Knowledge-Based Systems 125 (2017) 13–25

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A ranking-based strategy to prune variable selection ensembles

Chun-Xia Zhang[a,*], Jiang-She Zhang[a], Qing-Yan Yin[b]

[a] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an Shaanxi 710049, China
[b] School of Science, Xi'an University of Architecture and Technology, Xi'an Shaanxi 710055, China

ARTICLE INFO

Article history:
Received 7 September 2016
Revised 24 January 2017
Accepted 29 March 2017
Available online 7 April 2017

Keywords:
Variable selection ensemble
Ensemble pruning
Selection accuracy
Aggregation order
Variable ranking
Stochastic stepwise selection

ABSTRACT

Ensemble learning has attracted significant interest in the literature of variable selection due to its great potential to reduce false discovery rate and to stabilize selection results. In this paper, a novel ensemble pruning technique called PST2E (i.e., pruned ST2E) is introduced to obtain smaller but stronger variable selection ensembles. For the ensemble members generated by the ST2E algorithm [3], we sort them in descending order according to the prediction errors associated with their determined models. Subsequently, only a desired number of members ranked ahead are integrated to compose a subensemble. On the basis of the average importance measures produced by the pruned ensemble, all candidate variables are then ranked and decided to be important or not. In the context of linear and logistic regression models, the experiments conducted with both simulated and real-world data illustrate that PST2E significantly outperforms several other popular techniques in most cases when evaluating them with multiple measures. Another advantage of PST2E is that it admits easy implementation. As a result, PST2E can be deemed as an attractive alternative to tackle variable selection tasks in real applications.

1. Introduction

In statistical modelling field, variable selection has been an old but active topic due to its great capability to prevent model overfitting, to improve prediction accuracy and to enhance the interpretability of a model. In this work, we focus on variable selection methods in the framework of a linear regression model

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \cdots + \mathbf{x}_p\beta_p + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T \in \mathbb{R}^n$ is the response vector for the variable Y and $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p) \in \mathbb{R}^{n\times p}$ is the design matrix for the covariates X_1, X_2, $\cdots$, X_p, and $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ are n independent observations. In addition, $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T \in \mathbb{R}^p$ is a p-dimensional unknown coefficient vector while $\boldsymbol{\varepsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^T \in \mathbb{R}^n$ is a normally distributed error term, namely, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ in which σ is unknown. In model (1), there is no intercept term since the response and covariates are assumed to be centered.

Over the past decades, high-dimensional data become more and more prevalent in many disciplines. One of the common features of these data is *sparsity*, i.e., only a few variables have true influence on the response. The primary aim of variable selection is to identify these critical variables for better prediction or interpretation purpose. Just as argued by many researchers [1–5], variable selection actually serves two different objectives depending on whether the modelling purpose is for prediction or for interpretation. The task of the former is to seek a parsimonious model so that future data can be well forecast or *prediction accuracy* can be maximized. But for the latter, analysts would like to identify which variables are most important to the response so that their relationship can be easily explained. In other words, variable selection serves in explanatory modelling to maximize *selection accuracy*. The theme of this paper is concentrated on selection accuracy.

In recent years, ensemble learning [3,6–9] has attained much attention for performing variable selection due to its great potential to reduce false discovery rate and to stabilize selection results. The core idea of ensemble learning is to utilize many base machines to solve a problem so that they can complement each other. In the framework of variable selection, the constructed ensemble machines can be called *variable selection ensembles* (VSEs). The main reasons for the success of VSEs can be explained below. As far as some traditional variable selection methods such as genetic algorithm [6], lasso [10] are concerned, much evidence [6–8] has demonstrated that they often pick out more variables than necessary (i.e., having high false positive rate). In the meantime, some approaches like subset selection and lasso have been confirmed to be instable, namely, small changes in data result in much variation of the obtained results [11,12]. For these methods, if the selection process is repeated using slightly different data for a number of trials, the frequency that the truly important variables are chosen

* Corresponding author.
E-mail address: cxzhang@mail.xjtu.edu.cn (C.-X. Zhang).

http://dx.doi.org/10.1016/j.knosys.2017.03.031
0950-7051/

will be high while that of unimportant ones being falsely considered as important will be low. As a result, the important variables can be distinguished from the remaining ones more easily.

The main contribution of this article is to propose a novel algorithm by introducing *selective ensemble learning* into the process of building a variable selection ensemble so that better selection results can be obtained. Considering that ST2E [3] is an effective approach to construct a good VSE, an additional pruning phase is injected into it. The selective phase is executed by a ranking-based strategy, i.e., sorting the members of ST2E according to a criterion and keeping only half or even less members to generate importance measures for each variable. In comparison with full ST2E ensembles as well as several other techniques on some simulated and real-world data, the pruned subensembles exhibit better performance in both selecting important variables and excluding uninformative variables. Furthermore, the considered methods are extended to generalized linear models.

The outline of the paper is as follows. Section 2 presents some related works. In Section 3, the novel ensemble pruning method for variable ranking and selection is discussed in details. Sections 4 devotes to discussing how to extend the considered methods to generalized linear models. To examine the performance of the proposed method, some simulations and real-world examples are used to carry out experiments to compare it with several other techniques in Sections 5 and 6, respectively. Finally, Section 7 offers the conclusions together with some future work of the paper.

2. Related works

2.1. Variable selection

At present, there exist numerous techniques in literature to tackle variable selection problems under many different circumstances, such as stepwise selection [13,14], coefficient shrinkage [10,15–18], variable screening [19–21] and etc. Among these approaches, shrinkage-type ones have gained high popularity due to their good performance to simultaneously achieve variable selection and coefficient estimation. They work by minimizing an objective function composed of a likelihood function plus a penalty term so that the coefficients of unimportant variables are automatically shrunk to zero while those of important ones are not affected. Fan and Lv [17] presents a comprehensive overview of this type of methodologies and related theories. The least square shrinkage and selection operator (Lasso) [10], the smoothly clipped absolute deviation (SCAD) [16] and adaptive lasso [18] are outstanding representatives.

However, the performance of the above-mentioned approaches highly depends on some tuning parameters involved in them, which is even hard for statisticians to specify. As a result, some scholars [1,3] advocated variable ranking (i.e., sorting variables in line with their relative importance to the outcome) instead of variable selection since the latter can be realized by adopting a thresholding rule once the variables are ranked properly. In current work, we will follow this line to address variable selection issue in linear and generalized linear models.

2.2. Ensemble learning and variable selection ensembles

Ensemble learning, a relatively new paradigm in machine learning, seems to be an omnipotent tool to successfully resolve many difficult problems in a large variety of applications [3,22–27]. Here, we should notice that the corresponding ensembles differ substantially according to whether the ultimate goal is to predict or to explain. Regarding the popular ensemble techniques like bagging and boosting, most of them are developed to build a *prediction ensemble* (PE) so that future data can be better predicted. In the context of variable selection, however, the *variable selection ensembles* (VSEs) are constructed to more accurately detect the variables which are truly important to the response.

Nevertheless, the creation of a PE or a VSE can both be divided into two steps, that is, *ensemble generation* and *ensemble integration*. The first step addresses how to generate a series of accurate and diverse base models. Generally speaking, this is a critical step to ensure the good performance of the final constructed ensemble. Since accuracy and diversity are mutually exclusive, we often face with the dilemma about how to achieve a good trade-off between these two terms [22,26,27]. With respect to ensemble integration, it aims to fuse the base models in a suitable way so that prediction accuracy or selection accuracy can reach as high as possible. Taking into account the subject of this paper, we will briefly review the techniques related to VSEs.

Aiming at creating a series of accurate but diverse individuals of a VSE, the usual practice is to apply a base learner (i.e., a variable selection method) on multiple different training sets or to inject some randomness into the learner. Among the existing techniques belonging to the first type, researchers generally perform selection on some bootstrap samples. For example, stability selection [8] executes lasso on multiple subsamples which are randomly drawn from the given data. It greatly eliminates the dependence of lasso on its regularization parameter and also reduces false discovery rate. Bin et al. [28] systematically studied the efficiency of subsampling and bootstrapping to stabilize forward selection and backward elimination. More recently, Zhang et al. [9] proposed a novel method PBoostGA by introducing the idea of boosting into the construction of VSEs. In addition, this type of methods also include BoLasso [7], random lasso [29] and bagged stepwise search (BSS) [30]. As far as the generation techniques to manipulate base learner are concerned, the core idea is to use a stochastic search algorithm to perform variable selection. The approaches of parallel genetic algorithm (PGA) [6], stochastic stepwise ensembles (ST2E) [3], RandGA [31] belong to this class.

In ensemble integration step, a simple averaging rule is commonly utilized to get a more stable and reliable importance measure for each variable. Roughly speaking, the results produced by each individual of a VSE can be stored in a matrix, say, $\mathbf{E}$, of size $B \times p$ where B stands for the number of constituent members in the VSE and p denotes the number of variables. Each element $\mathbf{E}(b, j)$ $(b = 1, \cdots, B;\ \ j = 1, \cdots, p)$ often takes value 1 or 0 (e.g., ST2E [3], random lasso [29]), which means whether the bth member considers variable j as important or not. Sometimes, the value of $\mathbf{E}(b, j)$ may be a number from the interval [0, 1] (e.g., PGA [6], RandGA [31]), which can be deemed as the jth variable's importance that is estimated by the bth member. Usually, the adopted base learner determines what type of values $\mathbf{E}(b, j)$ takes.

In order to identify which variables are important, we can first calculate the average importance measure for each variable j as

$$R(j) = \frac{1}{B}\sum_{b=1}^{B}\mathbf{E}(b, j), \quad j = 1, 2, \cdots, p, \tag{2}$$

and order the p variables from most to least important according to $R(1)$, $R(2)$, $\cdots$, $R(p)$. Next, a thresholding rule like mean rule or searching for the largest gap on the scree plot [32] can be executed. The former means to select the variables which satisfy $R(j) > (1/p)\sum_{k=1}^{p} R(k)$. The latter can be realized in the following way: sort $R(1)$, $R(2)$, $\cdots$, $R(p)$ in descending order; search for the largest gap between any consecutive entries; and select the variables which locate above the gap. Similar to [3], the former scheme will be adopted in later experiments unless otherwise specified.