Knowledge-Based Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Sparse structural feature selection for multitarget regression

Haoliang Yuan*,[a], Junjie Zheng[a], Loi Lei Lai*,[a], Yuan Yan Tang[b,c]

[a] School of Automation, Guangdong University of Technology, Guangzhou 510006, China
[b] Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau 999078, China
[c] Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China

ARTICLE INFO

Keywords:
Multitarget regression
Feature selection
Structure matrix
Inter-target correlations

ABSTRACT

Feature selection is one of the most important dimensionality reduction techniques for its efficiency and interpretation. Recently, some sparse regression-based feature selection methods have obtained an increased attention from the research community. However, previous sparse regression-based feature selection methods are limited by the double-layer structure. To improve the learning performance, in this work, we propose a sparse structural feature selection model for multitarget regression, which utilizes a structure matrix to expand the double-layer to multi-layer structure. Our aim tries to explore the essential inter-target correlations. To enhance the robustness of our proposed method, we emphasize a joint $\ell_{2,1}$-norm minimization on the loss function, regression matrix, and structure matrix. An effective optimization method with provable convergence behavior is also proposed. Extensive experimental results on multivariate prediction demonstrate the effectiveness of our proposed method.

1. Introduction

Multitarget regression, also known as multivariate or multioutput regression, aims to predict multiple continuous variables using a common set of input variables [1,2,44,56,57,61]. The prediction outputs are in real values, as opposed to the closely related task of multilabel classification where the output variables are binary [21,33,37,47,53–55]. There are numerous applications for multitarget regression such as stock price forecasting [11], load forecasting [6], ecological modeling [24], and employment prediction [43]. To obtain the desired predictions, it is commonly required to collect lots of possible related features to form the high-dimensional inputs for representing the multiple outputs, which induce a complex input-output relationship [17,31]. However, high-dimensional inputs significantly increase the time and space requirements to process the data. Moreover, some features in the inputs may be irrelevant and redundant, which may result in low efficiency, overfitting, and poor prediction performance in learning tasks [40]. To address these issues, feature selection [23,39] can be adopted as an important data preprocessing technique to reduce the dimensionality of the high-dimensional data by finding relevant low-dimensional features.

Feature selection is designed to select a subset of features from the high-dimensional data. Since feature selection does not change the original semantics of the variables, hence it offers the advantage of interpretability. According to the way in which the learning methods is incorporated in evaluating and selecting features, feature selection methods can be roughly classified into three categories [15], i.e., filter methods [7,16,35,41], wrapper methods [25,48,49], and embedded methods [3,46,50].

In many practical applications, multitarget variables often exhibit statistical dependencies [2,44]. Regarding to multitarget regression, exploring the inter-target correlations is used to exhibit statistical dependencies. However, most of traditional feature selection methods [14] ignore to the consideration of the inter-target correlations to exhibit statistical dependencies. Currently, Zhen et al. [56,57] propose a multi-layer multitarget regression framework by deploying a structure matrix to construct a multi-layer structure. The aim of this multi-layer structure is to model the essential inter-target correlations. Moreover, multi-layer structure, which utilizes a structure matrix to explicitly explore the inter-target correlations, is more suitable to capture this complex input-output relationship for multi-target regression. In light of this multi-layer structure, we propose a sparse structural feature selection (SSFS) model, which emphasizes joint $\ell_{2,1}$-norm minimization on the loss function, regression matrix, and structure matrix. The $\ell_{2,1}$-norm based loss function is robust to outliers in data points [36]. While the $\ell_{2,1}$-norm constraint on regression matrix is to select features across all data points with row sparsity [14]. In addition to this, the $\ell_{2,1}$-norm constraint on the structure matrix is to find the intrinsic structure of

* Corresponding author.
E-mail addresses: hunteryuan@126.com (H. Yuan), l.l.lai@ieee.org (L.L. Lai).

https://doi.org/10.1016/j.knosys.2018.06.032
Received 13 January 2018; Received in revised form 20 June 2018; Accepted 23 June 2018

inter-target correlations [57]. The contributions of this paper are summarized as follows:

(1) We propose a new sparse structural feature selection model for multitarget regression by utilizing a multi-layer structure. Comparing with traditional feature selection methods, our method explores the intrinsic inter-target correlations to exhibit statistical dependencies for feature selection. Thus, it performs better than traditional methods.
(2) We emphasize a joint $\ell_{2,1}$-norm minimization on the loss function, regression matrix, and structure matrix to improve the robustness of our proposed model. To optimize this new minimization problem, we devise an efficient optimization strategy to solve this joint $\ell_{2,1}$-norm optimization problem. Some theoretical discussions are presented to show the convergence behavior and computational complexity of the optimization strategy. Extensive experimental results also confirm the effectiveness of our proposed method.

The remainder of this paper is organized as below. We describe the background about feature selection and different notations in Section 2. In Section 3, we briefly introduce the related works. In Section 4, we first introduce the concrete formulation of our model and then provide an effective algorithm to solve this problem. Then, we analyze the performance of SSFS in three aspects, i.e., mechanism of structure matrix, convergence behavior, and computational complexity. Section 5 provides some promising comparing results on various kinds of data sets, followed by the conclusions and future works in Section 6.

2. Background

2.1. Features selection

Filter methods select feature subsets based on a predefined criterion, which is completely independent on the learning methods. The widely used filter methods include PCA score (PCAscore) [32] and Laplacian score (LAPscore) [19]. PCAscore method assumes that larger variance means better representation ability. LAPscore method ranks the features by evaluating the power of locality preservation of each feature. It should be noted that the filter methods are relatively computationally efficient but may fail to select the most informative features for a particular learning task. Wrapper methods choose the features through learning methods, where a predefined classifier is usually desired. Dy and Brodley [10] explore the feature selection problem by using Expectation-Maximization clustering under two different performance criteria for evaluating candidate feature subsets. Maldonado and Weber [34] propose to find a subset of all the features by maximizing the performance of a Support Vector Machine classifier. Wrapper-based methods commonly perform better than filter models. However, the wrapper models are usually with more expensive computation and prone to the issue of overfitting. Embedded methods involve the feature selection into a joint framework of model construction. Thus, it usually regards feature selection as a part of the learning process, where the useful features are obtained by optimizing the objective function. Embedded methods receive increasing interests due to its superior performance. Constantinopoulos et al. [8] present a Bayesian method for mixture model training that simultaneously treats the feature selection and the model selection problem. Law et al. [26] propose a solution to the feature selection problem by casting it as an estimation problem.

Sparse representation has been a powerful tool for signal processing applications where the entity signal can be reconstructed based on the sparse signal [9,51]. Recently, some embedded feature selection methods based on the sparse regression framework have been proposed for specific applications [14]. Liu et al. [30] propose a multitask feature selection method (MTFS), which uses $\ell_{2,1}$-norm instead of ℓ_1-norm as the penalty. Nie et al. [36] propose a new robust feature selection method (RFS) with emphasizing joint $\ell_{2,1}$-norm minimization on both loss function and regularization. Xiang et al. [52] present a framework of discriminative least squares regression for feature selection. He et al. [18] study the problem of robust feature extraction based on $\ell_{2,1}$ regularized correntropy. Zhu et al. [59] propose a regularized self-representation model for feature selection, where each feature can be represented as the linear combination of its relevant features. Hou et al. [20] propose a novel unsupervised feature selection framework, in which the embedding learning and sparse regression are jointly performed. Li et al. [27–29] propose an unsupervised feature selection approach, which jointly exploits nonnegative spectral analysis and feature selection. Zhu et al. [60] propose a feature selection method by joint graph sparse coding, which considers both manifold learning and regression simultaneously to perform feature selection. Cai et al. [4] propose a feature selection approach, which has one $\ell_{2,1}$-norm loss function with an explicit $\ell_{2,0}$-norm equality constraint. Zhu et al. [58] propose an embedded feature selection method, in which feature selection can be conducted during the process of label recovery. Chang et al. [5] propose a joint feature selection framework based on sparsity and semi-supervised learning. Although these aforementioned methods have yielded the good performance for feature selection, their performances may also be further improved since their regression frameworks are limited by a double-layer structure, i.e., the input-output structure.

2.2. Notations and definitions

We employ the notations as usual throughout this paper. Reals are written as lowercase letters. Vectors are denoted by boldface lowercase letters, while matrices are presented as uppercase letters. We give an input data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]^T$, where $\mathbf{x}_i$ is the ith sample, n is the total number of samples, and d is the dimensionality of input space. Let $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n]^T$ be the target matrix, where $\mathbf{y}_i$ is the multivariate output of $\mathbf{x}_i$ and c is the dimensionality of output space. The matrix $\mathbf{I}$ denotes the identity matrix with an appropriate size. Given a matrix $\mathbf{A}$, $A_{i,j}$ denotes the ith row and jth colum element of $\mathbf{A}$, $\mathbf{A}_{i,*}$ denotes the ith row of $\mathbf{A}$, and $\mathbf{A}_{*,i}$ denotes the ith column of $\mathbf{A}$. The $\ell_{2,1}$-norm of matrix $\mathbf{A} \in \mathbb{R}^{d\times c}$ is defined as $\|\mathbf{A}\|_{2,1} = \Sigma_{i=1}^{d}\sqrt{\Sigma_{j=1}^{c}A_{ij}^2}$, and the Frobenius-norm of matrix $\mathbf{A}$ is denotes as $\|\mathbf{A}\|_F^2 = \Sigma_{i=1}^{d}\Sigma_{j=1}^{c}A_{ij}^2$.

3. Related works

In this section, we first introduce two widely-used filters methods, i.e., PCA score (PCAscore) [32] and Laplacian score (LAPscore) [19], and then describe two classic sparsity-inducing feature selection methods, i.e., multitask feature selection (MTFS) [30] and robust feature selection (RFS) [36].

3.1. PCAscore and LAPscore

PCAscore [32] is the simplest evaluation criterion for feature selection, as reflected by its representative power. We denote the variance of rth feature as V_r, which is defined as:

$$V_r = \frac{1}{n}\Sigma_{i=1}^{n}(X_{i,r} - \mu_r)^2 \tag{1}$$

where $\mu_r = \frac{1}{n}\Sigma_{i=1}^{n}X_{i,r}$. The magnitude of V_r reflects the representative ability of the feature. The larger V_r is, the more powerful representative ability of the feature will be.

LAPscore evaluates the feature with its ability of preserving the local structure. The measurement L_r of the rth feature is computed as below:

$$L_r = \frac{\Sigma_{i=1}^{n}\Sigma_{j=1}^{n}(X_{i,r} - X_{j,r})^2 S_{i,j}}{\Sigma_{i=1}^{n}(X_{i,r} - \mu_r)^2 B_{i,i}} \tag{2}$$

where $\mathbf{B}$ is a diagonal matrix with elements $B_{i,i} = \Sigma_{j=1}^{n}S_{i,j}$ and $S_{i,j}$ is the neighborhood relationship between $\mathbf{x}_i$ and $\mathbf{x}_j$, which is defined as: