Knowledge-Based Systems 000 (2017) 1–10

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Unsupervised feature selection via low-rank approximation and structure learning

Shiping Wang[a,b], Han Wang[b,*]

[a] College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China
[b] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 3 April 2016
Revised 19 October 2016
Accepted 1 March 2017
Available online xxx

Keywords:
Machine learning
Feature selection
Unsupervised learning
Low-rank approximation
Structure learning

ABSTRACT

Feature selection is an important research topic in machine learning and computer vision in that it can reduce the dimensionality of input data and improve the performance of learning algorithms. Low-rank approximation techniques can well exploit the low-rank property of input data, which coincides with the internal consistency of dimensionality reduction. In this paper, we propose an efficient unsupervised feature selection algorithm, which incorporates low-rank approximation as well as structure learning. First, using the self-representation of data matrix, we formalize the feature selection problem as a matrix factorization with low-rank constraints. This matrix factorization formulation also embeds structure learning regularization as well as a sparse regularized term. Second, we present an effective technique to approximate low-rank constraints and propose a convergent algorithm in a batch mode. This technique can serve as an algorithmic framework for general low-rank recovery problems as well. Finally, the proposed algorithm is validated in twelve publicly available datasets from machine learning repository. Extensive experimental results demonstrate that the proposed method is capable to achieve competitive performance compared to existing state-of-the-art feature selection methods in terms of clustering performance.

1. Introduction

Data are represented with high-dimensional feature vectors in a variety of machine learning and computer vision tasks, such as text document clustering, gene expression, speech recognition and image classification. These huge features are more likely to describe the data in an overlapping and even repeated mode, which implies the redundancy, inconsistency and irrelevance of features. In fact, an excess of features may weaken the inherent relation among features and increase the risk of over-fitting, which could lower the performance of learning tasks and deteriorate the generalization ability of learning algorithms. Various of dimensionality reduction methods have been proposed to capture consistent and discriminative features. Dimensionality reduction techniques can be mainly classified into two types: feature extraction and feature selection. Feature extraction is to search an optimal transformation to project original feature space into a low-dimensional subspace by a transformation satisfying certain criterion [1,2]. Some typical representatives of feature extraction are principal component analysis (PCA) [3], singular value decomposition (SVD) [4] and linear discriminant analysis (LDA) [5]. Feature selection aims at choosing an optimal feature subset that can best represent the original feature space [6,7]. Feature selection has always been a popular research topic in that it can improve model interpretability, shorten model training time and enhance model generalization ability [8,9].

By whether the class label information is available, feature selection methods can be categorized into supervised and unsupervised methods. Supervised feature selection is to search the most discriminative feature subset with the guidance of class label information [10–12]. Due to the explicit and sufficient available information, this type of feature selection technique can efficiently find an optimal feature subset for learning tasks. However, in many cases, labeling all samples is often time-consuming and costly unaffordable, which requires us to select the most relevant feature subset for learning algorithms in an unsupervised manner. Due to the unavailability of class label information, unsupervised feature selection of10 aims to exploit the intrinsic structures of data points, where the learned certain structure serves as pseudo supervised information to search the best feature subset [13,14]. Since the learned structure information is often indeterminate, insufficient and ambiguous, unsupervised feature selection is a more challenging task [15–17].

* Corresponding author.
E-mail addresses: shipingwangphd@gmail.com (S. Wang), hw@ntu.edu.sg (H. Wang).

http://dx.doi.org/10.1016/j.knosys.2017.03.002
0950-7051/

Please cite this article as: S. Wang, H. Wang, Unsupervised feature selection via low-rank approximation and structure learning, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.03.002

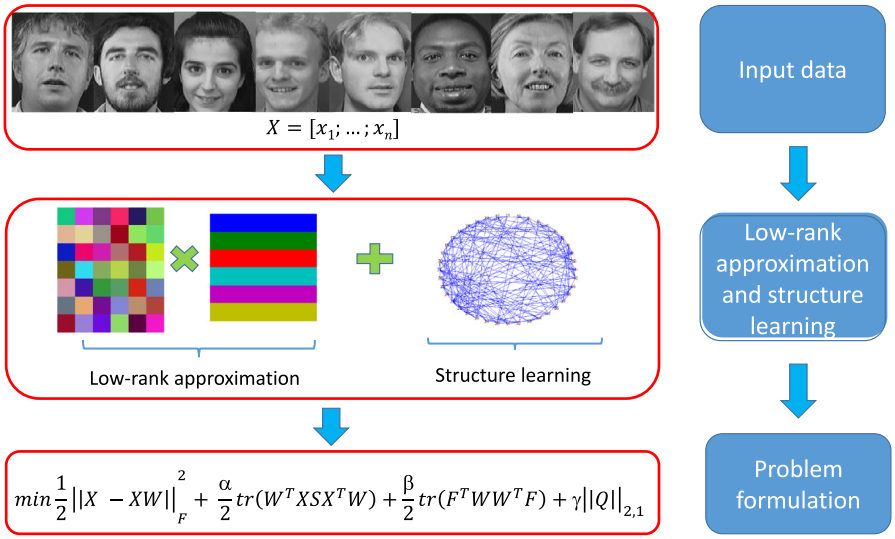

Fig. 1. Framework of this paper. With given data points, the feature selection problem is formulated as a form of matrix factorization, which integrates low-rank approximation and structure learning.

Due to the absence of class label information, structure learning acting as a pseudo supervised information has been widely used in unsupervised feature selection. Some typical structure learning algorithms of a feature space include locally local learning [18–20], locally linear embedding [21,22], global structure preserving [23,24], $\ell_{2,1}$-based sparsity preserving [25,26], graph structure learning [27], sparse regression [28–30], spectral analysis [31,32] and manifold structure embedding [33,34]. Besides, feature selection problems can be regarded as searching the most relevant feature subset that spans the low-dimensional subspace with the best characterization to the original feature space, which has an inherent consistency with low-rank approximation. Therefore, the aforementioned analyses motivate us to exploit the common advantages of low-rank approximation and structure learning, and develop an effective algorithm for feature selection.

In this paper, we propose an efficient unsupervised feature selection via a unified framework of low-rank approximation and structure learning. First, using the self-representation of data matrix, we formalize feature selection problems as matrix factorizations that exploit the advantages of structure learning regularization and low-rank approximation. Simultaneously, $\ell_{2,1}$-norm is used to capture the sparsity of the feature selection indicator matrix. Second, we present two techniques to deal with low-rank constraints and $\ell_{2,1}$-norm minimization. Based on the two techniques, an efficient algorithm with proved convergence is developed. These two techniques also provide a framework for low-rank recovery problems and $\ell_{2,1}$-norm minimization problems. The framework of this paper is demonstrated in Fig. 1. Finally, the proposed algorithm is compared with the Baseline algorithm and several state-of-the-art feature selection methods in twelve publicly available datasets from machine learning repository. Ranging from face images, object images, speech data and text documents to gene expression, these datasets serve as a good test bed for feature selection methods. Comprehensive experiments demonstrate that the proposed method come with competitive performance over the compared state-of-the-arts in terms of clustering performance in the tested datasets.

The rest of this paper is arranged as follows. The feature selection problem is formulated in Section 2, and the corresponding algorithm with proved convergence is proposed in Section 3. Experimental results are displayed and analyzed in Section 4, and discussing remarks are provided in Section 5. Finally, this paper is concluded in Section 6.

2. Problem formulation

We first recall some notations to be used throughout this paper. For any vector $v = (v_1, \cdots, v_n)^T \in \mathbb{R}^n$, $||v||_2 = \sqrt{\sum_{i=1}^n v_i^2}$. For an arbitrary matrix $M = (M_{ij})_{n\times m} \in \mathbb{R}^{n\times m}$, $||M||_F$ denotes the Frobenius norm, i.e., $||M||_F = \sqrt{\sum_{j=1}^m \sum_{i=1}^n M_{ij}^2}$, $||M||_{2,1}$ is the $\ell_{2,1}$-norm, i.e., $||M||_{2,1} = \sum_{i=1}^n \sqrt{\sum_{j=1}^m M_{ij}^2}$, and $rank(M)$ is the rank of M. For a square matrix $M \in \mathbb{R}^{n\times n}$, $Tr(M)$ is the trace of M, i.e., $Tr(M) = \sum_{i=1}^n M_{ii}$.

Denote data matrix as $X = (x_1; \cdots; x_n) = (f_1, \cdots, f_d) \in \mathbb{R}^{n\times d}$ where $x_i \in \mathbb{R}^d$ is a sample and $f_j \in \mathbb{R}^n$ is a feature vector. Using the regularized self-representation of data matrix, unsupervised feature selection problem can be formalized as

$$\arg\min_W \frac{1}{2}||X - XW||_F^2 + \gamma R(W) \tag{1}$$

where $W \in \mathbb{R}^{d\times d}$ represents the coefficient matrix to indicate which features are selected, $R(W)$ denotes the regularized term and γ is the coefficient to keep a trade-off between fitting accuracy and regularization. Let $W = (w_1; \cdots; w_d)$ where w_i is a d-dimensional column vector. Hereby, $||w_i||_2$ serves as a weight of ith feature. The bigger the weighted coefficient $||w_i||_2$ is, the more important the ith feature is.

On one hand, in order to avoid the trivial solutions to the aforementioned optimization problem, we attempt to choose an appropriate regularization term $R(W)$ from the viewpoint of low-rank approximation. The principle of unsupervised feature selection is to search a low-dimensional feature subset that best characterizes the original feature space. Therefore, the indicator matrix of a feature selector is assumed to be constrained by low-rank. On the other hand, structure learning is of critical importance to varying unsupervised learning problems [35,36], however, the above optimization problem does not consider the intrinsic structure of the input data. Incorporating the low-rank approximation and structure learning into unsupervised feature selection, the corresponding objective optimization problem is written as

$$\arg\min_W \frac{1}{2}||X - XW||_F^2 + \frac{\alpha}{2}Tr(W^TX^TSXW) + \gamma R(W) \quad \text{subject to } rank(W) = k \tag{2}$$

where S is a structure learning regularizer, α is a coefficient to adjust the weight of structure learning, γ is a regularization pa-

Please cite this article as: S. Wang, H. Wang, Unsupervised feature selection via low-rank approximation and structure learning, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.03.002