Knowledge-Based Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Consensus learning guided multi-view unsupervised feature selection☆

Chang Tang[1,a], Jiajia Chen[1,b], Xinwang Liu[c], Miaomiao Li[c], Pichao Wang[d], Minhui Wang[*,e], Peng Lu[*,f]

[a] School of Computer Science, China University of Geosciences, Wuhan, 430074, PR China
[b] Department of Pharmacy, Huai'an Second People's Hospital Affiliated to Xuzhou Medical College, Huai'an 223002, China
[c] School of Computer Science, National University of Defense Technology, Changsha 410073, PR China
[d] School of Computing and Information Technology, University of Wollongong, NSW 2500, Australia
[e] Department of Pharmacy, People's Hospital of Lian'shui County, Huai'an 223300, PR China
[f] Pharmacy Department, Huai'an Maternal and Child Health Hospital, Huai'an 223002, PR China

ARTICLE INFO

Keywords:
Multi-view unsupervised feature selection
Sparse constrain
Non-negative matrix factorization
Consensus learning

MSC:
00-01
99-00

ABSTRACT

Multi-view unsupervised feature selection has been proven to be an effective approach to reduce the dimensionality of multi-view data. One of its key issues is how to exploit the underlying common structures across different views. In this paper, we propose a consensus learning guided multi-view unsupervised feature selection method, which embeds multi-view feature selection into a non-negative matrix factorization based clustering with sparse constrain. The proposed method learns latent feature matrices from all the views, and optimizes a consensus matrix such that the difference between the cluster indicator matrix of each view and the consensus matrix is minimized. The parameters for balancing the weights of different views are automatically adjusted, and a sparse constraint is imposed on the latent feature matrices to perform feature selection. After that, we design an effective iterative algorithm to solve the resultant optimization problem. Extensive experiments have been conducted on six publicly multi-view datasets, and the results demonstrate that the proposed algorithm outperforms several other state-of-the-art single view and multi-view unsupervised feature selection methods in terms of clustering tasks, validating the effectiveness of the proposed multi-view unsupervised feature selection method. The source code of our algorithm will be available on our on-line page: http://tangchang.net/.

1. Introduction

With the explosive use of different data acquisition sensors and social media, objects in real-world are often represented by multiple heterogenous features from various representations. For instance, in image/video data processing, different visual descriptors such as such as Scale Invariant Feature Transform (SIFT) [1], Local Binary Patterns (LBP) [2], Histogram of Oriented Gradient (HOG) [3] and Fourier Shape Descriptors (FSD) [4] are often deployed to represent each image or video frame. In pharmaceutical research, a drug can be represented by its chemical structure and chemical response in different cells. A protein can be represented by its sequence and gene expression values in different cells [5,6]. Each view of feature representation characterizes the data in one specific feature space and has particular physical meaning and statistic property. Generally, this type of data is the so-called multi-view data. Compared to traditional single view learning methods which take the single representation of data as input, multi-view learning, which aims to learn from data represented by multiple distinct feature sets, can fully exploit the diversity and consistency among different representations to obtain better performance than single view methods. In multi-view learning, multiple representations are as the input and it is crucial to employ the complement information of all the heterogeneous features to create powerful models. Many multi-view learning methods have been proposed in past decade, such as feature learning [7–9], semi-supervised learning [10,11], ensemble learning [12,13], transfer learning [14,15] and active learning [16,17]. In practice, different modalities of data are usually represented in a high-dimensional feature space, this frequently leads to the curse of dimensionality problem [18–24]. In addition, obtaining the labels of data is a challenging and laborious task. Among

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

* Corresponding authors.
E-mail addresses: tangchang@cug.edu.cn (C. Tang), jjiachen@outlook.com (J. Chen), xinwangliu@nudt.edu.cn (X. Liu), miaomiaolinudt@gmail.com (M. Li), pw212@uowmail.edu.au (P. Wang), minhwang@163.com (M. Wang), penglumedical@outlook.com (P. Lu).
[1] Equal contribution as first author.

https://doi.org/10.1016/j.knosys.2018.06.016
Received 31 December 2017; Received in revised form 26 June 2018; Accepted 28 June 2018

those multi-view learning methods, multi-view unsupervised feature selection, which aims to reduce the dimensionality of high dimensional multi-view data by selecting a compact subset of representative features from all the features without data labels, has received widespread attention and has been used in many applications, e.g., visual concept recognition [25], human motion retrieval [26], web data clustering [27], image annotation [28] and action recognition [29].

In the past decades, a variety of multi-view unsupervised feature selection methods have been proposed. These methods can be mainly classified into two distinct categories. Methods in the first category combine features from multiple views into a single view and then employ traditional single-view unsupervised feature selection methods on the concentrated data [22–24,30–32]. The other category of methods process multi-view data directly for feature selection [25,26,29,33,34], these methods usually deploy the structure of data for completing this task. Though demonstrating promising performance in various applications, most of previous multi-view feature selection methods aim to learn a common similarity matrix of different views. However, due to the different feature distributions in different feature spaces, the sample similarity matrices in different feature spaces may differ. On the other hand, the sample labels of different feature views should be the same, and are not dependent with different feature distributions.

In this paper, we propose a novel consensus guided multi-view unsupervised feature selection method, referred to as CGMV-UFS briefly. Different from previous methods which try to preserve the local geometrical structure in each data view or across different views. CGMV-UFS embeds feature selection into a non-negative matrix factorization based clustering framework with sparse constraints, it tries to learn the latent feature matrices for all the views and simultaneously learn a consensus matrix to constrain that the difference between the learned cluster label matrix of each view and the consensus matrix is minimized. The weights of different views are tuned automatically, and the learned latent feature matrices with sparse constraints can serve to select discriminative features. Comprehensive experiments demonstrate that the proposed algorithm outperforms several other state-of-the-art single view and multi-view unsupervised feature selection methods in terms of clustering performance. The major contributions of this paper can be summarized as follows.

- A novel embedded multi-view unsupervised feature selection approach is proposed. The proposed approach embeds multi-view unsupervised feature selection into a non-negative matrix factorization (NMF) based clustering framework with sparse constraints. A consensus matrix is learned to constrain that the cluster indicator matrix of each view stays in step with each other;
- The parameters which balance different views are tuned automatically and we impose the sparse constraints on the learned latent feature matrices for measuring the importance of features;
- We develop an efficient alternating scheme with Lagrange multipliers to optimize the objective function and comprehensive experiments are conducted to verify the validity of our proposed method.

2. Background

In this section, we give the background of multi-view unsupervised feature selection. Firstly, we present some notions used in this paper.

Throughout this paper, matrices are written as boldface capital letters and vectors are denoted as boldface lower case letters. For an arbitrary vector $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}_i$ represents its ith element. For an arbitrary matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, $\mathbf{M}_{ij}$ denotes its (i, j)th entry, $\mathbf{m}_i$ and $\mathbf{m}^j$ denotes the ith row and jth column of $\mathbf{M}$, respectively. $Tr(\mathbf{M})$ is the trace of $\mathbf{M}$ if $\mathbf{M}$ is square and $\mathbf{M}^T$ is the transpose of $\mathbf{M}$. $\langle \mathbf{A}, \mathbf{B}\rangle$ is the standard inner product between two matrices. $\mathbf{I}_m$ is the identity matrix with size $m \times m$ (denoted by $\mathbf{I}$ if the size is obviously known). The $l_{2,1}$-norm of matrix $\mathbf{M}$ is defined as $\|\mathbf{M}\|_{2,1} = \sum_{i=1}^{m}\|\mathbf{m}_i\| = \sum_{i=1}^{m}\sqrt{\sum_{j=1}^{n}\mathbf{M}_{ij}^2}$. $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{M}_{ij}^2}$ is the well-known Frobenius norm of $\mathbf{M}$.

Supposing we have N data samples $\{\mathbf{x}_i\}_{i=1}^N$, the data matrix is denoted as $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^{d\times N}$. The ith sample $\mathbf{x}_i = [(\mathbf{x}_i^{(1)})^T, \cdots, (\mathbf{x}_i^{(V)})^T]^T \in \mathbb{R}^d$ is composed of features from V views, where the vth view $\mathbf{x}_i^{(v)} \in \mathbb{R}^{d_v}$ has d_v features such that $d = \sum_{v=1}^{V} d_v$. Denote the data matrix of the vth view as $\mathbf{X}^{(v)} = [\mathbf{x}_i^{(v)}, \cdots, \mathbf{x}_N^{(v)}] \in \mathbb{R}^{d_v\times N}$, then $\mathbf{X} = [(\mathbf{X}^{(v)})^T, \cdots, (\mathbf{X}^{(v)})^T]^T$. Unsupervised multi-view feature selection aims to select the top K discriminative features from those d features without using the sample labels.

3. Related works and motivation

In this section, we review some representative multi-view unsupervised feature selection approaches. Then we give the motivation of our work.

3.1. AMFS

Adaptive Multi-View Feature Selection (AMFS) [26] is an unsupervised feature selection approach proposed for human motion retrieval. In AMFS, local descriptors are employed to characterize local geometric structure of human motion data and there are three steps to select discriminative features. At the first, for each single view, they generate the graph-Laplacian [35] as local descriptor. Then, these view-dependent Laplacian matrices are combined together linearly with non-negative view weights to explore the complementary information between different views. Finally, AMFS uses the trace ratio criteria as in traditional Fisher Score feature selection method to discard the redundant features.

In general, AMFS formulates the objective function as follows:

$$\min_{\mathbf{W},\boldsymbol{\alpha}} \frac{Tr(\mathbf{W}^T\mathbf{X}(\sum_{v=1}^{V}\boldsymbol{\alpha}_v{}^r\mathbf{L}^{(v)})\mathbf{X}^T\mathbf{W})}{Tr(\mathbf{W}^T\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{W})}$$
$$s.\ t.\ \sum_{v=1}^{V}\boldsymbol{\alpha}_v = 1,\ \boldsymbol{\alpha}_v \ge 0,\ \mathbf{W} \in \{0, 1\}^{d\times s}, \tag{1}$$

where $\mathbf{L}^{(v)}$ is the graph Laplacian corresponding for characterizing the vth view and $\mathbf{H}$ is the centralized matrix for centralizing the data. $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_V]^T$ is a weight coefficient vector to combine all the Laplacian matrices and $r > 1$ is a parameter to avoid trivial solution. $\mathbf{W} \in \{0, 1\}^{d\times s}$ is a weight matrix in performing feature selection. Since each row of $\mathbf{W}$ has one and only one non-zero element, the features corresponding to non-zero elements will be selected.

3.2. AUMFS

Adaptive Unsupervised Multi-view Feature Selection (AUMFS) [25] is proposed for visual concept recognition, in which three kinds of information, i.e., data cluster structure, data similarity and the correlations between different views are used for feature selection. In order to select discriminative features, it employs a robust sparse regression model with the $l_{2,1}$-norm regularization to predict the cluster labels of data. Similar to AMFS, the local structure is also preserved by linear combining the graph Laplacian matrices of different views. In addition, AUMFS adds a $l_{2,1}$-norm to regularize the transformation matrices for each view. Since the $l_{2,1}$-norm can induce row sparsity of a matrix, it can be used for feature selection. In fact, AUMFS can be regarded as an extension of traditional spectral regression [36], which is a dimensionality reduction framework. Different to spectral regression, the robust regression and sparse regularizer are combined in AUMFS, and joint optimization is accomplished for multiple view data. When the optimal transformation matrix $\mathbf{W}$ is obtained, AUMFS employ the l_2-norm of each row of $\mathbf{W}$ to rank the importance of each feature. The larger this value is, the more important the corresponding feature is. The objective function of AUMFS is described as follows: