Image and Vision Computing 63 (2017) 10–23

Contents lists available at ScienceDirect

Image and Vision Computing

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

Semi-supervised multi-label feature selection via label correlation analysis with l_1-norm graph embedding☆

Xiao-dong Wang[a,b,*], Rung-Ching Chen[b,**], Chao-qun Hong[a], Zhi-qiang Zeng[a], Zhi-li Zhou[c]

[a] College of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China
[b] Department of Information Management, Chaoyang University of Technology, Taichung, Taiwan
[c] Jiangsu Engineering Center of Network Monitoring & School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China

ARTICLE INFO

Article history:
Received 28 October 2016
Received in revised form 3 March 2017
Accepted 13 May 2017
Available online 22 May 2017

Keywords:
Semi-supervised learning
Feature selection
Multi-label learning
Shared-subspace learning

ABSTRACT

In this paper, we propose a novel semi-supervised multi-label feature selection algorithm and apply it to three different applications: natural scene classification, web page annotation, and yeast gene functional classification. Compared with the previous works, there are two advantages of our algorithm: (1) Manifold learning which leverages the underlying geometric structure of the training data is imposed to utilize both labeled and unlabeled data. Besides, the underlying manifold structure is guaranteed to be clear by using the l_1-norm regularization. (2) Shared subspace learning which has shown its efficiency in multi-label learning scenarios, is also considered in our feature learning algorithm. The proposed objective function involves $l_{2,1}$-norm and l_1-norm, making it non-smooth and difficult to solve. We also design an efficient iterative algorithm to optimize it. Experimental results demonstrate the effectiveness of our algorithm compared with sate-of-the-art algorithms on different tasks.

1. Introduction

High dimensional data, which is frequently confronted in real world applications nowadays, has presented a great challenge to existing technologies [1,2]. It's hard to directly deal with this high dimensional data for its huge time and space cost [3,4]. Besides, the noises and relevant features within the data also put an obstacle to the pattern recognition tasks based on these data. According to the previous works [5], there are mainly two popular methods to tackle this problem, e.g. feature extraction and feature selection. Feature extraction aims to translate the high dimensional data into a low dimensional space through a translation matrix. In contrast, feature selection selects the most representative features from the original feature data, which could preserve the underlying meaning of the original data. However, it may be difficult for feature extraction to understand the relationship between the new translated feature data and the original data. Under this circumstance, feature selection has gained many researcher's interests in recent years.

According to whether there are labels or not, existing feature selection algorithms can be roughly categorized into three groups: supervised, unsupervised and semi-supervised algorithms. Generally speaking, supervised feature selection is widely used in many research areas. Given sufficient label information, supervised feature selection usually can utilize the discriminative information provided by labels and achieve high accuracy. However, the most significant problem of the supervised feature selection is collecting large number of labeled data which is not practical and requires especially expensive human labor. Besides, the supervised learning methods often suffer from over-fitting due to the small number of labeled data [6] . Unsupervised feature selection is designed to handle unlabeled data [7,8]. Unsupervised feature selection [9,10] selects features using the underlying correlation information among the data points, e.g., data variance and adjacent relation. Data variance [9] is a commonly used unsupervised feature selection, which evaluates the features by the variance along a dimension, and the features with maximum variance will be selected. Laplacian score [10] extends data variance by choosing features with the largest variance and considering the local structure of the data space. However, these methods neglect the inter-dependency between features when multiple features need to be selected. Recent years, a popular used criterion is to select the representative features by imposing some kinds of sparse constraints on the whole feature set [11]. Nevertheless, in unsupervised scenarios, there is no label information

☆ This paper has been recommended for acceptance by Matthew Turk.

* Correspondence to: X.-d. Wang, College of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China.
** Corresponding author.
E-mail addresses: xdwangjsj@xmut.edu.cn (X.-d. Wang), crching@cyut.edu.tw (R.-C. Chen).

http://dx.doi.org/10.1016/j.imavis.2017.05.004
0262-8856/

available, making it difficult to select discriminative features. When there are few labels available, semi-supervised feature selection is adapted, semi-supervised feature selection can be interpreted as a compromise between supervised and unsupervised feature selection. It is reported to be one of the best methods to cope with the "small-labeled-sample" problem [12].

The key issue of semi-supervised feature selection is to design a framework, where the labeled and unlabeled data can be sufficiently utilized to boost the learning performance [12–14]. In the past two decades, many semi-supervised feature selection algorithms have been proposed. Among the different methods, graph Laplacian based semi-supervised leaning has gained the most research interests [15–18]. A Flexible Manifold Embedding framework based on graph Laplacian was proposed in [15], and was demonstrated significant improvement over state-of-the-art semi-supervised algorithms. Liu and Nie et al. [16] proposed a graph based semi-supervised feature selection, where a new noise insensitive trace ratio criterion is implemented. In order to utilize the geometric information embedded in the original data, Yang et al. [19] proposed a two-level manifold learning method, in which several independent graphs for multimedia data were firstly constructed and were then combined to obtain the final data representation. However, the two-level manifold learning method is complex and impractical in the real world applications. To tackle this problem, Yang et al. [17] proposed a manifold semi-supervised multimedia retrieval framework, which ranking each sample with local regression and global alignment. Ma et al. [18]proposed a semi-supervised multimedia feature analysis algorithm, which makes use of the labels information and captures underlying manifold as well. However, all of these graph based algorithms neglect the clearness of the underlying structure, which has been shown its effectiveness in previous works [20–22].

In real-world classification applications, the samples we dealing with often belong to several classes simultaneously. For example, in image annotation, image for natural scene usually contains many subjects, such as "trees", "lands", and "sea". A simple way to cope with the multi-label problem is to translate it into several single-label problems. However, this approach neglects the correlations among multiple labels, which have been proven its effectiveness in many works [23–26]. For instance, the concept "trees" may be often accompanied with the concept "mountains", such relationship between these concepts is potentially helpful to improve the classification performance. In [23], a framework for multi-label classification was built, in which a shared subspace was considered among multiple labels. Ma et al. [24] proposed a feature selection by uncovering the shared feature subspace among multiple classes. These methods have been demonstrated their validity on several large-scale multimedia datasets. Nevertheless, both of these methods are designed in a supervised approach. Yang et al. [25] proposed a semi-supervised method for image annotation that mines label correlation and visual similarity jointly. Similar to [25], Chang et al. proposed to utilize the label correlations by imposing a new group sparse constraint [26].

In this paper, we propose a semi-supervised multi-label feature selection algorithm and apply it to several kinds of applications. The proposed algorithm firstly estimates the feature relevance by constructing a graph-based semi-supervised learning framework, which can utilize the labeled and unlabeled data concurrently. A l_1-norm based constraint is then imposed to ensure the underlying manifold structure is clear. To the end, the correlations between different labels are also taken into consideration.

The most important contributions of our method are as follows

1. The proposed algorithm combines semi-supervised feature learning, multi-label feature learning and shared subspace learning jointly into a single framework, in which both labeled and unlabeled samples are utilized and the sharing information among multiple labels is leverage considered simultaneously.
2. The construction process of graph matrix is constrained by the l_1-norm, which is easy to capture a clear underlying manifold structure, and thus achieves a better feature selection performance.
3. The proposed objective function is non-smooth and is difficult to solve. We propose an efficient iterative algorithm to optimize it. We apply our algorithm to several kinds of real-world datasets, and experimental results demonstrate the effectiveness of our algorithm.

The rest of this paper is organized as follows. We first give a brief review of the previous works in Section 2. In Section 3 and Section 4, we derive and optimize our semi-supervised multi-label feature selection model. In Section 5, we present our experimental evaluation on the proposed method, followed by the conclusion in Section 6.

2. Related works

In this section, we briefly review the related works. This paper is closely related to l_1-norm graph learning and shared subspace learning.

2.1. l_1-norm graph learning

Graph-based learning has shown its efficiency for modeling data in clustering or classification problems for its ability to incorporate geometrical information between unlabeled data. In the past two decades, several graph based learning methods were proposed [27,28]. Among the different algorithms, spectral clustering has shown its excellent performance over the state-of-the-art algorithms.

The main tools for spectral clustering are graph Laplacian matrix. Given n data points $\{x_1,\ldots,x_n\} \in R^{d\times n}$, denote $G = (V,A)$ as an undirected weighted graph with a vertex set V and weight matrix $A = [A_1,A_2,\ldots,A_n] \in R^{n\times n}$, where $A_{ij} = A_{ji} \geq 0$. There are several popular methods to construct the weight matrix A. A common choice of A is defined by

$$A_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right), \text{if } x_i \text{ and } x_j \text{ are } k \text{ nearest neighbors} \\ 0, \text{otherwise} \end{cases} \tag{1}$$

where the parameter σ is used to control the spread of neighbors. The graph Laplacian matrix is defined as $L = D - A$, where D is a diagonal matrix with each diagonal element as $D_{ii} = \sum_{j=1}^{n} A_{ij}, \forall i$. The intuition of clustering is to separate the data points into several different connected components. A traditional method is to translate this problem into a Ratio cut problem, and then relax it into following maximization problem

$$\min_{Q^TQ=I} Tr(Q^TLQ) \tag{2}$$

where $Q = [q^1,q^2,\ldots,q^c] \in R^{n\times c}$ is the cluster indicator matrix, $q^k \in R^{n\times 1}$ is the k-th column of Q. The solution Q of the relaxed problem is formed as continuous value, which can not be directly used as clustering results. Hence, further clustering algorithm such as K-means need to be imposed to obtain the final clustering indicators. However, K-means suffers from the local optimum problem, which making the performance of clustering unstable. In order to fix this problem, Nie et al. [20] proposed a l_1-norm based learning model to get a clearer manifold structure. Furthermore, they also extended it with a fixed rank constraint and proposed a parameter-free learning