Neurocomputing 173 (2016) 102–109

Contents lists available at ScienceDirect

Neurocomputing

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

Semi-supervised feature selection based on local discriminative information ☆

Zhiqiang Zeng [a], Xiaodong Wang [a], Jian Zhang [b], Qun Wu [c,*]

[a] College of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China
[b] School of Science and Technology, Zhejiang International Studies University, Hangzhou 310012, China
[c] School of Art and Design, Zhejiang Sci-Tech University, Hangzhou 310018, China

ARTICLE INFO

Article history:
Received 19 March 2015
Received in revised form
18 May 2015
Accepted 24 May 2015
Available online 14 August 2015

Keywords:
Feature selection
Semi-supervised learning
Local discriminative
Information
$l_{2,1}$ norm
Linear regression

ABSTRACT

Feature selection has been an effective way to reduce the dimensionality of the high dimensional data. In this paper, we propose a novel feature selection method which achieves batch feature selection using both supervised and unsupervised data samples. The objective function includes three parts: first, under the assumption that each data sample has been assigned a class label, the ratio of between class scatter matrix and total scatter matrix should be minimized, where the scatter matrices are formed by the selected features of these data samples; second, we use linear regression to model the correlations between the data samples with supervision information and their class labels; last, we use $l_{2,1}$-norm to guarantee the sparsity of the feature selection matrix and exploit the sharing information between supervised and unsupervised data samples jointly. Different from existing methods, our approach exploits local discriminative information to construct the model, therefore we obtain better results from extensive experiments compared with the existing methods.

1. Introduction

Recent years have witnessed a surge of multimedia data with very high dimensionality, such as gene expression data, video surveillance data, and meteorological data. This puts an obstacle to the pattern recognition tasks based on these data. Under this circumstance, dimensionality reduction has become a research hotpot in the field of pattern recognition. Feature selection is an effective way to achieve dimensionality reduction.

Current feature selection methods can be roughly categorized into two classes, i.e., supervised methods, unsupervised methods. Fisher criterion [1–3] is one of the most important supervised methods. The fisher method computes the inter-class distance and inner-class distance with respect to each feature, and chooses the feature with maximum inter-class distance and minimum inner-class distance. However, it ignores the inter-dependency between features when multiple features need to be selected simultaneously. Therefore the selected feature subset cannot necessarily guarantee optimized objective value. Recently, some researchers aim to enhance fisher criterion by incorporating feature–feature dependency in feature selection [4,5]. In unsupervised scenario, class label information is unavailable directly, which makes the task of feature selection more challenging. A commonly used criterion in unsupervised feature learning is to select features best preserving data similarity or manifold structure constructed from the whole feature space [6–8], but they fail to incorporate discriminative information implied within data, though it has been shown to be important in data analysis [9,10].

The supervised methods are developed based on a number of data samples with known class label information, therefore provide higher accuracy rate when used for pattern recognition [11]. But labeling a large number of data samples needs extensive expertise and sometimes is not practical. On the contrary, unsupervised methods do not need any labeling information, but they have to assume the data set is in accordance with some specific distribution, e.g. manifold structure, and try to preserve this distribution on the whole data set. So the effect of unsupervised method is usually not very good. To this end, some researchers propose to use both supervised and unsupervised methods for feature selection, and this is the so-called semi-supervised method. Semi-supervised methods take advantage of the data samples with known class labels to provide prior knowledge for most data samples without class labels. Therefore it is a good compromise between supervised and unsupervised methods [12–14].

The key for designing an effective semi-supervised feature selection algorithm is to develop a framework, under which the

☆This paper is supported by the National Natural Science Foundation of China (No.61273290, No.61303143), Xiamen Science and Technology Planning Project (No. 3502Z20143030), Scientific Research Fund of Fujian Provincial Education Department (No. JA15385), and Scientific Research Fund of Zhejiang Provincial Education Department (No.Y201326609).

* Corresponding author.
E-mail address: wuq@zstu.edu.cn (Q. Wu).

http://dx.doi.org/10.1016/j.neucom.2015.05.119
0925-2312/

relevance of a feature can be evaluated by both labeled and unlabeled data in a natural way [15]. Feature ranking is a well studied semi-supervised feature selection method which ranks all features with respect to their relevance and chooses the top ranked features as the working feature vector [16]. Laplacian score improves the efficiency of feature selection by considering the underlying manifold structure [6]. However, without label information the Laplacian score can hardly select discriminative features. Based on Laplacian score, a semi-supervised feature selection method was proposed in [17] ,but this method also selects the most discriminating features individually and neglects the correlations among features. In order to identify relevant features, several spectral regression based methods have been proposed [18–20], i.e., xu et al. [18] and zhao et al.[19] proposed two semi-supervised algorithm based on spectral analysis, these proposed methods try to discover both geometrical and discriminant structure in the data. In [20], an embedded model has been proposed to minimize the feature Redundancy for Spectral Feature selection (MRSF). Although the spectral based methods can achieve good performance in many applications, nevertheless, they must construct a cluster indicator first in the original data with high dimension, which is sensible to the noise. In [21], a semi-supervised method based on graph Laplacian has been proposed for leveraging manifold regularization. Yang et al. [22] proposed a one-step feature selection method, which aims to select the most discriminative features for data representation, where manifold structure are also considered and feature correlations are evaluated by the $l_{2,1}$-norm regularization, which was first proposed in [23]. Similarly to [22], the method proposed in [24] has gotten more accuracy by imposing a non-negative constraint. However, without the label information, these methods will deteriorate the feature selection performance.

Recent works have shown that there exists some common components in multiple training resources and it is beneficial to leverage such shared information for multimedia analysis applications, such as event detection [25,26], feature selection [27,28]. In [25,26], two complicated event detecting algorithms have been proposed, in which several kinds of training resources are used and the sharing information between these training resources are also explored to improve the detection performance. In order to achieve a good feature selection performance, Yang et al. [27] and Wang et al. [28] proposed to utilize the sharing information among related tasks in their multi-task feature selection framework. However, all these feature selection methods mentioned above are designed in a supervised way, these methods will fail, if the size of training samples is too small. Intuitively, there should exist some sharing information between labeled and unlabeled data collected from the same resource. It will be advantageous to improve the performance of feature selection if such sharing information is exploited.

The above observations motivate us to design a feature selection method, which could select most discriminative features by utilizing the feature correlations and the sharing information between labeled and unlabeled data. Inspired by [22] and [27], in this paper, we propose a novel semi-supervised feature selection method through a multi-task way, i.e., the proposed method contains two tasks, the supervised part task and the unsupervised part task. For the unsupervised part task, like [22], we consider the most discriminative information by exploiting the underlying manifold structure and feature correlations. For the supervised part task we use the label information to guide the training process. To this end, the supervised task and unsupervised task are combined in a framework where the sharing information between them is exploited. In addition, we adopt the $l_{2,1}$-norm to ensure the sparsity of the feature selection matrix and reduce the impact of the outliers. It is worthwhile to highlight several aspects of the proposed approach:

1. We define a discriminative model for both supervised and unsupervised data, with an assumption that each unsupervised and supervised data contain some sharing information.
2. Local discriminative information, instead of global discriminative information, is used when building the model through fisher criterion and linear regression. This is because it has been observed that local information is more important than global one in pattern recognition.
3. The $l_{2,1}$-norm is exploited to regularize the feature selection matrix. The regularization ensures the sparsity of the feature selection matrix, more importantly, the sparsity is achieved by learning in unsupervised and supervised samples jointly.
4. The proposed object function is non-smooth and hard to resolve, we propose an efficient iterative approach to optimize the model.

The rest part of this paper is organized as follows. In Section 2, we introduce the related works. The proposed method is described in detail in Section 3. Section 4 demonstrates the experimental results and Section 5 concludes this paper.

2. Preliminary knowledge

2.1. $l_{2,1}$-norm

For an arbitrary matrix $D \in R^{r\times p}$, we denote its $l_{2,1}$-norm as

$$\|D\|_{2,1} = \sum_{i=1}^{r} \sqrt{\sum_{j=1}^{p} D_{ij}^2} \tag{1}$$

The $l_{2,1}$-norm ensures the sparsity of the matrix, which is an important virtue for feature selection [23]. In addition, the sparsity of D is computed with some of its rows shrunk to zero, in this respect, the $l_{2,1}$-norm will contribute to selecting discriminative features.

2.2. Unsupervised discriminative feature selection

Denote $X_u = \{x_1^u, x_2^u, \ldots, x_n^u\}$ as the unlabeled training data sample set where $x_i^u \in R^d (1 \le i \le n)$ is the i-th datum and n is the number of data samples.

Suppose the unlabeled data are sampled from c classes, with n_i data in the i-th class. For better description, we give each unlabeled data sample x_i^u a fictitious class label $y_i^u \in \{0,1\}^{c\times 1} (1 \le i \le n)$. The j-th element of y_i^u equals 1 when x_i^u belongs to the j-th class. We denote the label matrix as $Y_u = \{y_1^u, \ldots, y_n^u\}^T \in \{0,1\}^{n\times c}$. The total scatter matrix S_t and the between class scatter matrix S_b of the unlabeled data are represented as below:

$$S_t = \sum_{i=1}^{n} (\mu - x_i^u)(\mu - x_i^u)^T = \tilde{X_u}\tilde{X_u}^T$$
$$S_b = \sum_{i=1}^{c} n_i (\mu - \mu_i)(\mu - \mu_i)^T = \tilde{X_u} G G^T \tilde{X_u}^T \tag{2}$$

where μ is the average value of all the unlabeled samples, and μ_i is the average value of samples in class i, $\tilde{X_u} = X_u H_n$ is the centered data matrix with $H_n = I_n - (1/n)\mathbf{1}_n\mathbf{1}_n^T \in R^{n\times n}$, where $\mathbf{1}_n \in R^{n\times 1}$ is a unary vector with all its elements equals 1 and $I_n \in R^{n\times n}$ is the identity matrix. $G = [G_1, \ldots, G_n]^T = Y_u (Y_u^T Y_u)^{-1/2}$ is the scaled label matrix. A commonly used criterion to encode the discriminative information is to select a number of features which can make S_b reach its maximum while S_t reach its minimum. There are several criterions to achieve that goal [9], such as $\mathrm{tr}(S_t^{-1}S_b)$ and $\mathrm{tr}(S_b)/\mathrm{tr}(S_t)$. Fisher score [5] is one of the most popular methods for feature