

With efficient band extraction methods, the original hyperspectral data is mapped or transferred into a low-dimensional subspace (while it still retains certain necessary features of the original data) to address the curse-of-dimensionality concern so that subsequent processing tasks such as classification and clustering can be performed more computationally efficient and accurate [7]. Depending on the availability of label information of the samples, dimensionality reduction (DR) algorithms for hyperspectral data can be divided into different categories. In supervised learning, to obtain high classification accuracy, DR algorithms look for relevant bands with important contributions to classification by only using labeled samples [8]. However, in many practical applications, only a few labeled samples are available, making it extremely difficult to discriminate and eliminate redundant, unrelated bands from the original high-dimensional data. With the advanced development of data acquisition technologies, it is easy to acquire a large number of unlabeled samples, but obtaining manually-labeled samples is time-consuming and expensive [9]. Under such situations, unsupervised DR algorithms can be employed to extract effective bands by exploring hidden structures in the brand features of unlabeled samples, without requiring any prior label knowledge [10,11]. As a practical trade-off between supervised and unsupervised dimensionality reduction, some

* Corresponding author.
E-mail addresses: wangxuesongcumt@163.com (X. Wang), chengyuhu@163.com (Y. Cheng).

http://dx.doi.org/10.1016/j.neucom.2014.12.127
0925-2312/

researchers proposed semi-supervised DR algorithms that can simultaneously process both labeled and unlabeled samples [12].

In recent years, semi-supervised DR algorithm has been widely applied in face recognition, text classification and so on. For instance, Cai, et al. [13] extended the linear discriminant analysis (LDA) [14] algorithm, proposed the semi-supervised discriminant analysis (SDA) and successfully applied it to face recognition with 1-labeled sample. A popular design concept to extend unsupervised and supervised DR algorithms is to use the manifold regularization method, e.g., semi-supervised maximum margin criterion (SSMMC) and semi-supervised LDA (SSLDA) proposed by Song et al. [15] based on MMC and LDA respectively [16], subspace semi-supervised locality preserving (SSLPP) and subspace semi-supervised boundary Fisher analysis proposed by Yang et al. [17]. Such semi-supervised DR algorithms can all be interpreted within one general framework, i.e., using the Graph Laplacian method as the regularization term [17]. Such graph-based semi-supervised DR algorithms have been successfully applied to areas such as face recognition, but there are still several associated concerns: (1) They all assume the basis of manifold structure, meaning it is necessary to have adequate samples for representing the distribution of samples [18]. (2) As Zhu [19] pointed out, most of such algorithms as SDA and SSLPP build adjacency graphs using the k nearest neighborhood (KNN) method, but the nearest neighborhood criterion generally cannot acquire adequate discriminant information. (3) Parameters in KNN or sample-dependence neighborhood need to be heuristically pre-determined and all sampling points use the neighborhood number of a fixed sized or fixed neighborhood range. The distribution of real samples is basically ignored, and parameter selection remains a challenge. To address such concerns, Qiao et al. [20] built the adjacency graph using sparsity-preserving projection instead of the conventional Graph Laplacian method and proposed the sparsity-preserving discriminant analysis (SPDA). Wright et al. [21] showed that the sparse representation classification criterion is superior to the nearest neighborhood criterion, especially in processing high-dimensional data. However, [21] provides sparse representation only for the regularization term in the semi-supervised DR (i.e., building the sparse adjacency graph only for unlabeled samples), but ignores the discriminant term (i.e., sparse structure of labeled samples), while the intra-class scatter matrix and inter-class scatter matrix constructed by applying LDA directly are not real sparse representation algorithms. Gui et al. [22] proposed the discriminant sparse neighborhood preserving embedding (DSNPE) and showed that the classification performance of the sparse representation algorithm for labeled samples is much better than that of LDA. These graph-based algorithms may also have a negative weight in the process of graph-building. Wong proposed non-negative sparseness preserving embedding (NSPE) [23], pointing out that a negative weight fails to rationally carry out the information transfer between samples.

By jointly addressing the above sparse representation and non-negative weight concerns, in this paper, we propose a DR algorithm with reference to MMC, a non-negative sparse semi-supervised maximum margin criterion (NS^{3}MMC) algorithm, and we apply it for hyperspectral data. The objective function of the proposed NS^{3}MMC consists of two terms: a discriminant term and a regularization term. By making adequate use of a few labeled samples, the discriminant term constructs a separability model between different surface objects. Using the block non-negative sparse representation, the regularization term finds the samples with the discriminative capacity from the unlabeled samples and effectively improves the DR effect of the algorithm when combined with the separability model of the discriminant term.

In the remainder of this paper, Section 2 describes the proposed non-negative sparse semi-supervised DR algorithm. Section 3 demonstrates the performances of the proposed algorithm on two real hyperspectral data sets. Section 4 concludes the paper.

2. Non-negative sparse semi-supervised dimensionality reduction algorithm

2.1. Problem formulation

In the semi-supervised problem, we have the high-dimensional hyperspectral data set with unknown distribution, $O_m = [X, Y]$, where only some samples are labeled. We assume that the first l samples are labeled, while the remaining u samples are unlabeled, i.e., the high-dimensional sample set is $X = (X_L, X_U)$, where $X_L = (x_1, x_2, \cdots, x_l)$, $X_U = (x_{l+1}, x_{l+2}, \cdots, x_{l+u})$, and $\mathbf{x}_i \in R^m$ is a m-dimensional vector. The label set is $Y = Y_L = (y_1, y_2, \cdots, y_l)^T$ where $y_i = \{1, \cdots, c, \cdots, C\}$. The semi-supervised DR algorithm aims to find one projection matrix $W = (w_1, w_2, \cdots, w_d) \in R^{m \times d} (d << m)$, by using both labeled and unlabeled samples, so that

$$z = W^T x \in R^d \tag{1}$$

where z is a low-dimensional representation of the original high-dimensional hyperspectral data x.

2.2. Discriminant term

Unlike SDA based on the LDA discriminant term and semi-supervised DR algorithms based on the pairwise constrains discriminant term [24], the discriminant term here is based on non-negative sparse representation. The purpose of non-negative sparse representation is to use the minimum number of elements (atoms) in the over-complete dictionary $\boldsymbol{X}_{\mathrm{L}} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_l] \in R^{m \times l}, m < l$ to represent hyperspectral data $\boldsymbol{x}_i$:

$$\begin{aligned} &\min_{\boldsymbol{h}_i} \quad \|\boldsymbol{h}_i\|_0 \\ &\text{s.t. } \boldsymbol{x}_i = \boldsymbol{X}_{\mathrm{L}} \boldsymbol{h}_i;\ \boldsymbol{h}_i \geq 0 \end{aligned} \tag{2}$$

where $\|\boldsymbol{h}_i\|_0$ represents the l_0-norm of $\boldsymbol{h}_i$ (i.e., the number of non-zero elements in $\boldsymbol{h}_i$), with $\boldsymbol{h}_i = \left[h_{i1}, \cdots, h_{ii-1}, 0, h_{ii+1}, \cdots, h_{il}\right]^{\mathrm{T}} \in R^l$, and h_{ij} represents the contribution of the jth sample $\boldsymbol{x}_j$ to reconstruct the sample $\boldsymbol{x}_i$. Because $\boldsymbol{x}_i$ should be eliminated from $\boldsymbol{X}_{\mathrm{L}}$, the ith element in $\boldsymbol{h}_i$ is set to be 0 (i.e., $\boldsymbol{x}_i = h_{i1}\boldsymbol{x}_1 + \cdots + h_{ii-1}\boldsymbol{x}_{i-1} + h_{ii+1}\boldsymbol{x}_{i+1} + \cdots + h_{il}\boldsymbol{x}_l$).

Eq. (2) is a NP-hard non-convex combinatorial optimization problem. Under the condition that the solutions are adequately sparse, the solution of the l_0 minimization problem is approximate or equal to the solution of the l_1 minimization problem [25]. The l_1 minimization problem can be solved efficiently through LASSO [26] or Elastic Net [27].

$$\begin{aligned} &\min_{\boldsymbol{h}_i} \quad \|\boldsymbol{h}_i\|_1 \\ &\text{s.t. } \boldsymbol{x}_i = \boldsymbol{X}_{\mathrm{L}}^c \boldsymbol{h}_i;\ 1 = 1^{\mathrm{T}} \cdot \boldsymbol{h}_i;\ \boldsymbol{h}_i \geq 0;\ y_i = c \end{aligned} \tag{3}$$

where $1 \in R^{l_c}$ is the all-1 vector and l_c is the number of samples labeled as class c. Non-negative GLS (Generalized Least Squares) is used to solve Eq. (3) for obtaining the reconstruction error [27]:

$$\begin{aligned} &\min_{\boldsymbol{h}_i} E(\boldsymbol{H}^c) = \min_{\boldsymbol{h}_i} \|\boldsymbol{X}_{\mathrm{L}}^c - \boldsymbol{X}_{\mathrm{L}}^c \boldsymbol{H}^c\|_2^2 + \gamma \|\boldsymbol{h}_i\|_1 \\ &\text{s.t. } \boldsymbol{h}_i \geq 0;\ 1 = 1^{\mathrm{T}} \boldsymbol{h}_i;\ y_i = c \end{aligned} \tag{4}$$

where γ is the non-negative coefficient for balancing sparseness and reconstruction error. According to Eq. (4), we can compute the optimal non-negative sparse reconstruction weight vector $\tilde{\mathbf{h}}_i \in R^{l_c}$ for each hyperspectral data $\mathbf{x}_i$i n class c, and thenthediscriminant non-negative sparse reconstruction weight matrix for class cis$\mathbf{H}^c = \left[\tilde{\mathbf{h}}_i\right]_{l_c \times l_c}$. Thus, the discriminant non-negative sparse reconstruction weight matrix $\mathbf{H}_{\mathrm{D}}$ of the hyper spectral data can be