Neurocomputing xx (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Neurocomputing

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

Unsupervised feature selection for visual classification via feature-representation property

Wei He[1], Xiaofeng Zhu*, Debo Cheng[1], Rongyao Hu, Shichao Zhang*

Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin 541004, China

ARTICLE INFO

Keywords:
Feature selection
Self-representation
Sparse learning
Unsupervised learning

ABSTRACT

Feature selection is designed to select a subset of features for avoiding the issue of 'curse of dimensionality'. In this paper, we propose a new feature-level self-representation framework for unsupervised feature selection. Specifically, the proposed method first uses a feature-level self-representation loss function to sparsely represent each feature by other features, and then employs an $\ell_{2,p}$-norm regularization term to yield row-sparsity on the coefficient matrix for conducting feature selection. Experimental results on benchmark databases showed that the proposed method effectively selected the most relevant features than the state-of-the-art methods.

1. Introduction

High-dimensional data could lead to expensive computation cost as well as result in the issue of 'curse of dimensionality' so that affecting the performance of learning from the data [1–3]. In the past decades, dimensionality reduction (including feature selection and subspace learning) via reducing the dimensions has been becoming an efficient solution to high-dimensional data [4,5].

Feature selection directly removes a subset of features to output interpretable results, so that making it practical in real applications [6]. Previous feature selection methods can be classified into three categories, e.g., supervised feature selection, semi-supervised feature selection and unsupervised feature selection [1]. Supervised feature selection methods usually select features according to the labels of training data. For example, Gu et al. proposed to seek a subset of features by maximizing the lower bound of traditional Fisher score [4], while Zhang et al. proposed to use spectral-spatial feature combination for hyper spectral image analysis [7]. Since supervised feature selection methods enclose labels to conduct feature selection, they are able to select discriminative features.

Semi-supervised feature selection mainly utilizes a small number of labeled samples and a large number of unlabeled samples for the training stage [8]. For example, Lv et al. employed a manifold regularization term to conduct the discriminative semi-supervised feature selection [9]. Wang et al. proposed to first learn the class labels of unlabeled samples, and then to use the learned class labels to define the margins for feature weight learning [10].

However, due to all kinds of reasons such as unknown labels and time-consuming to obtain labels, it is difficult to obtain enough labels for learning from data, unsupervised feature selection thus is practical in alleviating irrelevant features [7,11]. Compared to either supervised feature selection method or semi-supervised feature selection method, unsupervised feature selection lacks the label information, so it is very challenging to conduct unsupervised feature selection [12]. Recently, unsupervised feature selection methods mainly utilized evaluation indicators to remove redundant features. For example, Liu et al. combined the Laplacian score with the distance-based entropy measure to conduct unsupervised feature selection [13], while Nie et al. proposed to use a corresponding score to conduct feature selection [14].

In this paper, we propose a new unsupervised feature selection method with the utilization of the property of feature self-representation, in which features can represent themselves to find representative feature ingredients. Motivated by the successful application of the self-similarity in subspace clustering [15,7,16], this paper first proposes a feature-level self-representation for unsupervised learning, and then adds an $\ell_{2,1}$-norm regularizer in the objective function to yield sparse feature selection. In our method, the proposed loss function is proposed to represent each feature by other features with the rationale of that the important features are usually used to represent other features and the unimportant features will be disused for all features. The group sparsity (i.e., 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. Besides, this paper also devises an novel and efficient optimization method to solve the resulting objective function as well as proves its convergence.

* Corresponding authors.
E-mail address: seanzhuxf@gmail.com (X. Zhu).
[1] Wei He and Debo Cheng have equally contributed to this work.

http://dx.doi.org/10.1016/j.neucom.2016.07.064
Received 25 February 2016; Received in revised form 8 July 2016; Accepted 14 July 2016
Available online xxxx
0925-2312/

It should be noted that the property of self-representation is not a new concept, which has been popularly used in machine learning and computer vision such as in the application of sparse coding [17] and low-rank [18]. However, previous literature [19,20] focused on the sample-level self-similarity where each sample is represented by all samples. In this paper, we propose to represent each feature by its relevant features. That is, we conduct feature selection via devising a feature-level self-representation loss function. The contribution of our method is described as follows:

- Unlike previous unsupervised feature selection methods mainly utilize a number of evaluation indicators to remove the redundant features, we propose a novel feature-level self-representation to remove the irrelevant features. The proposed feature-level self-representation is different from the sample-level self-similarity, which represent each sample by all samples.
- We propose a novel iterative optimization algorithm to solve the resulting objective function, which is also testified to efficiently converge to the optimum solution.

The left parts of this paper are organized as follows: Section 2 introduces related work on feature selection methods and Section 3 gives the details of our proposed feature selection model. In Sections 5 and 6, respectively, we show our experimental results and conclude our paper.

2. Related work

Dimensionality reduction methods are usually divided into two groups: feature selection methods [21] and subspace learning methods [6,22]. Feature selection methods are widely used for reducing the dimensions of high-dimensional data to output interpretable results [23,24]. 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 filer methods [25–27], wrapper methods [28,29] and embedded methods [30,31].

Filter methods choose features without involving any learning algorithm, so it can use feature evaluation indices to rank features or evaluate feature subsets [25] and its selection process does not depend on the consequent process. For example, Tabakhi et al. proposed to select a representative feature subset with an iterative algorithm [26], while Cao et al. proposed to simultaneously use the q-value of false discovery rate to measure the statistical significant and decrease the influence of redundant genes [32].

Wrapper methods evaluate the goodness of features via learning algorithms, so they generally have better performance than filer methods. For example, Unler et al. proposed to first use a swarm intelligence algorithm for feature selection and then take use of Support Vector Machine (SVM) for classification [28], while Chyzhyk et al. employed extreme learning machines as a learning algorithm and also comprised a genetic algorithm to explore the feature combination space [33]. Unfortunately, the computation cost of wrapper methods is more expensive than filter methods.

Embedded methods generally take feature selection as a part of the learning process by searching for relevant features with the objective function of learning models. For example, Wen et al. proposed a feature selection method with robust classification via $\ell_{2,1}$-norms regularization [30]. Shi et al. presented a sparse regression model to combine the embedded learning with sparse regression under a common framework [31]. Imani et al. focused on using the combination of the genetic algorithm with the feedback mechanism of ant colony optimization to conduct feature selection [34].

3. Approach

In this section, we first define the notations used in this paper, and then describe the details of the proposed method, followed by the proposed optimization method to the resulting objective function.

3.1. Notations

In this paper, matrices are deonted as boldface uppercase letters and vectors are written as boldface lowercases letters. The i-th row and j-th column of a data matrix $\mathbf{X}$ are denoted as $\mathbf{x}^i$ and $\mathbf{x}_j$, respectively. The Frobenius norm of a matrix $\mathbf{X}$ is defined as $\| \mathbf{X} \|_F = \sqrt{\sum_i \|\mathbf{x}_i\|_2^2}$, and the $\ell_{2,1}$-norm of $\mathbf{X}$ is denoted as $\| \mathbf{X} \|_{2,1} = \sum_i \| \mathbf{x}_i \|_2 = \sum_i \sqrt{\sum_j x_{ij}^2}$. 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. Least square regression

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

$$\min_{\mathbf{W}} g(\mathbf{W}) = f(\mathbf{W}) + \lambda \phi(\mathbf{W}). \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d \times c}$ denotes the feature weight matrix, $\phi(\mathbf{W})$ denotes the regularization imposed on $\mathbf{W}$, $f(\mathbf{W})$ denotes the loss term, and λ denotes a positive constant. Usually, in the literature, e.g., [35–37], $f(\mathbf{W})$ is defined as $\mathbf{Y} - \mathbf{XW}$, aim at achieving the minimum regression error between the labels $\mathbf{Y}$ and their prediction $\mathbf{XW}$. In this scenario, the least square loss function between labels and the features is formulated as:

$$\min_{\mathbf{W}} l(\mathbf{Y} - \mathbf{XW}) + \lambda \phi(\mathbf{W}) \tag{2}$$

Obviously, Eq. (2) considers the sample similarity among samples to conduct regression. However, it prohibits unsupervised feature selection without label information.

3.3. Representative features

In this paper, motivated by the successful use of self-similarity in machine learning [38] and computer vision [39], we reveal the feature-level relation among features to characterize the property that each feature can be linearly approximated by a subset of other features in unsupervised feature selection.

Self-similarity has been widely used in computer vision and machine learning [38,39]. In computer vision, non-local self-similarity means that patches at different locations in an image may be similar to each other. In machine learning, self-similarity can also be modeled as a sparse representation model or a low-rank representation model depending on tasks [40]. However, the goal of self-similarity is to sparsely represent each sample by other samples (e.g., $\|\mathbf{X}^T - \mathbf{X}^T\mathbf{M}\|_F^2 \Rightarrow \mathbf{x}^i \approx \sum_{j=1}^n \mathbf{x}^j m_{ij}$, where m_{ij} is the similarity between the i-th sample $\mathbf{x}^i$ and the j-th sample $\mathbf{x}^j$, $i, j = 1, \ldots, n$.). In this paper, our goal is to sparsely represent each feature by other features, i.e., $\mathbf{x}_i \approx \sum_{j=1}^d \mathbf{x}_j w_{ji}$, $i, j = 1, \ldots, d$. This indicates our model to conduct a feature-level feature selection.

In the proposed feature-level self-representation, representative features are a subset of the given d features. Intuitively, a representative feature is one that other features are relevant to, and can be used to describe or represent other features. Formally, if the r-th feature is selected by the g-th feature as a representative feature, it is expected that the model parameters for the g-th feature (i.e., $\mathbf{w}^g$) is similar to those of the r-th feature (i.e., $\mathbf{w}^r$). To describe one feature in an accurate way, one representative feature can be insufficient to capture all important characteristics of the feature. Furthermore, the similarity