Neurocomputing 220 (2017) 130–137

Contents lists available at ScienceDirect

Neurocomputing

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

Graph self-representation method for unsupervised feature selection

Rongyao Hu [a,1], Xiaofeng Zhu [a,*], Debo Cheng [a,1], Wei He [a], Yan Yan [b], Jingkuan Song [b], Shichao Zhang [a,*]

[a] Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin, Guangxi 541004, China
[b] University of Trento, Italy

ARTICLE INFO

Article history:
Received 15 January 2016
Received in revised form
28 March 2016
Accepted 15 May 2016
Available online 24 June 2016

Keywords:
Dimensionality reduction
Alternating direction method multipliers
Locality preserving projection
Self-representation

ABSTRACT

Both subspace learning methods and feature selection methods are often used for removing irrelative features from high-dimensional data. Studies have shown that feature selection methods have interpretation ability and subspace learning methods output stable performance. This paper proposes a new unsupervised feature selection by integrating a subspace learning method (i.e., Locality Preserving Projection (LPP)) into a new feature selection method (i.e., a sparse feature-level self-representation method), aim at simultaneously receiving stable performance and interpretation ability. Different from traditional sample-level self-representation where each sample is represented by all samples and has been popularly used in machine learning and computer vision. In this paper, we propose to represent each feature by its relevant features to conduct feature selection via devising a feature-level self-representation loss function plus an $\ell_{2,1}$-norm regularization term. Then we add a graph regularization term (i.e., LPP) into the resulting feature selection model to simultaneously conduct feature selection and subspace learning. The rationale of the LPP regularization term is that LPP preserves the original distribution of data after removing irrelative features. Finally, we conducted experiments on UCI data sets and other real data sets and the experimental results showed that the proposed approach outperformed all comparison algorithms.

1. Introduction

In data mining and machine learning, high-dimensional data are often extracted to describe the diversity of data, the process readily leads to the issue of curse of dimensionality [1,2]. However, the high-dimensional data usually have a low-dimensional structure. Thus a lot of efforts have designed dimensionality reduction methods (including feature selection methods and subspace learning methods) to search for such low-dimensional structure by reducing the dimensions of high-dimensional data [3–5].

Dimensionality reduction methods are usually divided into feature selection methods [6,7] and subspace learning methods [8,9]. Feature selection methods are widely used for reducing the dimensions of data to output a subset of features [10,11]. That is, feature selection methods select a subset of features in accordance with criteria, such as distinguishing features with good characteristics and correlating to the predefined goal. The state-of-the-art feature selection methods include filter methods [12–14], wrapper methods [15,16] and embedded methods [17–20]. Feature selection methods directly removing a subset of features lead to interpretation but may lose information, so the performance of feature selection methods are unstable. Fortunately, subspace learning methods [9,21,22] have been designed to map all features into a low-dimensional space and also remove the noise and outliers inhere in data. In this way, subspace learning methods can achieve stable performance to deal with high-dimensional data. For example, Zhu et al. proposed to first conduct subspace learning to convert original data into low-dimensional Hamming subspace, and then to consider the correlations between the original space and the group effect of the features in training data [22]. Motivated by the above observation, one can integrate subspace learning methods into the framework of feature selection, to yield both the interpretation ability and stable performance.

In this paper, we propose a new unsupervised Graph Self-Representation Sparse Feature Selection (shorted for GSR_SFS) method, to address the above limitation. We first propose to represent each feature by its relevant features, which includes a feature-level self-representation loss function and an $\ell_{2,1}$-norm regularization term in a sparse way. The rationale of the feature-level self-representation loss function is that the more important the feature is, the more it has to the chance that represents other features jointly. Similarly, the unimportant features should not

[*] Corresponding authors.
E-mail addresses: hu_No1@126.com (R. Hu), seanzhuxf@gmail.com (X. Zhu), 15676209686@163.com (D. Cheng), risehnhew@163.com (W. He), yan.yan@unitn.it (Y. Yan), jingkuan.song1@unitn.it (J. Song), zhangsc@mailbox.gxnu.edu.cn, hhwwboy@163.com (S. Zhang).
[1] These authors have contributed equally to this work.

http://dx.doi.org/10.1016/j.neucom.2016.05.081
0925-2312/© 2016 Elsevier B.V. All rights reserved.

participate to represent other features. The $\ell_{2,1}$-norm regularization term penalizes all coefficients in the same row of the regression matrix together for joint selection or un-selection in predicting the response variables. Then, we add a graph regularization term into the resulting feature selection model to conduct subspace learning. The goal of the graph regularization term is to improve the stability of our feature selection model by preserving the local structures of data in the low-dimensional space. Furthermore, we propose an efficient Alternating Direction Method Multipliers (ADMM) method [23] to solve the proposed objective function. The proposed optimization method enables the GSR_SFS method to be used in the large-scale data sets.

The contributions of this paper are two-fold:

- The property of self-representation is not a new concept and has been popularly used in machine learning and computer vision, such as sparse coding [22,24] and low-rank [25,26]. However, previous literatures [27,28] focused on the sample-level self-representation where each sample is represented by all samples. In this paper, we propose to represent each feature by its relevant features to conduct feature selection.
- This paper combines subspace learning with feature selection with the goal of outputting a stable feature selection model and also interpreted ability. To embed two different conceptual topics (i.e., subspace learning and feature selection) into a unified framework is very challenging in data mining and machine learning. To address this, this paper embeds a subspace learning based regularizer into a new devised feature selection model, where the LPP enables the data distribution to be preserved after removing the features. This naturally leads to the improvement of the stability of feature selection.

The rest of this paper is organized as follows: Section 2 summarizes recent studies of feature selection methods and Section 3 gives the detail of the proposed algorithm and its corresponding optimization algorithm. Section 4 presents our experimental results, followed by the conclusion of this paper in Section 5.

2. Related work

During the past decades, a large number of feature selection methods have been proposed to overcome the high-dimensional issue. Based on the availability of class labels, feature selection methods can be parted into unsupervised method, supervised method, and semi-supervised method.

In real application, it is usually expensive to obtain label information, so it makes unsupervised feature selection practical [29,30,19]. For example, the maximum variance method selected top ranked features with maximum variance [31]. However, the selected features by the maximum variance method can not guarantee to be discriminate for classification [32]. Then, the Laplacian score method was proposed to select features by preserving the local manifold structure of the data set [32], while the Multi-cluster feature selection (MCFS) first performed regression that using the eigenvector of graph Laplacian and then selected features with maximum spectral regression coefficients [33]. Recently, Bhadra and Bandyopadhyay proposed an unsupervised feature selection method by using an improved version of differential evolution [34].

Supervised methods usually select features according to the availability of the class labels and evaluate by known class labels. Interestingly, many supervised methods can be formulated into a similar framework, where a regularization term is appended to the loss function for feature selection [35]. For example, Wu et al. proposed to conduct feature selection by a least square loss function plus a group sparse regularizer [36], while Ma et al. proposed to conduct feature selection by a mixture loss function and two sparse regularizers [37]. Since there is a label information, supervised feature selection methods are usually able to output discriminative features. For example, Feng et al. [38] proposed a supervised feature selection method to make use of label information for constructing good discriminative framework of dictionary learning.

Semi-supervised methods are proposed to utilize both the limited number of labeled samples and a large number of unlabeled samples for training [39]. For example, Zhao et al. proposed to compute the importance score of each feature on the graph Laplacians of both labeled and unlabeled data for semi-supervised learning [40]. However, both of them were unable to select a large number of features simultaneously and ignored the interdependencies between features. To address these issues, Ma et al. proposed a semi-supervised framework for automatic image annotation [41]. Recently, Han et al. proposed a semi-supervised feature selection method with spline regression for video semantic recognition [42].

3. Method

In this section, we first give some notations used in this paper and describe the detail of the proposed GSR_FS method, in Sections 3.1 and 3.2, respectively, and then explain the proposed optimization method to the resulting objection in Section 3.3.

3.1. Notations

In this paper, matrices and vectors are written as boldface uppercase letters and boldface lowercase letters, respectively. For a matrix $\mathbf{X} = [x_{ij}]$, its i-th row and j-th column are denoted as $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. We also denote the Frobenius norm, $\ell_{2,1}$-norm of a matrix $\mathbf{X}$, respectively, as $\|\mathbf{X}\|_F = \sqrt{\sum_i \|\mathbf{x}_i\|_2^2} = \sqrt{\sum_j \|\mathbf{x}_j\|_2^2}$ and $\|\mathbf{X}\|_{2,1} = \sum_i \|\mathbf{x}_i\|_2 = \sum_i \sqrt{\sum_j x_{ij}^2}$. We further denote the transpose operator, the trace operator, and the inverse of a matrix $\mathbf{X}$ as $\mathbf{X}^T$, $tr(\mathbf{X})$, and $\mathbf{X}^{-1}$, respectively.

3.2. Graph self-representation sparse feature selection

Given a data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, where n and d are the number of samples and features, respectively, where $\mathbf{x}_j \in \mathbb{R}^n$ stands for a feature vector. Given a response matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_c] \in \mathbb{R}^{n\times c}$, we use the following formulation to construct a linear relationship between $\mathbf{X}$ and $\mathbf{Y}$:

$$\min_{\mathbf{Z}} l(\mathbf{Y} - \mathbf{XZ}) + \lambda\phi(\mathbf{Z}) \tag{1}$$

where $\mathbf{Z} \in \mathbb{R}^{d\times c}$ denotes the feature weight matrix, $\phi(\mathbf{Z})$ denotes the regularization imposing on $\mathbf{Z}$, $l(\mathbf{Y} - \mathbf{XZ})$ denotes the loss term, and λ denotes a positive constant. However, in real applications, labels are usually difficult to be obtained due to all kinds of reasons, such as budget limitation and unavailable labels, so the assumption of unsupervised feature selection does not have label information $\mathbf{Y}$, which implies that Eq. (1) does not make sense.

In this paper, we assume that the features are dependent and each feature can be represented by other features. Note that such assumption can be found in machine learning and computer vision. Specifically, we define a linear regression model such that each feature $\mathbf{x}_i$ in $\mathbf{X}$ can be represented as a linear combination of other features (including itself):

$$\mathbf{x}_i \approx \sum_{j=1}^{d} \mathbf{x}_j z_{ji}, \quad i, j = 1, \ldots, d. \tag{2}$$

where the element z_{ji} of matrix $\mathbf{Z}$ (where $\mathbf{Z} \in \mathbb{R}^{d\times d}$) is a weight