Information Sciences 507 (2020) 240–255

Contents lists available at ScienceDirect

Information Sciences

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

Regularization-based model tree for multi-output regression

Jun-Yong Jeong[a], Ju-Seok Kang[b], Chi-Hyuck Jun[a,*]

[a] Industrial Management and Engineering, Pohang University of Science and Technology, Pohang, 37673, Republic of Korea
[b] Technical Research Laboratories Research Project Department, POSCO, Pohang, 37859, Republic of Korea

ARTICLE INFO

Article history:
Received 7 November 2017
Revised 19 June 2019
Accepted 11 August 2019
Available online 12 August 2019

Keywords:
Multi-output regression
Multi-target regression
Model trees
Sparse representation

ABSTRACT

Multi-output regression refers to the simultaneous prediction of several real-valued output variables to improve generalization performance by exploiting output relatedness. We propose a multi-output model tree that utilizes a regularization-based method to exploit the output relatedness when estimating linear models at leaf nodes. The proposed method can explain nonlinear input–output relation and provides easy interpretation of its mechanism based on input space partitioning and models at leaf nodes. The models exploit output relatedness by selecting common input variables to explain related output variables. We also present a computationally efficient two-stage splitting procedure that decreases the number of model estimations by analyzing residuals. We verify the effectiveness of the proposed method in a simulation study and demonstrate that it outperforms existing methods on several benchmark datasets. Furthermore, we apply the proposed method to real industry data as a case study to predict tensile qualities of plates.

1. Introduction

Multi-output regression, also known as multivariate regression and multi-target regression, refers to the simultaneous prediction of several continuous output variables to improve generalization performance [4,34]. Generalization performance can be improved by considering input–output relations and/or exploiting output relatedness. Typically, the input–output relation is nonlinear and varies over datasets. Thus, predictive models in a fixed form show limited generalization performance. The output relatedness represents that several input variables may have shared effects on the related output variables. These shared effects could be estimated more accurately by simultaneous estimation, which increases the corresponding effective size of observations [7].

Multi-output linear methods have focused on exploiting output relatedness when estimating a coefficient matrix by selecting input variables [22,35] or assuming a low-dimensional structure [1,20,28]. The input variable selection methods select the common input variables to explain related output variables. Basic methods use L1/L2 or L1/Linf norm regularization to select the common input variables to explain all output variables [35,39]. Advanced methods focus on enhancing the flexibility of variable selection by combining L1/Linf norm and L1 norm [19] or using an overlapping group structure of output variables [22]. The low-dimensional structure methods assume that a coefficient matrix has a low rank and coefficient vectors lie within low-dimensional space. A low rank property is achieved by penalizing the trace norm regularization [20,36] or factorizing a coefficient matrix [1,8,28]. However, a linear form cannot explain a nonlinear input–output relation.

* Corresponding author.
E-mail addresses: june0227@postech.ac.kr (J.-Y. Jeong), jskang80@posco.com (J.-S. Kang), chjun@postech.ac.kr (C.-H. Jun).

https://doi.org/10.1016/j.ins.2019.08.034
0020-0255/

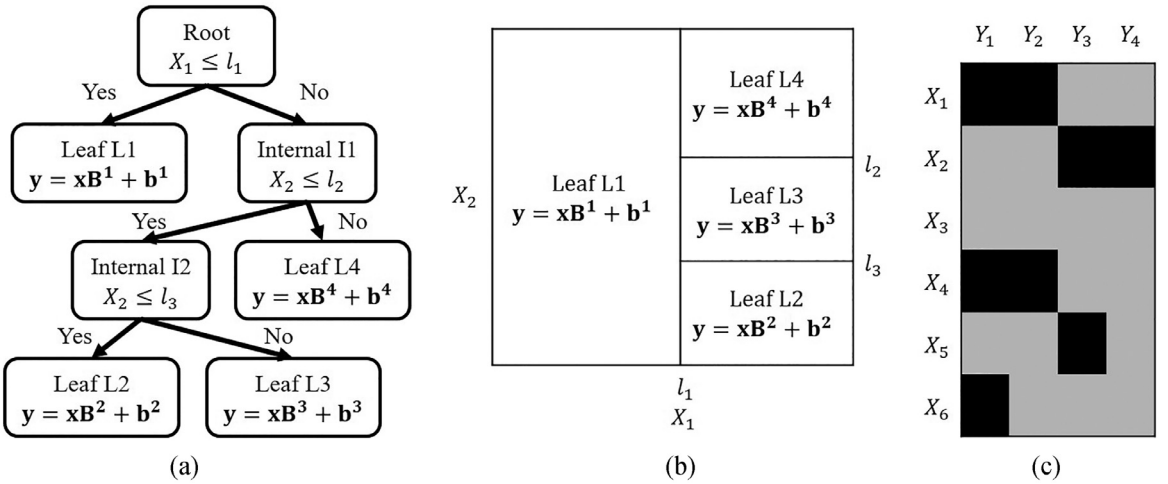

Fig. 1. Example of the proposed method: (a) tree structure, (b) linear models at leaf nodes, (c) coefficient matrix $\mathbf{B}^1$ at leaf node L1. Gray and black entries represent zero and non-zero values, respectively.

Multi-output tree methods have also been studied [10,11,31]. Single tree methods attempt to exploit output relatedness by defining splitting criteria, such as the Mahalanobis distance error [26], intra-cluster variation [41], and the Mallow distance [10], or by applying Chi-square tests on residuals [32]. Single tree methods have been extended to ensemble trees [23,24,27,38] and model trees [2,18]. The model trees are branches of tree methods where trees contain linear models at leaf nodes rather than constant forecasts [12,31,33,45]. Therefore, they can explain a nonlinear input–output relation and provide easy interpretation of its mechanism. However, existing multi-output model trees estimate a single-output linear model separately for each output variable and therefore exploit output relatedness only when determining split points [2,18]. Consequently, they induce trees that are not more accurate but only smaller than those from single-output model trees.

We propose a multi-output model tree that exploits output relatedness when determining split points as well as when estimating linear models. At leaf nodes, the proposed method utilizes a regularization-based method to exploit output relatedness by selecting the common input variables to explain related output variables. We present a computationally efficient two-stage splitting procedure that decreases the number of model estimations. First, we select promising candidate splits by analyzing the residuals of the current model and then evaluate them by estimating models. To the best of our knowledge, the proposed method is the first multi-output model tree that exploits output relatedness when estimating models at leaf nodes. We verify the effectiveness of the proposed method in a simulation study and show that the proposed method outperformed existing prediction methods on nine benchmark datasets. Furthermore, we apply the proposed method to a real industry data to predict tensile properties of plates collected from a world-leading steel company.

The remainder of the paper is organized as follows. In Section 2, we propose a model tree based on a regularization-based method and a two-stage splitting procedure. In Section 3, we first perform a simulation study to determine the effect of the number of observations and the correlation structure in output variables on the generalization performance of the proposed method. Then, we compare the proposed method to other multi-output regression methods on nine benchmark datasets. In Section 4, we present a case study with manufacturing data for predicting the tensile properties of plates. Conclusions and suggestions for future work are provided in Section 5.

2. Method

Consider D continuous input variables $X_1, \ldots, X_D$ and M continuous output variables $Y_1, \ldots, Y_M$. Each observation consists of a D-dimensional input vector $\mathbf{x} = (x_1, \ldots x_D)$ and an M-dimensional output vector $\mathbf{y} = (y_1, \ldots, y_M)$, where x_i $(i = 1, \ldots, D)$ and y_j $(j = 1, \ldots, M)$ are the realization of the input variable X_i and output variable Y_j, respectively. Given an input matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times D}$ and an output matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots \mathbf{y}_N]^T \in \mathbb{R}^{N \times M}$, we aim to learn function $\mathbf{h}$ that takes input vector $\mathbf{x}$ and predicts output vector $\mathbf{y}$ through the form of a model tree as follows.

$$\mathbf{h}: \mathbf{x} \in \mathbb{R}^D \rightarrow \mathbf{y} \in \mathbb{R}^M$$

We show an example of the proposed method in Fig. 1. As shown in Fig. 1(a), the proposed method uses a tree structure to partition input space recursively and utilizes a regularization-based method in estimating models at the resulting leaf nodes. The corresponding input space partitioning and linear models are shown in Fig. 1(b). At leaf nodes, a regularization-based method exploits output relatedness by selecting the common input variables to explain related output variables. Therefore, the corresponding coefficient matrix has sparsity. For example, in Fig. 1(c) where gray and black entries represent