Neurocomputing 175 (2016) 686–697

Contents lists available at ScienceDirect

Neurocomputing

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

Cluster structure preserving unsupervised feature selection for multi-view tasks

Hong Shi [a,b], Yin Li [a,b], Yahong Han [a,b,*], Qinghua Hu [a,b]

[a] School of Computer Science and Technology, Tianjin University, Tianjin 300072, PR China
[b] Tianjin Key Lab of Cognitive Computing and Applications, Tianjin 300072, PR China

ARTICLE INFO

Article history:
Received 19 March 2015
Received in revised form
21 September 2015
Accepted 1 November 2015
Communicated by Feiping Nie
Available online 7 November 2015

Keywords:
Multi-view
Feature selection
Unsupervised
Cluster structure

ABSTRACT

Multi-view or multi-modal tasks exist in many areas of pattern analysis as the advancement of feature acquisition or extraction. These tasks are usually confronted with the issue of curse of dimensionality. In this work we consider the unsupervised feature selection problem for multi-view tasks. As most of the existing feature selection methods can only handle single-view data, we develop a new algorithm, called Cluster Structure Preserving Unsupervised Feature Selection (CSP-UFS). To leverage the complementary information between multiple views in unsupervised scenarios, we incorporate discriminative analysis, spectral clustering and correlation information between multiple views into a unified framework. Intuitionally speaking, the cluster structures of data in feature spaces reflect the discriminative information of distinct classes. Thus we introduce spectral clustering to discover the cluster structure and use discriminative analysis to preserve the structure. We design an alternating optimization algorithm to solve the proposed objective function. Experimental results on different datasets show the effectiveness of the proposed algorithm.

1. Introduction

In the era of Big Data, large-scale data exist in many domains. The term big not only refers to the fact that the size of instances are very large, but also mean that the ways of representing information are various [1]. In some tasks, objects are represented by different feature descriptors produced by different sources or different viewpoints. Thus there are plenty of approaches to generate features, and each approach reflects different aspects of the object and has different semantics and statistical properties. For instance, in video indexing and video content understanding, a variety of feature descriptors have been developed. For the key frames of an online video, multiple types of information can be used to represent the images, including color information, shape information, and texture information. For the video itself, motion information, audio information and text information can also provide hints for classification [1]. In order to differentiate from traditional single-view data, this kind of tasks is called as the multi-view data [2–4]. Multi-view data are widely used in machine learning and pattern recognition. Multi-view learning is now attracting much attention in recent years [5].

As to multi-view learning, objects are usually represented by high-dimensional feature vectors. Moreover, large amounts of data are unlabeled. In this case, there is an urgent requirement of the reduction of the feature dimension in machine learning. Feature selection aims to generate a relatively small feature subset compared with the original high dimensional data. It can make the computation much more efficient and even more accurate, and the redundancy or noisy information will be eliminated for a better model [6,7]. Feature selection has been proved to be an effective approach to process high-dimensional data [8,9].

In recent years, a variety of feature selection algorithms have been proposed. Feature selection algorithms can be roughly divided into two main groups according to whether the labels of samples are available: supervised and unsupervised feature selection. Discriminative information is often encoded in labels, and it plays a critically important role in many machine learning tasks. So supervised feature selection methods can find the discriminative features relatively easy by considering the label information. Based on discriminative information, lots of algorithms have been proposed, e.g., Fisher Score [10], trace ratio [11], sparse multi-output regression [12] and robust regression [13]. Liu et al. recently proposed a new supervised feature selection method [14]. It works in an agglomerative way, and a new evaluation function is introduced to measure the goodness of features.

However, in many practical applications, the label information is much expensive to obtain. We have vast volume of unlabeled

* Corresponding author at: School of Computer Science and Technology, Tianjin University, Tianjin 300072, PR China.
E-mail address: yahong@tju.edu.cn (Y. Han).

http://dx.doi.org/10.1016/j.neucom.2015.11.001
0925-2312/

samples, but we do not have their labels. In this situation, supervised feature selection will lose the benefits and fail to find the discriminative information encoded in labels. This motivates us to use semi-supervised or unsupervised feature selection methods to solve this problem. Semi-supervised feature selection methods utilize both labeled and unlabeled data to select the relevant features. Zhao and Liu [15] proposed a semi-supervised method by utilizing the spectral analysis. However, it is a filter-based method, and this kind of methods may discard important features [16]. Xu et al. [17] proposed an embedded feature selection algorithm based on manifold regularization, and it can discover more discriminative information. Based on spline regression, Han et al. proposed S^2FS^2R in [18], which utilize both the discriminative information and local geometry structure to implement this task. In addition, Chang et al. [19] introduced a novel convex semi-supervised method for multi-label data and it can be dealt with large-scale datasets.

For unsupervised feature selection methods, how to select the discriminative features in unsupervised scenarios is a key issue in some applications. Most algorithms address this problem by assigning each feature a score, which reflects the capability of the feature in preserving the structure of data or the power of dividing different categories. Then they select the top ranked features as the new expression of the original data. The representative algorithms include Principal component analysis (PCA), Laplacian Score [20]. In addition, there is another kind of approaches, which has caused many concerns over the recent years. These methods make use of spectral analysis on the graph Laplacian matrix to preserve the cluster structure of the original data. Meanwhile they apply some sparsity regularization models to select the features with large weights [21]. The typical methods include MCFS [22], UDFS [7], NDFS [23], RUFS [24] and RSFS [25], where RUFS and RSFS are the robust methods, and they can handle the noise on the data. Beyond that Zhu et al. [26] proposed a regularized self-representation (RSR) modal based on self-similarity of objects in nature.

In addition, Wang et al. [27] proposed to address the unsupervised feature selection problem by incorporating the unsupervised trace ratio formulation and sparsity regularization model, and they proved that the trace ratio formulation they used is a unified form of trace ratio linear discriminant analysis (LDA) and K-means clustering. Recently, Wang et al. [28] introduced a new method which attempts to maximize class margins of the transformed data and utilize K-means clustering to get the pseudo labels. These two methods integrate the linear discriminant analysis and K-means clustering to deal with the problem with the sparsity regularization model. The same is that our method also utilizes the LDA, and the difference is that we use spectral clustering to get the cluster structure as its good performance. Besides, our method can handle the multi-view tasks.

Unfortunately, most of the existing unsupervised feature selection algorithms can only handle the single-view task. There are two strategies to solve this problem. One is to concatenate different features into one feature space, and then apply the existing feature selection algorithms on the new vectors. This strategy ignores the difference of multiple views, while different views may have different semantics and statistical properties. Another strategy is to apply existing feature selection methods on each view separately and finally combine the selected features from different views of data. Clearly, this strategy has its disadvantage as well. We know the information from different views are relevant and complementary. The complementary information between multiple views is ignored if this strategy is introduced. AUMFS [2] and MVFS [9] can select features from multiple views by exploiting the complementary information among different views, simultaneously. However those methods predict the class labels of unsupervised data with a linear transformation. In fact, the assumption is not held in real-world applications as the transformation is usually nonlinear [29]. AMFS [30] utilize a local linear regression modal to learn the view-based Laplacian graphs and build a trace ratio objective function to solve this problem.

Inspired by the previous works on feature selection methods [2,7,11,23], we propose a new method in this work, called Cluster Structure Preserving Unsupervised Feature Selection (CSP-UFS). As we know that the discriminative information is generally encoded in the data labels, and it can distinguish different categories as much as possible. In unsupervised scenarios, we use the cluster structure to represent the label information. During the feature selection process, we try to preserve the cluster structure. So we can preserve the discriminative information, and the selected features can also have the capability to distinguish different categories. We simultaneously incorporate the discriminative analysis, cluster structure and correlation information between multiple views to implement an unsupervised feature selection algorithm for multi-view tasks. Since different views may complement and reinforce each other, a nonnegative weight is imposed on each view independently. The views with large weights should have large contributions to the final learning task. In this case, we can estimate the cluster label leveraging all the views. In addition, the discriminant analysis is introduced to preserve the cluster structure of the data, which is obtained by spectral clustering. We then select features via a structural sparsity regularization model. The proposed algorithm integrates discriminant analysis, spectral clustering and a sparsity regularization model into the same framework. We design a new optimization algorithm to solve it. Experimental results on public data sets demonstrate the effectiveness of the proposed algorithm.

The rest of this paper is organized as follows. In Section 2, we introduce the preliminary knowledge of this paper. In Section 3, we present the details of our method and the solution of the optimization problem. Experimental results on five datasets are shown in Section 4, and Section 5 concludes this work.

2. Preliminaries

In machine learning and pattern recognition, spectral clustering has been demonstrated as a simple and effective clustering approach. This technique has drawn many concerns. It can extract complex clustering structures [31]. We introduce spectral clustering to discover the cluster structures in this work. We introduce a robust Laplacian matrix using local regression and global alignment (LRGA) [32].

2.1. Spectral clustering

Let $\mathcal{X}=\{x_1,x_2,\ldots,x_n\}$ be the set of n samples, where $x_i\in\mathbb{R}^d$ $(1\le i\le n)$ is the feature vector of the i-th sample. The goal of clustering is to partition the samples into c clusters $\{C_j\}_{j=1}^{c}$, where c is the number of clusters and C_j is the j-th cluster. $Y=[y_1,y_2,\ldots,y_n]^T\in\{0,1\}^{n\times c}$ is the indicator matrix of the data. Each element y_{ij} in Y indicates whether $x_i\in C_j$. If $x_i\in C_j$, $y_{ij}=1$; and $y_{ij}=0$, otherwise. The indicator matrix represents the cluster structure and contains the discriminative information. In order to obtain balanced clusters, the scaled indicator matrix is usually used which has been shown to be better performance in practice [33]. Following [23,34], the scaled cluster indicator matrix F is defined as

$$F=[f_1,f_2,\ldots,f_n]^T=Y\left(Y^TY\right)^{-1/2}, \tag{1}$$