Neurocomputing 000 (2018) 1–14

Contents lists available at ScienceDirect

Neurocomputing

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

Graph autoencoder-based unsupervised feature selection with broad and local data structure preservation

Siwei Feng, Marco F. Duarte*

Department of Electrical and Computer Engineering, University of Massachusetts Amherst, Amherst, MA 01003, USA

ARTICLE INFO

Article history:
Received 7 January 2018
Revised 20 April 2018
Accepted 21 May 2018
Available online xxx

Communicated by Domenico Ciuonzo

Keywords:
Unsupervised feature selection
Autoencoder
Manifold learning
Spectral graph analysis
Column sparsity

ABSTRACT

Feature selection is a dimensionality reduction technique that selects a subset of representative features from high-dimensional data by eliminating irrelevant and redundant features. Recently, feature selection combined with sparse learning has attracted significant attention due to its outstanding performance compared with traditional feature selection methods that ignores correlation between features. These works first map data onto a low-dimensional subspace and then select features by posing a sparsity constraint on the transformation matrix. However, they are restricted by design to linear data transformation, a potential drawback given that the underlying correlation structures of data are often non-linear. To leverage a more sophisticated embedding, we propose an autoencoder-based unsupervised feature selection approach that leverages a single-layer autoencoder for a joint framework of feature selection and manifold learning. More specifically, we enforce column sparsity on the weight matrix connecting the input layer and the hidden layer, as in previous work. Additionally, we include spectral graph analysis on the projected data into the learning process to achieve local data geometry preservation from the original data space to the low-dimensional feature space. Extensive experiments are conducted on image, audio, text, and biological data. The promising experimental results validate the superiority of the proposed method.

1. Introduction

In recent years, high-dimensional data can be found in many areas such as computer vision [1–3], pattern recognition [4–7], data mining [8], etc. High dimensionality enables data to include more information. However, learning high-dimensional data often suffer from several issues. For example, with a fixed number of training data, a large data dimensionality can cause the so-called Hughes phenomenon, i.e., a reduction in the generalization of the learned models due to overfitting during the training procedure compared with lower dimensional data [9]. Moreover, high-dimensional data tend to include significant redundancy in adjacent features, or even noise, which leads to large amounts of useless or even harmful information being processed, stored, and transmitted [10,11]. All these issues present challenges to many conventional data analysis problems. Moreover, several papers in the literature have shown that the intrinsic dimensionality of high-dimensional data is actually small [7,12–14]. Thus, dimensionality reduction is a popular preprocessing step for high-dimensional data analysis, which decreases time for data processing and also improves generalization of learned models. Feature selection [15–20] is a set of frequently used dimensionality reduction approaches that aim at selecting a subset of features. Feature selection has the advantage of preserving the same feature space as that of raw data. Feature selection methods can be categorized into groups based on different criteria summarized below; refer to [21] for a detailed survey on feature selection.

- *Label availability.* Based on the availability of label information, feature selection algorithms can be classified into supervised [15–17], semi-supervised [18–20], and unsupervised [22–38] methods. Since labeled data are usually expensive and time-consuming to acquire [39,40], unsupervised feature selection has been gaining more and more attention recently and is the subject of our focus in this work.
- *Search strategy.* In terms of selection strategies, feature selection methods can be categorized into wrapper, filter, and embedded methods. Wrapper methods [41,42] are seldom used in practice since they rely on a repetition of feature subset searching and selected feature subset evaluation until some stopping criteria or some desired performance are reached, which requires an exponential search space and thus is computationally prohibitive when feature dimensionality is high. Filter feature

* Corresponding author.
E-mail addresses: siwei@umass.edu (S. Feng), mduarte@ecs.umass.edu (M.F. Duarte).

https://doi.org/10.1016/j.neucom.2018.05.117
0925-2312/© 2018 Published by Elsevier B.V.

selection methods, e.g., Laplacian score [22] and SPEC [23], assign a score (measuring task relevance, redundancy, etc.) to each feature and select those with the best scores. Though convenient to computation, these methods are often tailored specifically for a given task and may not provide an appropriate match to the specific application of interest [21]. Embedded methods combine feature selection and model learning and provide a compromise between the two earlier extremes, as they are more efficient than wrapper methods and more task-specific than filter methods. In this paper, we focus on embedded feature selection methods.

In recent years, feature selection algorithms aiming at selecting features that preserve intrinsic data structure (such as subspace or manifold structure) [24–38] have attracted significant attention due to their good performance and interpretability [21]. In these methods, data are linearly projected onto new spaces through a transformation matrix, with fitting errors being minimized along with some sparse regularization terms. Feature importance is usually scored using the norms of corresponding rows/columns in the transformation matrix. In some methods [28–33,36–38], the local data structure, which is usually characterized by nearest neighbor graphs, is also preserved in the low-dimensional projection space. A more detailed discussion on this type of methods is in Section 2.1. One basic assumption of these methods is that the data to be processed lie in or near a completely linear low-dimensional manifold, which is then modeled as a linear subspace.[1] However, this is not always true in practice, in particular with more sophisticated data.

In the case when data lies on or close to more generalized or non-linear manifolds, many approaches for dimensionality reduction have been proposed that leverage the data local geometry using neighborhood graphs, such as ISOMAP [43], Laplacian eigenmaps [44], locally linear embedding [45], etc., but few developments have been reported in feature selection. In this paper, we propose a novel algorithm for graph and autoencoder-based feature selection (GAFS). The reason we choose an autoencoder for the underlying manifold learning is because of its broader goal of data reconstruction, which is a good match in spirit for an unsupervised feature selection framework: we expect to be able to infer the entire data vector from just a few of its dimensions. In this method, we integrate three objective functions into a single optimization framework: (*i*) we use a single-layer autoencoder to reconstruct the input data; (*ii*) we use an $\ell_{2,1}$-norm penalty on the columns of the weight matrix connecting the autoencoder's input layer and hidden layer to provide feature selection; and (*iii*) we preserve the local geometric structure of the data through to the corresponding hidden layer activations. To the best of our knowledge, we are the first to combine unsupervised feature selection with an autoencoder design and the preservation of local data structure. Extensive experiments are conducted on image data, audio data, text data, and biological data. Many experimental results are provided to demonstrate the outstanding performance achieved by the proposed method compared with other state-of-the-art unsupervised feature selection algorithms. The key contributions of this paper are highlighted as follows.

- We propose a novel unsupervised feature selection framework which is based on an autoencoder and graph data regularization. By using this framework, the information of the underlying data subspace can be leveraged, which loosens the assumption of linear manifold in many relevant techniques.
- We present an efficient solver for the optimization problem underlying the proposed unsupervised feature selection scheme. Our approach relies on an iterative scheme based on the gradient descent of the proposed objective function.
- We provide multiple numerical experiments that showcase the advantages of the flexible models used in our feature selection approach with respect to the state-of-the-art approaches from the literature.

The rest of this paper is organized as follows. Section 2 overviews related work. The proposed framework and the corresponding optimization scheme are presented in Section 3. Experimental results and the corresponding analysis are provided in Section 4. Section 5 includes conclusion and future work.

2. Related work

In this section, we provide a review of literature related to our proposed method and introduce the paper's notation standard. Datasets are denoted by $\mathbf{X} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(n)}] \in \mathbb{R}^{d\times n}$, where $\mathbf{X}^{(i)} \in \mathbb{R}^d$ is the ith sample in $\mathbf{X}$ for $i = 1, 2, \ldots, n$, and where d and n denote data dimensionality and number of data points in $\mathbf{X}$, respectively. For a matrix $\mathbf{X}$, $\mathbf{X}^{(q)}$ denotes the qth column of the matrix, while $\mathbf{X}^{(p,q)}$ denotes the entry of the matrix at the pth row and qth column.

The $\ell_{r,p}$-norm for a matrix $\mathbf{W} \in \mathbb{R}^{a\times b}$ is denoted as

$$\|\mathbf{W}\|_{r,p} = \left(\sum_{j=1}^{b}\left(\sum_{i=1}^{a}|\mathbf{W}^{(i,j)}|^r\right)^{p/r}\right)^{1/p}. \tag{1}$$

Two common norm choices in optimization are the $\ell_{2,1}$-norm and the Frobenius norm (e.g., $r = p = 2$). Note that unlike most of the literature, our outer sum is performed over the ℓ_r-norms of the matrix columns instead of its rows; this is done for notation convenience of our subsequent mathematical expressions.

The trace of a matrix $\mathbf{L} \in \mathbb{R}^{a\times a}$ is defined as

$$\mathrm{Tr}(\mathbf{L}) = \sum_{i=1}^{a}\mathbf{L}^{(i,i)}, \tag{2}$$

which is the sum of elements on the main diagonal of $\mathbf{L}$.

We use $\mathbf{1}$ and $\mathbf{0}$ to denote an all-ones and all-zeros matrix or vector with of the appropriate size, respectively.

2.1. Sparse learning-based unsupervised feature selection

Many unsupervised feature selection methods based on subspace structure preservation have been proposed in the past decades. For classes missing labels, unsupervised feature selection methods select features that are representative of the underlying subspace structure of the data [24]. The basic idea is to use a transformation matrix to project data to a new space and guide feature selection based on the sparsity of the transformation matrix [25]. To be more specific, the generic framework of these methods is based on the optimization

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{Y}, \mathbf{WX}) + \lambda\mathcal{R}(\mathbf{W}), \tag{3}$$

where $\mathbf{Y} = [\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \ldots, \mathbf{Y}^{(n)}] \in \mathbb{R}^{m\times n}$ $(m < d)$ is an embedding matrix in which $\mathbf{Y}^{(i)} \in \mathbb{R}^m$ for $i = 1, 2, \ldots, n$ denotes the representation of data point $\mathbf{X}^{(i)}$ in the obtained low-dimensional subspace. $\mathcal{L}(\cdot)$ denotes a loss function, and $\mathcal{R}(\cdot)$ denotes a regularization function on the transformation matrix $\mathbf{W} \in \mathbb{R}^{m\times d}$. The methods differ in their choice of embedding $\mathbf{Y}$ and loss and regularization functions; some examples are presented below.

[1] People also refer to linear manifold as subspace or linear subspace in the literature. In the sequel, we refer to such a linear manifold or subspace as a subspace for conciseness.

Please cite this article as: S. Feng, M.F. Duarte, Graph autoencoder-based unsupervised feature selection with broad and local data structure preservation, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.05.117