Knowledge-Based Systems xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Selecting feature subset with sparsity and low redundancy for unsupervised learning

Jiuqi Han [1], Zhengya Sun *, Hongwei Hao

Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, 100190 Beijing, China

ARTICLE INFO

Article history:
Received 21 October 2014
Received in revised form 6 June 2015
Accepted 8 June 2015
Available online xxxx

Keywords:
Unsupervised feature selection
Nonnegative spectral analysis
Sparsity and low redundancy

ABSTRACT

Feature selection techniques are attracting more and more attention with the growing number of domains that produce high dimensional data. Due to the absence of class labels, many researchers focus on the unsupervised scenario, attempting to find an optimal feature subset that preserves the original data distribution. However, the existing methods either fail to achieve sparsity or ignore the potential redundancy among features. In this paper, we propose a novel unsupervised feature selection algorithm, which retains the preserving power, and implements high sparsity and low redundancy in a unified manner. On the one hand, to preserve the data structure of the whole feature set, we build the graph Laplacian matrix and learn the pseudo class labels through spectral analysis. By finding a feature weight matrix, we are allowed to map the original data into a low dimensional space based on the pseudo labels. On the other hand, to ensure the sparsity and low redundancy simultaneously, we introduce a novel regularization term into the objective function with the nonnegative constraints imposed, which can be viewed as the combination of the matrix norms $||\cdot||_{m_1}$ and $||\cdot||_{m_2}$ on the weights of features. An iterative multiplicative algorithm is accordingly designed with proved convergence to efficiently solve the constrained optimization problem. Extensive experimental results on different real world data sets demonstrate the promising performance of our proposed method over the state-of-the-arts.

1. Introduction

In many application domains, such as pattern recognition and computer vision, data is often represented by high dimensional feature vectors. In practice, not all the features are informative and discriminative, because many of them are redundant, trivial, or even noisy sometimes [1]. The superfluous features may bring some adverse effects, such as overfitting, low efficiency and poor performance [2]. Besides, a number of methods that are analytically or computationally manageable in low dimensional space may become completely intractable when the number of features reaches thousands or even more [3]. Therefore, reducing the data dimension is an indispensable part for data mining and machine learning tasks [4].

As an important direction for dimension reduction, feature selection aims to select an optimal feature subset from high dimensional data for a compact and accurate data representation [5]. That is to say, the important and informative features are kept, with the redundant and noisy ones removed. In recent years, many research efforts have been devoted to achieving this target. According to the availability of label information, these methods can be roughly grouped into three categories, i.e., supervised feature selection [6], semi-supervised feature selection [7] and unsupervised feature selection [8–10]. Typically, supervised and semi-supervised methods both require the label information to varying degree for feature selection [9]. As the scale of the accessible data is growing rapidly, the cost of manually labeling would increase excessively. Thus, how to select the salient features in unsupervised scenarios becomes appealing and demanding, while facing the challenge that there are no true labels to guide the search for a good feature subset [11–13]. Existing solutions mainly fall into two classes [14]: filters, which rely on the intrinsic properties of the data and are usually applied prior to any learning algorithms [10]; wrappers, in which feature selection is wrapped around the learning algorithms that is utilized to score the candidate features in the process [15]. In this paper, we are particularly interested in the unsupervised filter methods.

Lots of previous unsupervised filter methods have tackled the selection tasks by ranking features on the basis of whether they can preserve the structure of the original data. Since there are no

* Corresponding author. Tel.: +86 010 82544483.
E-mail addresses: jiuqi.han@ia.ac.cn (J. Han), zhengya.sun@ia.ac.cn (Z. Sun).
[1] Tel.: +86 010 82544483.

http://dx.doi.org/10.1016/j.knosys.2015.06.008
0950-7051/

class labels, researchers usually characterize this structure by building the graph Laplacian matrix based on the k-nearest neighbor graph [16], followed by learning the pseudo class labels through spectral analysis. In addition, either sparsity constraint (l_1-norm [8] or $l_{2,1}$-norm [17] regularization) or low redundancy constraint (minimum mutual information or Pearson correlation coefficient [18]) is imposed on the feature selection matrix, and merged into the learning process. However, in many cases, to gain the sparsity and low redundancy simultaneously is more desirable, both of which are essential and complementary for the improved performance. Generally speaking, the former ensures only a few informative features are picked at certain accuracy loss, while the latter helps select the features that have little correlation, even if they are uninformative. Apparently, integrating their advantages poses a great challenge for the conventional approaches.

For illustration purposes, we as well take $l_{2,1}$-norm as an example. $l_{2,1}$-norm regularization is able to ensure the matrix sparse in rows, i.e., some rows of the matrix shrink to zero if the corresponding features cannot well distinguish between the pseudo labels. Put it another way, different columns of the matrix tend to have nonzero entries on the same position, with the features having small scores for all the labels discarded. However, methods using $l_{2,1}$-norm regularization might get into trouble when dealing with some informative but redundant features. As the correlations between different features are neglected, redundant features would possibly be retained, which is not desired.

In light of this, we present a novel unsupervised feature selection method, named FSLR (Feature subset with Sparsity and Low Redundancy), which takes into account the sparsity and the redundancy among features besides the accurate representation of the data. On the one hand, to retain the structure of the data, we perform spectral analysis to guarantee the pseudo labels in accordance with the detected structure. Meanwhile, we find a weight matrix that maps the original data into a low dimensional space based on the pseudo labels. On the other hand, to ensure the sparsity and low redundancy simultaneously, we introduce a novel regularization term in formulating the objective function with the nonnegative constraints imposed, which can be viewed as the combination of the matrix norms $||\cdot||_{m_1}$ and $||\cdot||_{m_2}$ on the weights of features. In this way, we are allowed to get a compact and sound representation of the original data. Subsequently, a simple yet effective iterative multiplicative update rule is designed to solve the constrained optimization problem. We then analyze the convergence behavior as well as the computational complexity of the proposed algorithm. Finally, we evaluate the performance of the proposed method on different real world data sets, and the experimental results manifest its superiority over the state-of-the-arts.

Our key contributions are highlighted as follows.

- We design a novel measurement, which can be viewed as the combination of the matrix norms $||\cdot||_{m_1}$ and $||\cdot||_{m_2}$ on the weights of features, to estimate the sparsity and redundancy among features at the same time.
- We formulate the unsupervised feature selection problem which retains the preserving power, and implements high sparsity and low redundancy in a unified manner.
- We develop an efficient iterative multiplicative algorithm to solve the proposed objective, and analyze its convergence behavior as well as the computational complexity.

The remainder of this paper is organized as follows: in Section 2, we overview the related works. We formulate our method and provide an effective solution as well as its convergence behavior and complexity analysis in Section 3. After that, we discuss the similar characteristics of some existing methods and compare our proposed FSLR with them in Section 4. Description about the data sets, the details of the baselines, evaluation metrics and parameter settings are shown in Section 5. Section 6 gives the experimental results, followed by the conclusion and future work in Section 7.

2. Related works

Great efforts on filter models have been made to address the issue of unsupervised feature selection in the past decades [2,19]. Unsupervised filter approaches evaluate features according to the intrinsic properties of the data prior to any learning algorithms [20]. Initially, data variance is one of the most widely used and typical criterion. In this view, features with larger variances are more powerful in representing different classes. Based on the fact that the local structure of the data space is more important than the global structure in some learning problems, He et al. presented Laplacian Score (LS) [20] to evaluate the locality preserving power of features via the graph Laplacian. In the sequel, Zhao and Liu presented SPEC as a framework [21], which generated families of algorithms based on spectral graph theory, and took LS as a special case.

In essence, LS and SPEC estimate the quality of features independently, ignoring the sparsity and the redundancy among them. Hence, several algorithms have been presented to overcome these weaknesses. To decrease the redundancy among features, Wang et al. proposed a framework named maximum weight and minimum redundancy (MWMR) [18], under which an external feature weighting algorithm, such as LS, and an external redundancy measurement, such as mutual information, were combined in a unified criterion. And the optimal feature subset was selected via maximizing this criterion. Zhao et al. proposed a framework named Similarity Preserving Feature Selection (SPFS) [22], which encompassed several widely used feature selection criteria by virtue of linear kernel similarity preserving and multi-output regression.

Meanwhile, to ensure the sparsity of features, some methods resort to the regularization terms, such as l_1-norm, $l_{2,1}$-norm and l_2-norm. Cai et al. proposed Multi-Cluster Feature Selection (MCFS) [8], which used l_1-norm regularization to decrease the combination coefficients of different features, and spectral analysis to measure the importance of each feature for different classes. Compared with l_1-norm, $l_{2,1}$-norm regularization was used much more frequently by scholars. Nie et al. proposed Robust Feature Selection (RFS) [23], which emphasized $l_{2,1}$-norm minimization on both the loss function and the regularization term. Yang et al. incorporated discriminative analysis and $l_{2,1}$-norm minimization into a joint framework, referred to as Unsupervised Discriminative Feature Selectioin (UDFS) [5]. To indicate the discriminative ability of features, UDFS defined the local discriminative score based on the total scatter matrix and between class scatter matrix. Besides, Yan et al. proposed Joint Laplacian Feature Weights Learning (JLFWL) [16], which selected the important features based on l_2-norm regularization, and determined the optimal size of the feature subset according to the number of positive feature weights. They obtained the weights of features by solving the spectral composition problem and the least squares problem.

In recent years, several studies adopted the spectral analysis (manifold learning) and sparse regression jointly, for instance, Joint Embedding Learning and Sparse Regression (JELSR) [17], Nonnegative Discriminative Feature Selection (NDFS) [9] and Robust Unsupervised Feature Selection (RUFS) [10]. Broadly, JELSR and NDFS share the same objection function, yet differ in the graph Laplacian matrix and constraint on the feature selection

Please cite this article in press as: J. Han et al., Selecting feature subset with sparsity and low redundancy for unsupervised learning, Knowl. Based Syst. (2015), http://dx.doi.org/10.1016/j.knosys.2015.06.008