Neural Networks 69 (2015) 126–134

Contents lists available at ScienceDirect

Neural Networks

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

Locality preserving score for joint feature weights learning

Hui Yan *, Jian Yang

School of Computer Science and Engineering, Nanjing University of Science and Technology, 210094, China

ARTICLE INFO

Article history:
Received 13 January 2015
Received in revised form 6 May 2015
Accepted 3 June 2015
Available online 15 June 2015

Keywords:
Feature selection
Locality preserving
Adaptive neighbors

ABSTRACT

Locality preserving measurement criterion is frequently used for assessing the quality of features. However, locality preserving criterion based unsupervised feature selection algorithms have two widely acknowledged weaknesses: (1) The performance of feature selection heavily depends on the effectiveness of the similarity matrix, which is defined in the original space, and thus it is probably inconsistent with the one in the weighted space. (2) Greedy searching strategy neglects the correlation and redundancy among features. To alleviate these deficiencies, we propose a novel unsupervised feature selection algorithm by jointly learning adaptive nearest neighbors in the weighed space. An effective iterative algorithm is developed to solve the proposed formulation, where each iteration reduces to a convex subproblem which can be efficiently solved with some off-the-shelf toolboxes. The results of experiments on the UCI and face data sets demonstrate the effectiveness of the proposed algorithm, for outperforming many state-of-the-art unsupervised and supervised feature selection methods in terms of classification accuracy.

1. Introduction

In many fields of applications such as computer vision (Ma, Nie, Yang, Uijlings, & Sebe, 2012), pattern recognition and biological study (Yu et al., 2014), data are characterized as high dimensional feature vectors. In practice, only a small subset of features is really important, since most features are usually correlated or redundant to each other (Duda & Stork, 2008; Liu et al., 2011). Therefore, dimension reduction is often applied first to transform an original high feature space into its corresponding low feature space for a compact and accurate data representation. Feature selection is one of the fundamental problems in dimension reduction. Not only can it make the subsequential learning more efficient, but it can also improve results comprehensibility. In the past decade, feature selection has attracted more and more research attention (Brown, Pocock, Zhao, & Lujan, 2012; Deng & Runger, 2012; Gilad-Bachrach, Navot, & Tishby, 2004; Nguyen, Jeffrey, Simone, & James, 2014; Peng, Wang, Lei, Qing, & Simon, 2015; Yang, Shen, Ma, Huang, & Zhou, 2011).

There are three different kinds of feature selection algorithms: supervised algorithms (Duda & Stork, 2008; Nie, Huang, Cai, & Ding, 2010; Sikonja & Kononenko, 2003; Wolf & Shashua, 2005), semi-supervised algorithms (Xu, King, Lyu, & Jin, 2010; Zhao & Liu, 2007a), and unsupervised algorithms (Cai, Zhang, & He, 2010; Dy & Brodley, 2004; Wolf & Shashua, 2005; Zhao & Liu, 2007b), according to the way of utilizing label information (Liu & Yu, 2005). Supervised algorithms are able to select discriminative features by taking advantage of information encoded in the labels. Nevertheless, it costs quite an amount of time and labor to acquire labeled data. So it is common and easy to acquire a small quantity of labeled data and a large quantity of unlabeled data. Consequently, semi-supervised feature selection is developed to solve the so-called 'small labeled sample problem'. Since there is no label information available, unsupervised feature selection is much more challenging than a supervised one. Meanwhile relatively few investigations are dedicated to overcoming the problem of missed label information. In this paper, we focus on unsupervised feature selection algorithms.

Unsupervised feature selection mainly reflects two concerns: measurement criteria and searching strategies.

Unsupervised feature selection frequently stresses the importance of features based on data similarity criterion, manifold structure preserving criterion (He, Cai, & Niyogi, 2005), or information-theoretic criterion (Nguyen et al., 2014; Peng, Long, & Ding, 2005). Some researches demonstrate that exploring the local discriminative information of data is a successful strategy for dimension reduction (Sugiyama, 2006; Yang, Xu, Nie, Yan, & Zhuang, 2010). The frequently used feature measurement criterion is based on the capability of keeping data similarity or manifold structure in the original feature space. For instance, the key idea of Relief (Kira & Rendell, 1992) is to iteratively estimate feature importance according to the discriminative ability to distinguish neighboring samples

* Corresponding author. Tel.: +86 84317297 3038.
E-mail address: yanhui@mail.njust.edu.cn (H. Yan).

http://dx.doi.org/10.1016/j.neunet.2015.06.001
0893-6080/

from different classes. Recent studies have shown that many real world data have a distribution which lies near a low-dimensional manifold embedded in a high-dimensional ambient space (Belkin & Niyogi, 2003). This characteristic serves as a motivation for Laplacian Score (He et al., 2005). Laplacian Score is based on the locality preserving criterion, which says if two data points are close, their selected features should be close as well.

Feature searching strategy is an essentially combinatorial computational problem, which is often NP-hard. Fortunately, this issue can be alleviated to some extent by using a feature weighting strategy. Traditional feature weighting algorithms, e.g., Relief and Laplacian Score, evaluate statistical properties of the features, rank them, and then select features individually from original features set. More recently, researchers apply sparse regularization based model into joint feature selection. Cai et al. (2010) propose a two-step approach, coined Multi-Cluster Feature Selection (MCFS), which incorporates spectral regression and l_1-norm minimization. There are two major deficiencies in Cai et al. (2010). Firstly, the flat embedding for the data points in the original feature space is highly unlikely to be the one in the weighted space. Secondly, MCFS does not lead to a proper feature coefficient matrix in which each column of this matrix is optimized individually, and their sparsity patterns are independent. Thus in MCFS how to select features is not clear. Accordingly Nie et al. (2010) propose $l_{2,1}$-norm minimization model to achieve a row-sparse feature coefficient matrix, which then guides the features searching process. These efforts have shown that it is better to select features jointly than individually.

1.1. Motivation

- The locality preserving criterion describes the intrinsic geometric structure by constructing a k-nearest neighbors graph. Therefore, the performance of feature selection heavily depends on the effectiveness of graph construction. Because similarity measurement and feature search are often conducted in two separated steps, it suffers that the learnt data similarity in the original feature space may not be the optimal one in the weighted space, leading to a suboptimal result.
- Traditional feature weights learning algorithms determine the useful feature subset by first computing scores for each feature independently according to some criteria, and then adding features one by one into the feature pool with high scores. On the one hand, since this greedy (incremental) algorithm computes scores for each feature individually, it neglects the correlation among features. On the other hand, they cannot delete redundant features. For example, some features with top scores may be highly correlated to each other.

1.2. Contribution

In order to overcome the aforementioned disadvantages, and inspired by Nie, Wang, and Huang (2014), we aim to learn feature weights jointly in the space of weighted features. The main contributions of the paper are threefold:

- Feature selection and adaptive neighbors preserving are combined into a single framework, which can select the most informative features with the capability of keeping local data similarity.
- Instead of simply assessing each feature importance separately, our algorithm selects the optimal feature subset in batch mode by joint feature weights learning.
- An effective and efficient iterative algorithm is developed to solve the proposed formulation. The results of experiments on the UCI and face data sets show that the proposed algorithm outperforms many state-of-the-art unsupervised and supervised feature selection methods in terms of classification accuracy. Moreover it is verified that the proposed algorithm converges within very few iterations, and it is not sensitive to parameters.

The remainder of this paper is organized as follows: we review a typical feature selection algorithm based on local preserving criterion, i.e., Laplacian Score, in Section 2. We present the proposed Locality Preserving Score for joint feature weights learning, and its optimization in Section 3. In Section 4, comparative experiments are conducted and analyzed to show the performance of the proposed method. Finally, conclusions are drawn in Section 5.

2. Laplacian score

In this section, we introduce a famous feature selection approach based on locality preserving criterion, i.e., Laplacian Score, which is the closest to the proposed algorithm.

Given data samples $x_1, x_2, \ldots, x_n$, we denote $X = [x_1, \ldots, x_n] \in R^{d\times n}$ as the data matrix, where d is the number of features and n is the number of samples. Let x_{ir} denote the rth feature of the ith sample, $i = 1, \ldots, n$, $r = 1, \ldots, d$.

Laplacian Score is based on the observation that, two data points are probably related to the same class if they are close to each other. It defines the neighbors of x_i as the k-nearest data samples in the data set to x_i, and constructs an affinity matrix $S \in R^{n\times n}$ as follows:

$$S_{ij} = \begin{cases} d(x_i, x_j), & \text{if } x_i \in \mathcal{N}(x_j) \text{ or } x_j \in \mathcal{N}(x_i) \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $\mathcal{N}(x_j)$ denotes the set of k-nearest neighbors of x_i, and $d(x_i, x_j)$ measures the distance between x_i and x_j. The matrix S can be viewed as a similarity matrix of the graph with the n data samples as the nodes.

For the rth feature, we define $f_r = [x_{ir}, x_{2r}, \ldots, x_{nr}]^T$, $D = diag(S\mathbf{1})$, $\mathbf{1} = [1, \ldots, 1]^T$, and $L = D - S$ where L is named the graph Laplacian (Chung, 2007) in graph theory. To remove the feature mean, we define

$$\widetilde{f}_r = f_r - \frac{f_r^T D\mathbf{1}}{\mathbf{1}^T D\mathbf{1}}\mathbf{1}.$$

Like traditional greedy feature selection approaches that consider each feature individually, Laplacian Score defines the weight of the rth feature as follows:

$$LS_r = \frac{\widetilde{f}_r^T L\widetilde{f}_r}{\widetilde{f}_r^T D\widetilde{f}_r}. \tag{2}$$

We can rank each feature according to its Laplacian Score, and then select the top d' features with the lowest scores.

3. Locality preserving score for joint feature weights learning

3.1. Ideas and algorithm

Two major drawbacks of traditional feature selection methods based on locality preserving criterion become clear from the algorithm described in Section 2. Firstly, the similarity matrix S is defined in the original feature space, which may be inconsistent with the one in the weighted feature space. Secondly, they select features according to feature weights individually, which leads to the selected features as a whole suboptimal. As for the first problem, we derive the probability distributions of data in the weighted space (Nie et al., 2014; Sun, 2007), instead of direct pairwise distances in the original space. The purpose is to iteratively learn the probabilistic neighbors in the weighted space,