Pattern Recognition 46 (2013) 2798–2806

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Large Margin Subspace Learning for feature selection

Bo Liu [a,b,*], Bin Fang [a], Xinwang Liu [c], Jie Chen [d], Zhenghong Huang [b], Xiping He [b]

[a] *College of Computer Science, Chongqing University, Chongqing 400044, PR China*
[b] *School of Computer Science and Information Engineering, Chongqing Technology and Business University, Chongqing, PR China*
[c] *School of Computer Science, National University of Defense Technology, Changsha, Hunan 410073, PR China*
[d] *Institut Charles Delaunay, Université de Technologie de Troyes, France*

ARTICLE INFO

Article history:
Received 2 August 2012
Received in revised form
5 December 2012
Accepted 21 February 2013
Available online 7 March 2013

Keywords:
Feature selection
$\ell_{2,1}$-norm regularization
Large margin maximization
Subspace learning

ABSTRACT

Recent research has shown the benefits of large margin framework for feature selection. In this paper, we propose a novel feature selection algorithm, termed as Large Margin Subspace Learning (LMSL), which seeks a projection matrix to maximize the margin of a given sample, defined as the distance between the nearest missing (the nearest neighbor with the different label) and the nearest hit (the nearest neighbor with the same label) of the given sample. Instead of calculating the nearest neighbor of the given sample directly, we treat each sample with different (same) labels with the given sample as a potential nearest missing (hint), with the probability estimated by kernel density estimation. By this way, the nearest missing (hint) is calculated as an expectation of all different (same) class samples. In order to perform feature selection, an $\ell_{2,1}$-norm is imposed on the projection matrix to enforce row-sparsity. An efficient algorithm is then proposed to solve the resultant optimization problem. Comprehensive experiments are conducted to compare the performance of the proposed algorithm with the other five state-of-the-art algorithms RFS, SPFS, mRMR, TR and LLFS, it achieves better performance than the former four. Compared with the algorithm LLFS, the proposed algorithm has a competitive performance with however a significantly faster computational.

1. Introduction

Feature selection is to choose a subset of the original features according to some selection algorithm, it has a wide range of applications, including bioinformatics [1], object detection [2], computer vision [3]. It aims to reduce the data dimensionality by removing the redundancy and the correlation of extracted features. It has been proven in both theory and practice effective in enhancing learning efficiency, increasing predictive accuracy, and reducing complexity of learned results, due to the fact that accuracy of most classification algorithms, such as SVM, can be affected notably when applying on tasks with a small number of training data or with high-dimensional inputs [4,5]. Feature selection for high-dimensional data is one of the most important topics in machine learning research. Recently feature selection based on large margin has been widely investigated [6,7]. Feature selection based on subspace learning algorithm for high dimensional data was also proposed [8], and achieved good results. However, research which combines subspace learning with large margin has not appeared. The main contributions of this paper include:

- An efficient and novel feature selection algorithm termed as Large Margin Subspace Learning (LMSL) is proposed. Different from traditional feature selection algorithms, LMSL is a subspace learning algorithm based on large margin framework. Firstly, we utilize the expectation to estimate the nearest missing (the nearest neighbor with the different label) and the nearest hit (the nearest neighbor with the same label) of the given sample. Then these nearest neighbors are projected into the subspace $\boldsymbol{W}$ $(\in \mathbb{R}^{d\times p}, d \gg p$, and to be described in Section 3.2). After that, we define a novel metric function based on large margin in the subspace. The objective function is formulated by the metric function. To obtain row-sparsity of the solution, an $\ell_{2,1}$-norm regularization is incorporated into the objective function.
- As the proposed objective function and constraints are nonconvex. In general, a global optimal solution is hard to be obtained. We propose an efficient algorithm that obtain suboptimal solution and give detailed description of the algorithm.
- Extensive experiments are conducted to evaluate the proposed algorithm. Experimental results confirm the efficiency of our algorithm compared with different types of feature selection algorithms.

[*] Corresponding author at: College of Computer Science, Chongqing University, Chongqing 400044, PR China. Tel.: +86 13996283578.
E-mail addresses: flyinsky723@gmail.com (B. Liu), fb@cqu.edu.cn (B. Fang).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.02.012

The rest of this paper is organized as follows: Section 2 reviews prior work on the feature selection. In Section 3, we describe a basic introduction on large margin theory and a novel feature selection algorithm. Experimental evaluation is reported and discussed in Section 4. Finally, we conclude the paper and give a perspective of future work in Section 5.

2. Related work

According to the way of utilizing label information, feature selection algorithms can be divided into supervised [9], unsupervised [10–12] two classes. The first class includes Fisher score, ReliefF [13], SVM-RFE [14], etc. From the perspective of the selection strategy, feature selection algorithms are divided into three categories [15]: filter, wrapper and embedded algorithms. The filter algorithms compute some score of a selected feature subset by information of each feature, the algorithms are computationally much cheaper and more efficient, such as: Fisher Score [16], Laplacian Score [17], Trace Ratio [18]. These algorithms can be encompassed by algorithm named SPFS (Similarity Preserving Feature Selection) [15]. Wrapper algorithms use a procedure that wraps around a learning algorithm, and repeatedly calls the learning algorithm to evaluate how well it does using different feature subsets. The wrapper algorithm was firstly proposed in [19]. One serious problem with these wrapper algorithms is their high computational complexity because they need to train a large number of classifiers. To alleviate this problem, forward and backward selection were proposed in [20]. The algorithms are heuristic algorithms, none of them guarantee the optimal solution.

Feature selection of embedded algorithms was recently emerged in [4,21]. For large-scale feature selection problems, the novel embedded algorithms were proposed by performing feature selection directly in the SVM formulation in [4,5]. For microarray data, the embedded algorithm named RFE was proposed in [14]. In this algorithm, an SVM classifier was iteratively trained with the set of features, and those with small weights were then removed from the set. In order to obtain the sparse solution and to improve the computability, ℓ_1-SVM with a linear kernel was adopted in [22]. To jointly perform feature selection and SVM parameter learning for linear and nonlinear kernels, authors in [23] proposed a convex framework with ℓ_1-SVM. However, for an arbitrarily complex nonlinear problem, these algorithms are still no better performance. In [6], the authors proposed a notable algorithm. The main idea of the algorithm is to decompose complex nonlinear problem into a set of locally linear through local learning, and then to learn feature selection in the margin theory. To establish margin-based error function in weighted feature space, the benefits of the introduction of the Expectation–Maximization algorithm are to solve the nearest neighbor of a given sample, which is unknown before learning. To improve the performance of feature selection, the large margin principle in [6] is adopted into our paper.

Recently, the problem of subspace learning has received a lot of interests in dimensionality reduction and feature selection for high-dimensional data. Popular dimensionality reduction algorithms include principal component analysis (PCA) [24], linear discriminant analysis (LDA) [16], locality preserving projection (LPP) [25], neighborhood preserving embedding (NPE) [26], graph optimization for dimensionality reduction with sparsity constraints (GODRSC) [27]. These algorithms can be interpreted in a unified graph embedding framework based on Laplacian matrix in [28]. A framework for joint feature selection and subspace learning was presented in [8], where authors reformulated the subspace learning problem and used $\ell_{2,1}$-norm on the projection matrix to obtain row-sparsity of the solution, this enabled to select relevant features and learn transformation simultaneously.

Feature selection is also closely related to distance metric learning. In [29], large margin component analysis (LMCA) for the low-dimensional projection of the inputs was proposed. The algorithm aimed at separating points in different classes by a large margin. Authors in [30] use the maximum margin score for discriminatively optimizing the structure of Bayesian network classifiers. For k-nearest neighbor (k-NN) classification from labeled samples, a Mahanalobis distance metric was learned by semidefinite programming in [31]. The metric was trained with the goal that the k-nearest neighbors belong to the same class while examples from different classes were separated by a large margin. In [7], authors introduced a margin based on feature selection criterion and applied it to measure the quality of sets of features. Experiments showed that the algorithms based on the large margin were effectiveness for feature selection.

In order to obtain the sparse solution, regularization was introduced into most algorithms previously mentioned. In [32,33], feature selection regularized by ℓ_1-norm showed interesting performance. However, due to the non-differentiability of ℓ_1-norm, the regularized problem was solved using sub-gradient method, which was complex and inefficient. In [34], a so-called Hybrid Huberized SVM (HHSVM) algorithm was proposed by compositing ℓ_1 and ℓ_2 norms ($\ell_{2,1}$-norm). Instead of individually using one of these two norms, this composite regularization had more favorable properties as it investigated the structure of the problem. This type of regularization was also introduced into the multi-task feature selection [35]. In [36], the Nesterov method was used to optimize the objective function with $\ell_{2,1}$-norm regularization, and an Euclidean space projection algorithm with a linear time complexity was proposed to improve the computational efficiency of the Nesterov method. In [15], the $\ell_{2,1}$-norm regularized objective function is formulated and the Nesterov method in [36] was used to solve the objective function. In [1], another efficient algorithm for the $\ell_{2,1}$-norm regularization was proposed. This algorithm did not require the gradient of the objective function and experiments showed its efficiency and fast convergence rate. Feature selection algorithm based on subspace learning and $\ell_{2,1}$-norm was proposed in [8], and the algorithm in [1] was used to solve the objective function.

3. The proposed algorithm

In this section, we will propose an algorithm termed as Large Margin Subspace Learning (LMSL). This algorithm considers the large margin of samples. It is more suitable to feature selection for high dimensional data.

In what follows, we will firstly introduce margin-based metric function. Then the motivation and theoretical basis of LMSL will be proposed. After that the objective function will be formulated and the solution method will be proposed. Finally, the implementation and analysis of the algorithm will be discussed.

3.1. Margin-based metric function

The margin plays a crucial role in current machine learning research. The basic idea of marginal feature selection is to measure the importance of features by the margin of samples.

Let $A=[x_1,\dots,x_n]^T\in\mathbb{R}^{n\times d}$ be a training data set, where x_n is the nth data sample containing d features, $Y=[y_1,\dots,y_n]$ is the corresponding class labels, and $d\gg n$. The margin of x_j was defined in [7] as the following:

Definition 1. Let A be a training data set, x_j be a sample from A, and w be a weight vector, then the margin of x_j about w is

$$\rho_j(w)=\sum_{k\in M_j}d_w(x_j,x_k)-\sum_{k\in H_j}d_w(x_j,x_k), \tag{1}$$