Contents lists available at ScienceDirect

Signal Processing

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

Multiple graph unsupervised feature selection

Xingzhong Du [a], Yan Yan [a,*], Pingbo Pan [b], Guodong Long [c], Lei Zhao [d]

[a] School of Information Technology and Electrical Engineering, The University of Queensland, QLD 4072, Australia
[b] Zhejiang University, Hangzhou 310027, China
[c] Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney, Ultimo NSW 2007, Australia
[d] Advanced Data Analytics Research Center, Soochow University, Soochow 215006, China

ARTICLE INFO

Article history:
Received 14 October 2014
Received in revised form
19 December 2014
Accepted 22 December 2014

Keywords:
Multiple graph
$l_{2,p}$-norm
Unsupervised learning
Feature selection

ABSTRACT

Feature selection improves the quality of the model by filtering out the noisy or redundant part. In the unsupervised scenarios, the selection is challenging due to the unavailability of the labels. To overcome that, the graphs which can unfold the geometry structure on the manifold are usually used to regularize the selection process. These graphs can be constructed either in the local view or the global view. As the local graph is more discriminative, previous methods tended to use the local graph rather than the global graph. But the global graph also has useful information. In light of this, in this paper, we propose a multiple graph unsupervised feature selection method to leverage the information from both local and global graphs. Besides that, we enforce the $l_{2,p}$ norm to achieve more flexible sparse learning. The experiments which inspect the effects of multiple graph and $l_{2,p}$ norm are conducted respectively on various datasets, and the comparisons to other mainstream methods are also presented in this paper. The results support that the multiple graph could be better than the single graph in the unsupervised feature selection, and the overall performance of the proposed method is higher than the other comparisons.

1. Introduction

Feature selection has been a hot topic recently in many research areas where the data are represented by the high dimensional features, such as computer vision [1], information retrieval [2], biology informatics [3] and multimedia [4,5]. Due to that the original features may have irrelevant or noisy dimensions for the specific tasks, the key idea is to map all the features into a low dimensional representation and keep the original informativeness and interpretability [6]. To achieve this, the majority select the most discriminative features from the original ones. Compared to PCA which projects the data into a low dimensional space, the dimension in feature selection can be reduced directly by indexing when the discriminative part is identified. With these selected features, the efficiency of the model training and the performance of the testing are improved [7–9].

In general, the feature selection methods can be grouped into two types based on the availability of label information, namely supervised methods and unsupervised methods [10,11]. The supervised methods select the discriminative features according to the correlations between the features and the labels, while the unsupervised methods only leverage the information derived from the data themselves due to the unavailability of the labels [11,12]. Among these, the graphs which can unfold the geometry structure in the original space have become an indispensable component in recent works [13–17]. This is based on the assumption that all the data points locate on a manifold in a high dimensional space. According to the definition of manifold, the tangent plane on each datum point which reveals the local

[*] Corresponding author.
E-mail address: y.yan3@uq.edu.au (Y. Yan).

http://dx.doi.org/10.1016/j.sigpro.2014.12.027
0165-1684/© 2014 Elsevier B.V. All rights reserved.

geometry structure is stable when the manifold is unfolded. Therefore, many graphs are constructed to depict the local geometry structure, and the most widely used construction methods are k nearest neighbors and ϵ-ball [18]. It is also feasible to construct the graphs to depict the global geometry structures which reflect the correlations between any data points. However, in most recent works [19,13–17], the features selected according to the local geometry structures can always outperform the ones selected according to the global geometry structures.

However, there exist some flaws in unfolding manifold with single local graph. First of all, according to the properties of the manifold, the topological information is incomplete when we only use one graph. Secondly, the capacity of discrimination from local graph is high only if all points in the original space can be locally reconstructed by its neighbors [18], which can not be fully ensured in real world. Therefore, we propose to leverage the information from the global geometry based on the assumption that the local geometry is not sufficient to reconstruct the space in the low dimensional space. That is, when we unfold the manifold by the graph, we combine the local and global graphs together to get a more comprehensive understanding of the original space. As a result, in this paper, we propose a unsupervised feature selection framework (MGFS) to regularize the objective function based on multiple graph. In the proposed framework, the local and global graphs are combined, so the solution will take the advantages of both sides, which results in better performance. Besides that, since a typical unsupervised feature method usually consists of graph regularization and sparse learning [19], the $\ell_{2,p}$ norm is added into the objective function to control the sparsity of the feature selection matrix. Compared to the $\ell_{2,1}$ norm which has been extensively used in previous feature selection algorithms, the $\ell_{2,p}$ norm has more flexibility in controlling the sparsity of the feature selection matrix, which is another improvement to existing unsupervised feature selection methods.

Following [13–15], we evaluate MGFS's performance by the clustering results based on the selected features. We first fix the p as 1 which is equivalent to the $\ell_{2,1}$ norm, and adjust the weights on local and global graphs. Under these settings, we compare the performance changes caused by the multiple graph. Then, we fix the settings on the multiple graph, and test whether $\ell_{2,p}$ norm is more flexible than the $\ell_{2,1}$ norm. Finally, we compare MGFS with other unsupervised feature selection methods. The results show that multiple graph can is better than single graph, and the $\ell_{2,p}$ norm is better than the $\ell_{2,1}$ norm. Furthermore, MGFS cis superior to the comparison feature selection methods.

2. Related work

In the recent decade, manifold learning has gradually attracted much research attention in the field [20]. It has been experimentally demonstrated that exploiting the manifold structure would be beneficial for a variety of applications, such as multimedia retrieval [21], person identification in video streams [22], image classification [23], visual concept recognition [24], photo cropping [25,26], action recognition [27,28] and aerial image categorization [29,30] and segmentation [31]. In the meantime, a few studies on feature selection have also used the local structure in semi-supervised or unsupervised setting, e.g., [32–34,13,35]. In these works, the manifold structure is usually represented by a graph, which is then transformed to a Laplacian matrix. While there are a few different ways to exploit the local geometry, most of the existing papers only use a single graph. In [36], Han et al. proposed to construct a graph according to the label information, which is then combined with the graph obtained by local spline regression. The study in [36] has shown that it would be better if we utilize multiple graph for feature selection. However, in the unsupervised learning, there is no label information to construct the discriminative graph for feature selection. Instead, we propose to adopt a clustering based approach to obtain the discriminate graph [37] for feature selection in unsupervised learning. The discriminative graph obtained via unsupervised approach is then combined with the graph learned by local spline regression. The two graphs provide different views of the local geometry and the combination would improve the performance of feature selection.

3. Proposed framework

3.1. Preliminaries

In this paper, given a matrix A, the element which locates in the ith row and jth column is denoted as a_{ij}. The ith column vector of A is represented as a_i and the jth row vector of A is represented as a^j. We use function $diag(A)$ to calculate a diagonal matrix D where $d_{ii} = \sum_i^n a_{ij}$. Following [38,39], the $l_{2,p}$ norm of a matrix $A \in \mathbb{R}^{u \times v}$ is defined as

$$\|A\|_{2,p} = \left(\sum_{i=1}^{u} \left(\sum_{j=1}^{v} a_{ij}^2 \right)^{p/2} \right)^{1/p}. \tag{1}$$

A datum is represented by a column vector $x_i \in \mathbb{R}^{d \times 1}$ where d is its dimension. All the data together form a data matrix $X = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}^{d \times n}$, in which n is the total number of the data. I_d denotes an identity matrix of dimension d. For a constant m, $\mathbf{1}_m$ is a column vector whose dimension is m and all elements are 1. Then, $H_m \in \mathbb{R}^{m \times m}$ is defined as $I_m - (1/m)\mathbf{1}_m\mathbf{1}_m^T$. Noticing that $H_m\mathbf{1}_m = \mathbf{0}_m$ and $\mathbf{1}_m^T H_m = \mathbf{0}_m^T$. Besides that, suppose that the training data are from c classes, we define $y_i \in \{0,1\}^{c \times 1} (1 \le i \le n)$ as the label vector of x_i. If the jth $(1 \le j \le c)$ element of y_i is 1, it indicates that x_i is belonging to the jth class. Similar to X, we construct $Y \in \{0,1\}^{n \times c}$ by putting the label vectors together. Based on the definition of Y, we further construct the scaled label matrix $G = \{G_1, G_2, \dots, G_n\}^T \in \mathbb{R}^{n \times c}$ as follows:

$$G = Y(Y^TY)^{-1/2}$$

Noticing that, given the expression above, $G^TG = I_c$. Assuming that n_i samples are belonging to the ith class, the total scatter matrix S_t and between class scatter matrix then are defined as follows [40]:

$$S_t = \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T = \tilde{X}\tilde{X}^T$$