Image and Vision Computing xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Image and Vision Computing

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

Non-convex regularized self-representation for unsupervised feature selection☆

Pengfei Zhu[b], Wencheng Zhu[b], Weizhi Wang[a], Wangmeng Zuo[a,*], Qinghua Hu[b]

[a]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[b]School of Computer Science and Technology, Tianjin University, Tianjin 300072, China

ARTICLE INFO

Article history:
Received 2 May 2016
Received in revised form 8 September 2016
Accepted 16 November 2016
Available online xxxx

Keywords:
Self-representation
Unsupervised feature selection
Sparse representation
Group sparsity

ABSTRACT

Feature selection aims to select a subset of features to decrease time complexity, reduce storage burden and improve the generalization ability of classification or clustering. For the countless unlabeled high dimensional data, unsupervised feature selection is effective in alleviating the curse of dimensionality and can find applications in various fields. In this paper, we propose a non-convex regularized self-representation (RSR) model where features can be represented by a linear combination of other features, and propose to impose $L_{2,p}$-norm ($0 \le p < 1$) regularization on self-representation coefficients for unsupervised feature selection. Compared with the conventional $L_{2,1}$-norm regularization, when $p < 1$, much sparser solution is obtained on the self-representation coefficients, and it is also more effective in selecting salient features. To solve the non-convex ($0 < p < 1$) RSR model, we further propose an efficient iterative reweighted least square (IRLS) algorithm with guaranteed convergence to a stationary point. When $p = 0$, we exploit the augmented Lagrangian method (ALM) to solve the RSR model. Extensive experimental results on nine datasets show that our feature selection method with small p is more effective. It mostly outperforms RSR with $p = 1$ and other state-of-the-art unsupervised feature selection methods in terms of classification accuracy and clustering performance.

1. Introduction

With the booming of electronic sensors and networks, a huge amount of high-dimensional data have been produced [1]. The high-dimensional data not only increase the processing time and space complexity, but also deteriorate the performance of clustering or classification due to the curse of dimensionality [2]. Feature selection, which can effectively remove the irrelevant and redundant features, reduce the computational and storage complexity, and enhance the model generalization capability by selecting a subset of features, has become a necessary step to build a promising machine learning model for classification, clustering, and other tasks [3,4]. In recent years, many efforts have been devoted to developing new feature selection algorithms [5–8].

In general, feature selection methods fall into two categories: supervised and unsupervised, based on label availability. As there is no label information in unsupervised feature selection method, it is usually more difficult than that of supervised scenario. Besides, with the wide use of network and social media, there are large amounts of data which are unlabeled, and cannot be directly processed using supervised feature learning methods. Therefore, unsupervised feature learning has drawn much attention, and that is also what we focus on in this paper.

Unsupervised feature learning methods can be roughly categorized into three groups: filter models, wrapper models and embedding models. Filter models [7] select a subset of features by using some feature evaluation indices, i.e. some statistical properties of data, while wrapper models [9,10] search in the space of feature subset and take classification performance as evaluation criterion for feature selection. Wrapper models can be quite computationally intensive, especially for large-scale problems. In contrast to that, embedding models [11] incorporate the selection process in the learning model to simultaneously learn the optimal classifier while finding salient features. Leaving all the differences of the previous methods unconsidered, early studies on unsupervised feature selection mainly use some evaluation indices to evaluate the importance of each individual feature or feature subset. These important indices can be calculated from clustering performance, redundancy, sample similarity, manifold structure, and some representative indices like Laplacian score [5], variance [3], and trace ratio [12]. However, the dependence on searching makes these methods computationally expensive. To reduce computation, a no-searching feature

☆ This paper has been recommended for acceptance by Jiwen Lu.
* Corresponding author.

http://dx.doi.org/10.1016/j.imavis.2016.11.014
0262-8856/

clustering method based on feature similarity is proposed to find the representative features [13]. To best preserve sample similarity, a series of spectral clustering based feature selection methods have been developed [14–16]. Recently, Zhu et al. [17] innovatively proposed a regularized self-representation method for unsupervised feature selection. In their method, one feature is represented as a linear combination of other features, which is called self-representation property of features. By minimizing the self-representation error, a feature weight matrix is learned and a feature subset can be selected.

Meanwhile, sparsity regularization methods have been widely utilized in pattern recognition and computer vision area. They have been employed to dimensionality reduction and feature selection, and achieved some favorable results. By imposing L_1-norm regularization, L_1-SVM [18] was proposed to perform feature selection. By modeling feature selection as a loss minimization problem, the $L_{2,1}$-norm group sparsity has also been introduced to feature selection [4,19,20] to remove the redundancy among features. And it is also employed in the method of Zhu et al. [17], where $L_{2,1}$-norm was used to regularize the feature weight matrix and self-representation error, and has led to the state-of-the-art results.

In this paper, we propose to use $L_{2,p}$-norm regularization to select features with emphasis on small p ($0 \leq p < 1$) values. Just like the situation in vector l_1-norm vs. l_p-norm, when $0 < p < 1$, the non-zero rows of the resolved representation coefficient matrix will become sparser than that of the standard $L_{2,1}$-norm. To further impose sparsity on coefficient matrix, the limit case of $p = 0$ will also be considered, where we define the induced regularization as $L_{2,0}$-norm, although it is indeed not a genuine norm. At the same time, to eliminate the adverse effect of outliers, we use the standard $L_{2,1}$-norm to regularize the loss term. An improved Iterative Reweighted Lease Square (IRLS) algorithm is proposed to solve the $0 < p < 1$ model whose convergence can be ensured. On the other hand, the $p = 0$ model is quite non-convex, and non-differentiable; therefore, the IRLS method is not applicable to this situation. Considering that, the effective augmented Lagrange method (ALM) [21] is utilized to solve this problem, which can make sure that our iterative computation is locally convergent. Experiments are conducted on real-world datasets, and validate that features selected by our models are more effective than that of the standard $L_{2,1}$-norm regularization and other popular feature selection methods in terms classification and clustering metrics.

This paper is an extended version of our conference paper [22]. In this work, we consider the case when $p = 0$ and compare the impact of p values on feature selection. The remaining of this paper is organized as follows: Section 2 introduces the model of unsupervised feature selection in this paper; Section 3 describes the optimization procedure and algorithms; Section 4 presents the comparative experiments and Section 5 concludes this paper.

2. Non-convex regularized self-representation model

2.1. Problem statement

In general, the real-world datasets are very redundant and may contain outliers. And a promising unsupervised feature learning algorithm is expected to select a desired feature subset from the given unlabeled dataset, which can effectively describe the dataset and is helpful for subsequent tasks.

Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ be the given dataset matrix, where m is the sample number, and n is the feature dimension. We use $\mathbf{f}_i \in \mathbb{R}^m$ to represent the ith feature vector of $\mathbf{X}$, then $\mathbf{X} = [\mathbf{f}_1, \ldots, \mathbf{f}_i, \ldots \mathbf{f}_n]$. The purpose of feature selection is to select k features, and use them for classification, clustering or other tasks. The previous feature selection methods use some indices, e.g., Laplacian score [5], trace ratio [12] as mentioned before, while recent methods tend to construct a response matrix by using the sample similarity or sample manifold structure, thereafter, the feature selection problem can be formulated as a multi-output regression problem:

$$J_0(\mathbf{X}_K) = \min_{K \subset D, \mathbf{W}} l(\mathbf{Y} - \mathbf{X}_K \mathbf{W}) \tag{1}$$

where $D = \{1, 2, \ldots, n\}$ is the dimension, and K is the selected subsets, $\mathbf{X}_K$ is the corresponding K columns of $\mathbf{X}$, $\mathbf{W}$ is the corresponding feature weight matrix, and $l(\cdot)$ is the loss term, which is used to evaluate the performance of feature selection.

Obviously, this is a discrete optimization problem and the number of feasible feature subsets is $C_n^k = n!/k!(n-k)!$. It is a NP-hard problem, which will be intractable using brute force computation with the feature dimension increasing. Rather than directly solving this challenging discrete optimization problem, we incorporate some regularization on $\mathbf{W}$, resulting in the following formulation:

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

where, $l(\mathbf{Y} - \mathbf{X}\mathbf{W})$ is the loss term just like that in Eq. (1), the newly introduced term $R(\mathbf{W})$ is the regularization that combined with the constant parameter λ helps dynamically choose the optimal feature subset while helps choose and calculate the optimal weight matrix.

2.2. Loss term and regularization term

Inspired by the sample representation models [17], we utilize the property of feature self-representation to realize feature selection. Like RSR [17], we use the original space data matrix $\mathbf{X}$ as the response matrix, i.e., $\mathbf{Y} = \mathbf{X}$, then each feature can be linearly represented by all the features, that is, for each feature vector $\mathbf{f}_i$ in $\mathbf{X}$, it can be represented as follows:

$$\mathbf{f}_i = \sum_{j=1}^{n} \mathbf{f}_j \mathbf{W}_{ji} + \mathbf{b}_i \tag{3}$$

where $\mathbf{W}_{ji}$ is the ji-th component of $\mathbf{W}$ and $\mathbf{b}_i \in \mathbb{R}^m$ is the bias vector. Putting the whole features together, we have

$$\mathbf{X} = \mathbf{X}\mathbf{W} + \mathbf{B} \tag{4}$$

where $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n] \in \mathbb{R}^{m \times n}$.

In this model, we will use the learned weight matrix $\mathbf{W}$ to reflect the importance of different features when the bias is small. Let $\mathbf{W} = \left[\mathbf{w}_1^T, \mathbf{w}_2^T, \ldots, \mathbf{w}_n^T\right]^T$, where $\mathbf{w}_i$ is the ith row of $\mathbf{W}$, while T represents the matrix/vector transpose operation. $\|\mathbf{w}_i\|_2$ can imply the importance of the ith feature in the representation, that is, if the ith feature contributes nothing to feature representation, then $\|\mathbf{w}_i\|_2$ will be 0; on the contrary, if the ith feature is frequently used to represent most of the features, then $\|\mathbf{w}_i\|_2$ must be significant. Apparently, the row-sparsity is expected to describe the property of weight matrix $\mathbf{W}$.

As we all know that, in the theory of sparse representation, l_0-norm can lead to sparse results, while it is non-convex. Therefore, l_1-norm is widely used as the convex alternative to l_0-norm, and under some conditions they are equivalent [22]. However, as some researchers [23] point out, l_p-norm with $0 < p < 1$ is more similar to l_0-norm and can produce better sparse representation results than that of l_1-norm. Considering that, we select the $l_{2,p}$-norm regularizer, where $0 < p < 1$ or $p = 0$ to construct the regularization term of $\mathbf{W}$. With this setting, the solution of representation weight matrix $\mathbf{W}$ will be even sparser in rows. It is worth noting that, in this paper, we explicitly take into account the situation of $p = 0$, which is different from that of $0 < p < 1$. Therefore, in the following, we will take these two settings separately.

Please cite this article as: P. Zhu et al., Non-convex regularized self-representation for unsupervised feature selection, Image and Vision Computing (2016), http://dx.doi.org/10.1016/j.imavis.2016.11.014