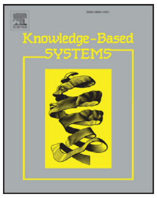
Knowledge-Based Systems 145 (2018) 109–120

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Robust unsupervised feature selection via dual self-representation and manifold regularization☆

Chang Tang[a], Xinwang Liu[b], Miaomiao Li[b], Pichao Wang[c,*], Jiajia Chen[d,*], Lizhe Wang[a], Wanqing Li[c]

[a] School of Computer Science, China University of Geosciences, Wuhan 430074, PR China
[b] School of Computer, National University of Defense Technology, Changsha 410073, PR China
[c] School of Computing and Information Technology, University of Wollongong, NSW, 2500, Australia
[d] Department of Pharmacy, Huai'an Second People's Hospital Affiliated to Xuzhou Medical College, Huai'an, 223002, PR China

ARTICLE INFO

Article history:
Received 22 August 2017
Revised 25 December 2017
Accepted 4 January 2018
Available online 5 January 2018

MSC:
00-01
99-00

Keywords:
Unsupervised feature selection
Local geometric structure
Similarity preservation
Self-representation
Graph learning

ABSTRACT

Unsupervised feature selection has become an important and challenging pre-processing step in machine learning and data mining since large amount of unlabelled high dimensional data are often required to be processed. In this paper, we propose an efficient method for robust unsupervised feature selection via dual self-representation and manifold regularization, referred to as DSRMR briefly. On the one hand, a feature self-representation term is used to learn the feature representation coefficient matrix to measure the importance of different feature dimensions. On the other hand, a sample self-representation term is used to automatically learn the sample similarity graph to preserve the local geometrical structure of data which has been verified critical in unsupervised feature selection. By using $l_{2,1}$-norm to regularize the feature representation residual matrix and representation coefficient matrix, our method is robustness to outliers, and the row sparsity of the feature coefficient matrix induced by $l_{2,1}$-norm can effectively select representative features. During the optimization process, the feature coefficient matrix and sample similarity graph constrain each other to obtain optimal solution. Experimental results on ten real-world data sets demonstrate that the proposed method can effectively identify important features, outperforming many state-of-the-art unsupervised feature selection methods in terms of clustering accuracy (ACC) and normalized mutual information (NMI).

1. Introduction

Due to the rapid development of electronic sensors and social media, a huge amount of high-dimensional data have been produced in many practical applications [1–5]. Data with high dimensionality not only significantly increases the time and memory burden of the algorithms and computer hardware, but also degenerates the performance of many algorithms due to the curse of dimensionality and the existence of irrelevant, redundant and noisy dimensions. Feature selection [6], as an important step in many data processing tasks, aims to remove the irrelevant and redundant features from the original data, which can alleviate the curse of dimensionality, reduce the data storage space and time complexity of learning algorithms. Based on the availability of data labels, there are three categories of feature selection approaches: supervised [7,8], semi-supervised [9,10] and unsupervised [4,11–17]. Since obtaining the labels of data is a challenging and laborious task, as a special case of unsupervised feature learning [18–21], unsupervised feature selection, which aims to select a subset of features from high-dimensional data with unknown sample labels, has received more and more attention recently.

Various methods of unsupervised feature selection have been proposed and they can be mainly classified into three distinct types: filter methods [7,12], wrapper methods [22] and embedded methods [13,23–30]. Compared to the first two kinds of methods, embedded methods can achieve superior performance in many respects, and have been advanced rapidly. Among those embedded methods, most of them try to discover the local geometrical data structure which has been verified much more critical than global structure in unsupervised feature selection [31]. The most commonly used local geometrical structure preserving method is the graph Laplacian regularization [13,31–37]. There are two main in-

☆ Fully documented templates are available in the elsarticle package on CTAN.

* Corresponding authors.

E-mail addresses: tangchang@cug.edu.cn (C. Tang), 1022xinwang.liu@gmail.com (X. Liu), miaomiaolinudt@gmail.com (M. Li), pw212@uowmail.edu.au (P. Wang), jjiachen@outlook.com (J. Chen), lizhe.wang@gmail.com (L. Wang), wanqing@uow.edu.au (W. Li).

https://doi.org/10.1016/j.knosys.2018.01.009
0950-7051/© 2018 Elsevier B.V. All rights reserved.

dependent steps in classical graph based methods: (a) construct similarity matrix by exploring the local structure; (b) and then select representative features by graph regularization based learning. As a special case of representative learning, self-representation has been employed for unsupervised feature selection and achieved great success in recent years [26–30]. In these methods, each feature is represented as the linear combination of its relevant features, and some constraints are applied to the feature reconstruction term and feature representation coefficient matrix. In order to preserve the local geometrical structure of original data, the Laplacian graph regularization term has also been integrated into self-representation based unsupervised feature selection [13,17,30–32].

Though demonstrating promising performance in various applications, we have observed that there are at least two issues in the previous graph regularized self-representation based unsupervised feature selection methods. First, the Frobenius norm has been widely used to regularize the feature representation residual term [13,30–32], which makes many methods sensitive to data outliers, resulting in unsatisfactory feature selection performance. Second, the graph construction and feature representation coefficient matrix learning are independent, i.e., the similarity graph of data samples are computed in advance by using a set of predefined or hand-crafted distance functions, then the local geometrical data structure are preserved using the fixed graph. However, the quality of the graph will be significantly influenced by the parameters of the manually set distance functions, such as the number of neighbor assignment and the kernel width in traditional Gaussian kernel function.

In order to address the above mentioned issues, in this paper, we propose a robust unsupervised feature selection method via dual self-representation and manifold regularization, referred to as DSRMR briefly. Specifically, a feature self-representation term is used to learn the feature representation coefficient matrix which captures the importance of different feature dimensions. A sample self-representation term is used to automatically learn the similarity graph of data samples. Meanwhile, unlike the existing graph regularized unsupervised feature selection methods which use the fixed graph, we use the learnt similarity graph to preserve the local geometrical structure. In order to ensure the robustness to outliers, an $l_{2,1}$-norm is used to regularize the feature reconstruction residual term. For measuring the importance of different feature dimensions, the feature representation coefficient matrix is also constrained by the $l_{2,1}$-norm for row-sparsity. Clustering performance with the selected features on ten benchmark data sets demonstrates the effectiveness of the proposed method. In summary, the main contributions of this paper are highlighted as follows:

- We propose a robust unsupervised feature selection method by using dual self-representation and manifold regularization.
- A feature self-representation term and a sample self-representation term are used to learn the feature representation coefficient matrix and sample similarity graph, respectively.
- Different to previous graph regularized methods which use the fixed graph, we use the learnt similarity graph for local geometrical structure preservation. An $l_{2,1}$-norm is used to regularize the feature reconstruction residual term for robustness to outliers.
- We derive an iterative approach to solve the proposed optimization problem. The feature representation coefficient matrix and the sample similarity matrix can constrain each other to obtain optimal solution iteratively during the optimization process. Comprehensive experiments on ten benchmark data sets are conducted to show the effectiveness of the proposed method, and demonstrate the advantage over other state-of-the-art methods.

The rest of this paper is arranged as follows. Several related unsupervised feature selection works are reviewed in Section 2. In Section 3, we detail the propose unsupervised feature selection model. Section 4 presents the optimization algorithm for solving the proposed model, and the convergence and computational analysis are also provided in this section. Experimental results and parameters sensitivity analysis are shown in Section 5. Finally, we conclude this paper in Section 6.

2. Related works

During the past decades, a large number of unsupervised feature selection methods have been proposed to overcome the high-dimensional issue without sample label information. For example, the top ranked features with maximum variance are selected by the maximum variance method [38]. However, the features selected by the maximum variance method are not discriminative enough for classification. In order to select features with local manifold structure of data being preserved, the Laplacian score [12] has been proposed for unsupervised feature selection. Other metrics such as feature similarity [11] and trace ratio [39] have also been used for selecting valuable features. Cai et al. proposed Multi-Cluster Feature Selection (MCFS) [23], which can captures the local manifold structure via spectral analysis, and then selects the features which can best preserve the clustering structure. As a general framework for dimensionality reduction, flexible manifold embedding [8] has also been adopted by many feature selection methods. Qian et al. [40] proposed a robust unsupervised feature selection framework which uses flexible manifold embedding, nonnegative matrix factorization and $l_{2,1}$-norm to perform robust clustering and robust feature selection simultaneously. Hou et al. proposed a general framework for unsupervised feature selection by joint embedding learning and sparse regression [25]. In recent years, many self-representation based methods have been explored and shown promising results [26–28,30,33,41]. The assumption behind these methods is that each feature can be well approximated by the linear combination of its relevant features and the representation coefficient matrix with sparsity constrain can be used as the feature weights.

Recent research has indicated that preserving local geometric data structure is especially important for unsupervised feature selection [31]. To this end, the graph Laplacian regularization term has been widely used for preserving local geometric structure in many embedded unsupervised feature selection methods [13,30–33,41].

In the previous self-representation and graph regularized unsupervised feature selection methods, the Frobenius norm is usually used for regularizing the feature reconstruction residual term, which makes the models sensitive to outliers. In addition, most of existing graph regularized methods use a fixed similarity graph which is manually set in advance for local geometrical structure preservation. The unreliable similarity graph and improper number of neighbor assignment eventually lead to suboptimal result. In this paper, we propose a robust unsupervised feature selection via dual self-representation and manifold regularization to solve above mentioned two issues and experimental results on several benchmark data sets demonstrate the effectiveness of the proposed method and the advantages over other state-of-the-art ones.

3. Proposed method

In this section, we first give some notations used in this paper and then the details of the proposed DSRMR method will be elaborated hereafter.