Neurocomputing 93 (2012) 115–125

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Graph embedding based feature selection

Dan Wei [a], Shutao Li [a,*], Mingkui Tan [b]

[a] College of Electrical and Information Engineering, Hunan University, Changsha 410082, China
[b] School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore

ARTICLE INFO

Article history:
Received 8 September 2011
Received in revised form
31 December 2011
Accepted 29 March 2012
Communicated by J. Kwok
Available online 17 May 2012

Keywords:
Feature selection
Graph embedding
Recursive feature elimination
Manifold learning
Gene selection

ABSTRACT

Usually many real datasets in pattern recognition applications contain a large quantity of noisy and redundant features that are irrelevant to the intrinsic characteristics of the dataset. The irrelevant features may seriously deteriorate the learning performance. Hence feature selection which aims to select the most informative features from the original dataset plays an important role in data mining, image recognition and microarray data analysis. In this paper, we developed a new feature selection technique based on the recently developed graph embedding framework for manifold learning. We first show that the recently developed feature scores such as Linear Discriminant Analysis score and Marginal Fisher Analysis score can be seen as a direct application of the graph preserving criterion. And then, we investigate the negative influence brought by the large noise features and propose two recursive feature elimination (RFE) methods based on feature score and subset level score, respectively, for identifying the optimal feature subset. The experimental results both on toy dataset and real-world dataset verify the effectiveness and efficiency of the proposed methods.

1. Introduction

Many real datasets such as images and microarray data are represented as very high dimensional vectors which bring great challenge in data mining and further processing [1–3]. High dimensionality not only increases the learning cost, but also deteriorates the learning performance, known as the problem of "Curse of dimensionality" [4]. Hence dimensionality reduction has attracted great attentions in pattern recognition and machine learning applications such as computer vision and microarray data analysis. Generally speaking, there are mainly two kinds of dimension reduction techniques, i.e. feature extraction [5,6] and feature selection [7,8], to tackle with the "Curse of dimensionality". Feature extraction refers to the techniques that map the high dimension data (linearly or nonlinearly) to the lower dimensional subspace under some constraints. And feature selection refers to selecting the most informative features from the original dataset. Feature selection has received great attentions and is being widely used in recent years. One typical application of feature selection is the gene selection in the microarray data analysis. In general, the original microarray data contains thousands of genes (most of them are proved to be redundant) with a small number of samples, which causes the small sample size problem [6] and raises the difficulties in diagnosis. Hence, selecting high discriminative genes (or features) from the rude gene expression data can improve the performance of cancer classification and cut down the cost of medical diagnosis.

Many feature selection methods have been proposed in recent years. These methods can typically be categorized into two groups: wrapper method [9,10] and filter method [11–14]. The wrapper method selects the discriminative features dependently on the classifier used. The wrapper method can be expected to be of high performance, but it is difficult to scale to large datasets owing to the expensive computation cost. The wrapper methods, such as SVM-RFE can be expected of good performance in identifying optimal feature subset [9]. However, they are computationally more expensive compared with filter methods and lack of good generalization capability over classifiers [14]. What's more, if the classifier is not well trained, the performance of the wrapper methods may decline.

The filter method refers to selecting informative features according to their discriminative power without considering any knowledge of the classifier. The filter method possesses the advantages of high speed and capability of dealing with large datasets, but lack of abilities to find the optimal feature subset. Typical filter methods includes T-statistics [12], signal-to-noise ratio method [2] and Fisher score [13]. These methods have shown good performance on linear feature selection but poor performance on nonlinear feature identification owing to that they cannot reveal the mutual information among features. To solve this problem, some new feature scores have been proposed

* Corresponding author.
E-mail addresses: weiweidandan@163.com (D. Wei), shutao_li@yahoo.com.cn (S. Li), tanm0097@ntu.edu.sg (M. Tan).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.03.016

recently based on the graph constructed on the samples, such as Locality Sensitive Discriminant Feature (LSDF) score [1] and Laplacian score [15]. Recently, Nie et al. proposed a subset level (SL) score based method identifying the optimal feature. The SL method can be viewed as a special filter method but shows much better performance than traditional filter methods [14]. By exploring the intrinsic structure of the dataset, we can possibly find more informative features [1,14,15]. Particularly, via the intrinsic graph, some features with complex nonlinear structures can be identified, which is a hard problem for linear feature selection methods such as SVM-RFE. However, their performance may be declined as the noise features increase. Note that, in the traditional graph based feature selection methods, the graph is pre-computed with all features, including both informative and noninformative features. When doing feature selection, one assumes that only a small part of features are informative. Under this scenario, one can hardly build a stable graph when there are relatively large number of noise features. Correspondingly, the performance of the feature selection can no longer be guaranteed. An empirical study of this issue will be presented in Section 3.1.

Regarding the above ambiguity in graph based feature selection, in this paper, we assume that we can obtain a reasonable graph which can relatively describe the relationship among patterns with given features. Considering that with large number of features, the graph can be contaminated by the noise features, we start from all features and recursively build the graph with the remaining features and then remove the non-informative features with respect to the current graph. With this recursive strategy, we proposed two new feature selection methods, namely the feature score based recursive feature elimination method (FS-RFE) and the subset level score based recursive feature elimination method (SL-RFE). Although they are still local, the proposed methods can be expected to have better performance. In summary, the contributions of this paper are: (I) We reveal that the traditional graph based feature selection methods are sensitive to large noises. (II) To avoid the negative influence brought by the noise features to the graph, we proposed an FS-RFE method and an SL-RFE method for identifying the optimal feature subsets. The experimental results verified the performances.

The rest of this paper is organized as follows. A short introduction to the graph embedding framework is given in Section 2. In Section 3, we present a feature score recursive feature elimination method (FS-RFE) and a subset level score recursive feature elimination method (SL-RFE) for feature selection. The experimental results are presented in Section 4. The conclusions are finally discussed in Section 5.

2. Prior knowledge: graph embedding

For a general learning problem, let $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x_n}]$ denote the dataset and $\mathbf{x}_n\in\mathbf{R}^m$ is a sample with m dimensions. The dataset can also be written as $\mathbf{X}=[\mathbf{f}_1,\mathbf{f}_2,\ldots,\mathbf{f}_m]^T$, where $\mathbf{f}_i\in\mathbf{R}^n$ $(i=1,2,\ldots,m)$ are the feature vectors. In supervised learning tasks, a sample $\mathbf{x}_n$ is labeled by class label $c_i\in\{1,2,\ldots,n_c\}$, where n_c is the number of classes. Generally, the dimension m is always very large which increases the difficulties of learning. Yan et al. present a novel unifying framework, named graph embedding, to formulate various feature extraction methods and provide new perspective in designing new methods [6]. In graph embedding framework, an intrinsic graph G and a penalty graph G^p are adopted. Graph $G=\{\mathbf{X},\mathbf{S}\}$ and $G^p=\{\mathbf{X},\mathbf{S}^p\}$ are two undirected weighted graphs with similarity matrix $\mathbf{S}$ and $\mathbf{S}^p$ that can be the adjacency matrix or similarity matrix, depending on different applications. Let $\mathbf{L}=\mathbf{D}-\mathbf{S}$ be the Laplacian matrix of graph G, where $\mathbf{D}$ is a diagonal matrix with entries $D_{ii}=\sum_{i\neq j}S_{ij}$. Similarly we can get the Laplacian matrix $\mathbf{L}^p$ of G^p. The intrinsic graph G denotes the similarity characteristics to be strengthened while the intrinsic graph G^p refers to the similarity characteristics to be suppressed. Simply suppose we project to a one dimensional line, then the graph-preserving criterion of the graph embedding framework is formulated as follows:

$$\mathbf{y}=\arg\min_{\mathbf{y}^T\mathbf{B}\mathbf{y}=\Delta}\sum_{i\neq j}\|\mathbf{y}_i-\mathbf{y}_j\|^2S_{ij}, \tag{1}$$

where $\mathbf{y}$ is the lower dimensional representation of $\mathbf{X}$ with $\mathbf{y}=[y_1,y_2,\ldots,y_n]^T$. The above projection often appears in classification, where the data is projected to a direction that is perpendicular to the separating hyperplane [16]. By simple algebra calculation, we can get a simpler form with matrix formulations

$$\mathbf{y}=\arg\min_{\mathbf{y}^T\mathbf{B}\mathbf{y}=1}\mathbf{y}^T\mathbf{L}\mathbf{y}, \tag{2}$$

where matrix $\mathbf{B}$ can be the identity matrix $\mathbf{I}$ or the Laplacian matrix of the penalty graph G^p, that is $\mathbf{B}=\mathbf{L}^p$. The constrained minimization problem (1) and (2) can be interpreted as two aspects: on the one hand, for those vertices near to each other, we would like to make them be near in their lower representations, which can be realized by minimizing the objective function of (1) or (2); on the other hand, for those vertices far from each other, we would make them apart as far as possible, which can be realized by maximizing $\mathbf{y}^T\mathbf{B}\mathbf{y}=1$. By taking the two aspects together, it amounts to solve the constrained minimization problem (2) or the constrained maximization problem (3)

$$\mathbf{y}=\arg\max_{\mathbf{y}^T\mathbf{L}\mathbf{y}=1}\mathbf{y}^T\mathbf{B}\mathbf{y}. \tag{3}$$

There are three extensions of the above graph preserving criterion, i.e. linearization, kernelization and tensorization. In this paper, only the linear extension will be considered. In the linear extension, suppose that the high dimension data $\mathbf{X}$ will be linearly mapped to a lower dimensional subspace by linear projection $\mathbf{y}=\mathbf{w}^T\mathbf{X}$, where $\mathbf{y}\in\mathbf{R}^d$. Then the optimal projection direction $\mathbf{w}$ can be obtained by solving the following constrained maximization problem:

$$\mathbf{w}^*=\arg\max_{\mathbf{w}^T\mathbf{X}LX^T\mathbf{w}=1}\mathbf{w}^T\mathbf{X}BX^T\mathbf{w}. \tag{4}$$

The above constrained maximization problem can be reformulated as a general Rayleigh quotient problem [17]:

$$\mathbf{w}^*=\arg\max_{\mathbf{w}}\frac{\mathbf{w}^T\mathbf{X}BX^T\mathbf{w}}{\mathbf{w}^T\mathbf{X}LX^T\mathbf{w}}. \tag{5}$$

Most of the linear feature extraction methods, such as Linear Discriminant Analysis (LDA) [12], MFA [6] can be formulated within the graph embedding framework. The only difference among them just lies in the different definitions of the intrinsic graph G and the corresponding penalty graph G^p. Here we only present the graph definitions of LDA and MFA. LDA searches for the projections that minimize the intra-class scatter and at the same time maximize the inter-class scatter, which is equivalent to the problem (5) by defining the intrinsic graph and the penalty graph as

$$\begin{cases} S_{ij}=\delta_{c_i,c_j}/n_{c_i}, & i\neq j,\\ S^p_{ij}=1/N-S_{ij}, & i\neq j, \end{cases} \tag{6}$$

where $\delta_{c_i,c_j}=1$, if $c_i=c_j$, otherwise $\delta_{c_i,c_j}=0$. Obviously in the intrinsic graph of LDA, all the data points in the same class are interconnected with weight S_{ij}, while in the penalty graph the data points from different classes are interconnected with weight S^p_{ij}. Therefore, LDA fails to discover the local geometrical structure of the data manifold [6] and therefore can not deal with nonlinear problems. To preserve the local structure of the original data in the