Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

A novel bagging approach for variable ranking and selection via a mixed importance measure

Chun-Xia Zhang, Jiang-She Zhang, Guan-Wei Wang & Nan-Nan Ji

To cite this article: Chun-Xia Zhang, Jiang-She Zhang, Guan-Wei Wang & Nan-Nan Ji (2017): A novel bagging approach for variable ranking and selection via a mixed importance measure, Journal of Applied Statistics, DOI: 10.1080/02664763.2017.1391181

To link to this article: https://doi.org/10.1080/02664763.2017.1391181

Published online: 23 Oct 2017.

JOURNAL OF APPLIED STATISTICS, 2017
https://doi.org/10.1080/02664763.2017.1391181

Taylor & Francis
Taylor & Francis Group

A novel bagging approach for variable ranking and selection via a mixed importance measure

Chun-Xia Zhang[a], Jiang-She Zhang[a], Guan-Wei Wang[b] and Nan-Nan Ji[c]

[a]School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an Shaanxi, People's Republic of China; [b]School of Mechatronic Engineering and Shaanxi Key Laboratory of Non-Traditional Machining, Xi'an Technological University, Xi'an Shaanxi, People's Republic of China; [c]School of Science, Chang'an University, Xi'an Shaanxi, People's Republic of China

ABSTRACT

At present, ensemble learning has exhibited its great power in stabilizing and enhancing the performance of some traditional variable selection methods such as lasso and genetic algorithm. In this paper, a novel bagging ensemble method called BSSW is developed to implement variable ranking and selection in linear regression models. Its main idea is to execute stepwise search algorithm on multiple bootstrap samples. In each trial, a mixed importance measure is assigned to each variable according to the order that it is selected into final model as well as the improvement of model fitting resulted from its inclusion. Based on the importance measure averaged across some bootstrapping trials, all candidate variables are ranked and then decided to be important or not. To extend the scope of application, BSSW is extended to the situation of generalized linear models. Experiments carried out with some simulated and real data indicate that BSSW achieves better performance in most studied cases when compared with several other existing methods.

ARTICLE HISTORY
Received 27 April 2015
Accepted 8 October 2017

KEYWORDS
Variable selection; variable ranking; bagging; stepwise search algorithm; ensemble learning; selection accuracy; variable selection ensemble

AMS SUBJECT CLASSIFICATION
62J05; 68T05; 62F07

1. Introduction

Variable selection is always an important topic in statistical modeling since it can significantly enhance estimation and prediction accuracy, and improve interpretability. Currently, large amounts of high-dimensional data have emerged in many research and application areas because it is much more convenient to collect data than ever before. However, more variables do not necessarily imply that higher prediction accuracy can be obtained. At the same time, a large number of variables often prevent us from explaining how the exploratory variables influence our interested outcome. When facing with high-dimensional data, it is commonly believed that there are only a few variables playing an important role, namely, the model is usually sparse. As a result, it is particularly important to detect truly important variables to achieve high accuracy for prediction purpose or to identity relevant variables for interpretation purpose. In order to achieve this objective,

CONTACT Chun-Xia Zhang cxzhang@mail.xjtu.edu.cn School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an Shaanxi 710049, People's Republic of China