Pattern Recognition 48 (2015) 438–446

Contents lists available at ScienceDirect

Pattern Recognition

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

Unsupervised feature selection by regularized self-representation

Pengfei Zhu [a,*], Wangmeng Zuo [a,b], Lei Zhang [a], Qinghua Hu [c], Simon C.K. Shiu [a]

[a] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[b] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[c] School of Computer Science and Technology, Tianjin University, Tianjin 300072, China

ARTICLE INFO

Article history:
Received 8 January 2014
Received in revised form
16 July 2014
Accepted 9 August 2014
Available online 19 August 2014

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

ABSTRACT

By removing the irrelevant and redundant features, feature selection aims to find a compact representation of the original feature with good generalization ability. With the prevalence of unlabeled data, unsupervised feature selection has shown to be effective in alleviating the curse of dimensionality, and is essential for comprehensive analysis and understanding of myriads of unlabeled high dimensional data. Motivated by the success of low-rank representation in subspace clustering, we propose a regularized self-representation (RSR) model for unsupervised feature selection, where each feature can be represented as the linear combination of its relevant features. By using $L_{2,1}$-norm to characterize the representation coefficient matrix and the representation residual matrix, RSR is effective to select representative features and ensure the robustness to outliers. If a feature is important, then it will participate in the representation of most of other features, leading to a significant row of representation coefficients, and vice versa. Experimental analysis on synthetic and real-world data demonstrates that the proposed method can effectively identify the representative features, outperforming many state-of-the-art unsupervised feature selection methods in terms of clustering accuracy, redundancy reduction and classification accuracy.

1. Introduction

The explosive use of electronic sensors and social media produces a huge amount of high-dimensional data [1,2], and the high dimensionality greatly increases the time and space complexities for data processing, making the clustering and classification methods, which are analytically or computationally manageable in low-dimensional space, completely intractable [3]. Feature selection is an important step to remove the irrelevant and redundant features from the original data [4], alleviating the curse of dimensionality, reducing the storage space and time complexity, and building a compact data representation with good generalization ability [5,6]. In recent years, continuous efforts have been made to develop new feature selection algorithms [3,6–12].

Feature selection methods can be categorized into unsupervised and supervised ones [4,5,13]. Supervised feature selection methods include wrapper models and filter models. Wrapper models search in the space of feature subset, and employ one classifier to repeatedly evaluate the goodness of the selected feature subsets, making it computationally intensive and intractable for large-scale problems [5]. Filter models are independent of certain classifiers and they use some feature evaluation indices to rank features or evaluate feature subsets, e.g., Fisher score.

In many data mining applications, sample labels are unknown, therefore making unsupervised feature selection indispensable [14]. Early unsupervised feature selection methods mainly use some evaluation indices to evaluate each individual feature or feature subset, and then select the top K features or the best feature subset. These indices evaluate the clustering performance, redundancy, information loss, sample similarity or manifold structure, e.g., variance [5], Laplacian score [7], or trace ratio [15]. These methods, however, are computationally expensive in searching. To reduce the computational cost, a feature clustering method is proposed in [14] to find the representative features based on feature similarity without searching. Recently, a series of algorithms have been developed based on spectral clustering techniques to select a feature subset that best preserves the sample similarity [3,6,7,15–17]. In [7,15,16], features are selected one by one and the correlation between features is totally ignored [9], while in [3,6,17], the importance of features is evaluated jointly and features are selected in batch.

On the other hand, sparsity regularization has been widely used in feature selection and shown good effectiveness, robustness and efficiency, e.g., L_1-SVM [18] and sparse logistic regression [19]. Group sparsity, which is often used in multi-task learning [20] and joint representation [21], has also been applied to feature selection. By modeling feature selection as a loss minimization problem, in [8,9,6,17,22] group sparsity is imposed on the feature weights matrix to select features. The $L_{2,1}$-norm group sparsity

* Corresponding author. Tel.: +85267610177.
E-mail address: zhupengfeifly@gmail.com (P. Zhu).

http://dx.doi.org/10.1016/j.patcog.2014.08.006
0031-3203/

regularization has been adopted and shown good performance to remove the redundancy in feature selection [6,23].

Unlike supervised feature selection, in unsupervised feature selection the class label information is unavailable to guide the selection of minimal feature subset. In this paper, we find that the self-representation property of redundant features, which characterizes the property that each feature can be well approximated by the linear combination of its relevant features, also provides some insights on unsupervised feature selection. In nature, self-similarity widely exists, i.e., a part of an object is similar to other parts of itself, e.g., coastlines [24], stock market movements [25] and images [26]. Taking images for example, patches at different locations in an image perhaps are similar to each other, which is called non-local self-similarity. In image processing, the so-called non-local self-similarity has been successfully used in high performance image restoration and denoising [26]. Based on self-similarity of objects in nature, self-representation property generally holds for most high dimensional data, and has been extensively used in machine learning and computer vision fields. Just as sparsity leads to sparse representation, self-similarity results in self-representation.

With the above considerations, in this paper we propose a simple yet very effective unsupervised feature selection method by exploiting the self-representation ability of features. The feature matrix is represented over itself to find the representative feature components. The representation residual is minimized by $L_{2,1}$-norm loss to reduce the effect of outlier samples. Different from the other applications, in unsupervised feature selection, our goal is to identify a representative feature subset so that all the features can be well reconstructed by them. Thus, $L_{2,1}$-norm regularization is imposed on the representation coefficients to enforce group sparsity. With the proposed regularized self-representation model, if a feature is important, it will participate in the representation of other features and hence produce a significant row of representation coefficients and vice versa. To the best of our knowledge, this work is the first attempt to conduct unsupervised feature selection from the viewpoint of feature self-representation. Extensive experiments have been performed on synthetic and real-world data sets, and the results validate the leading performance of the proposed method in terms of clustering, redundancy and classification evaluation measures.

The rest of this paper is organized as follows: Section 2 introduces the unsupervised feature selection task; in Section 3, regularized self-representation is proposed; Section 4 presents the optimization and algorithms; Section 5 discusses the relationships between RSR and low rank representation; Section 6 conducts experiments and Section 7 concludes this paper.

2. Problem statement

The objective of unsupervised feature selection is to select a desired feature subset from a given dataset without label information. The real-world data are often very redundant in features and can have outlier samples. Fig. 1(a) illustrates a corrupted data matrix. Each row vector is a sample and each column vector represents one feature of all samples. The shaded central column represents a redundant feature, and the shaded central row represents an outlier sample. As shown in Fig. 1(b) and (c), a robust and effective feature selection algorithm should eliminate the effect of the outlier samples and indicate the redundant features.

Let $\boldsymbol{X} \in \Re^{n\times m}$ be a data matrix, where n and m are the numbers of samples and features, respectively. We use $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ to represent the n samples, $\boldsymbol{x}_i \in \Re^m$ and $\boldsymbol{X} = [\boldsymbol{x}_1; \boldsymbol{x}_2; \ldots; \boldsymbol{x}_n]$. We use $f_1, f_2, \ldots, f_m$ to denote the m features, and $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_m$ are the corresponding feature vectors, where $\boldsymbol{f}_i \in \Re^n$ and $\boldsymbol{X} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_m]$.

Early unsupervised feature selection methods use some metrics (e.g, variance, Laplacian score [7]) to evaluate each feature, and then rank the features by the evaluated metric values. The recently developed methods [17,6,9,3] usually first calculate the sample similarity or sample manifold structure, and then build a response matrix $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_m]$. The feature selection problem is then converted into a multi-output regression problem:

$$\min_{\boldsymbol{W}} l(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}) + \lambda R(\boldsymbol{W}) \tag{1}$$

where $\boldsymbol{W}$ is the feature weight matrix, $l(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W})$ is the loss item, $R(\boldsymbol{W})$ is the regularization imposed on $\boldsymbol{W}$ and λ is a positive constant.

In Eq. (1), the response matrix $\boldsymbol{Y}$ is known before the optimization phase and $\boldsymbol{W}$ is the variable. $\boldsymbol{Y}$ contains the sample similarity information and it is calculated differently in different methods. Taking minimum redundancy feature selection (MRFS) [6] for example, the sample similarity matrix $\boldsymbol{S}$ is first calculated, and then the elements of $\boldsymbol{Y}$ are determined as $\boldsymbol{y}_k = \lambda_k^{1/2}\boldsymbol{\xi}_k$, where λ_k and $\boldsymbol{\xi}_k$ are the kth eigenvalue and eigenvector of normalized similarity matrix $\hat{\boldsymbol{S}}$.

3. Regularized self-representation

The model in Eq. (1) considers the data similarity and selects features jointly. Though it is widely used in many feature selection methods, it is difficult to choose the proper response matrix. Thanks to self-representation property of features, in this section we propose a regularized self-representation (RSR) model for unsupervised feature selection. The proposed RSR model simply uses the data matrix $\boldsymbol{X}$ as the response matrix , i.e., $\boldsymbol{Y} = \boldsymbol{X}$, which is more natural and can be well interpreted by the self-representation principle, i.e., each feature can be well represented by all features. For each feature $\boldsymbol{f}_i$ in $\boldsymbol{X}$, we represent it as a linear combination of other features (including itself):

$$\boldsymbol{f}_i = \sum_{j=1}^{m} \boldsymbol{f}_j w_{ji} + \boldsymbol{e}_i \tag{2}$$

Then for all the features, we have

$$\boldsymbol{X} = \boldsymbol{X}\boldsymbol{W} + \boldsymbol{E} \tag{3}$$

where $\boldsymbol{W} = [w_{ji}] \in \Re^{m\times m}$ is the representation coefficients matrix. The above representation model is a kind of self-representation of features.

Clearly, the matrix $\boldsymbol{W}$ to be learned should reflect the importance of different features while making the representation residual $\boldsymbol{E}$ small. One may use the Frobenius norm to measure the residual, i.e., $\min_{\boldsymbol{W}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}\|_F^2$. However, as illustrated in Fig. 1, there can be some outlier samples in the data matrix $\boldsymbol{X}$, while the Frobenius norm is sensitive to outliers. Considering that an outlier sample is a row of the matrix $\boldsymbol{X}$, and its representation residual is a row in the matrix $\boldsymbol{E} = \boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}$, we propose to use the $L_{2,1}-$norm to characterize $\boldsymbol{E}$; that is, we impose row-sparsity on $\boldsymbol{E}$ to enforce robustness to outlier samples. Meanwhile, if we let $\boldsymbol{W}$ be an $m \times m$ identity matrix, a trivial solution will be obtained with the residual $\boldsymbol{E} = 0$. Thus, a regularization item $R(\boldsymbol{W})$ must be introduced to avoid the trivial solution of $\boldsymbol{W}$ and guide the selection of feature subset. Then we have the following minimization problem:

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}\|_{2,1} + \lambda R(\boldsymbol{W}) \tag{4}$$

Let $\boldsymbol{W} = [\boldsymbol{w}_1; \ldots; \boldsymbol{w}_i; \ldots; \boldsymbol{w}_m]$, where $\boldsymbol{w}_i$ is ith row of $\boldsymbol{W}$. $\|\boldsymbol{w}_i\|_2$ can be used as the feature weight because it reflects the importance of the ith feature in representation. For example, if $\|\boldsymbol{w}_i\|_2 = 0$, it means that the ith feature will contribute nothing to the representation of other features. If the ith feature take part in the representation of all features, then $\|\boldsymbol{w}_i\|_2$ must be significant. Therefore, the row-sparsity is expected for regularizing the coefficients matrix $\boldsymbol{W}$. We let