Knowledge-Based Systems 153 (2018) 1–11

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Early stopping aggregation in selective variable selection ensembles for high-dimensional linear regression models

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 3 November 2017
Revised 2 April 2018
Accepted 11 April 2018
Available online 14 April 2018

Keywords:
Variable selection ensemble
Ensemble pruning
Variable selection
Selection accuracy
Aggregation order
Ranking accuracy

ABSTRACT

Nowadays, variable selection has become the most popular and effective tool to analyze high-dimensional data. Among the existing approaches, variable selection ensembles (VSEs) have exhibited their great power in improving selection accuracy and stabilizing the results of a traditional selection method. The construction of a VSE generally consists of two phases, i.e., ensemble generation and ensemble aggregation. We study selective VSEs in this paper by inserting a pruning step before combining the generated members into a VSE. As a result, a smaller but more accurate subensemble can be obtained. By taking ST2E (stochastic stepwise ensemble) as our main example, we first extended it to handle high-dimensional data. On the basis of its individuals, the aggregation order is rearranged according to their corresponding RIC_c (corrected risk inflation criterion) values. Then, only some members ranked ahead are averaged to estimate the importance measures for each candidate variable. In terms of several variable ranking and selection metrics, experiments conducted with simulated and real-world high-dimensional data show that pruned ST2E is superior to several other benchmark methods in most cases. By analyzing the accuracy-diversity patterns of VSEs, the pruning step is found to exclude less accurate members and lead the reserved members to more concentrate on the true importance vector.

1. Introduction

With the emergence of high dimensional data in various applications, variable selection has become an increasingly important tool to handle related problems. In high dimensional situations [1–5], it is commonly assumed that the true model is sparse in the sense that only a small proportion of the covariates are truly influential to the response. The main purpose of variable selection is to accurately identify these important variables. In doing so, the coefficient estimation accuracy and prediction accuracy of the fitted model can be greatly enhanced. More importantly, we can easily interpret how the covariates affect our interested outcome. Here, we have to differ the two different objectives that variable selection serves [3,6,7]. In predictive modelling, variable selection aims to seek a parsimonious model to maximize *prediction accuracy* or generalization ability of the fitted model. In interpretative modelling, however, it attempts to identify the true sparse model, or to maximize *selection accuracy* (i.e., the frequency that truly important variable are correctly identified to be important). In this paper, we will address variable selection tasks in high-dimensional linear regression models with selection accuracy as the target.

When facing with (ultra-)high dimensional data, shrinkage methods may be the most natural choice. This type of approaches include but are not limited to the least absolute shrinkage and selection operator (lasso) [8], the smoothly clipped absolute deviation method (SCAD) [9], the adaptive lasso [10] and so on. Fan and Lv [2] present a comprehensive review for these methods. However, their performance relies heavily on the tuning parameters. To specify the parameters properly so that the model selection consistency (i.e., as n increases to infinity, the probability that a method correctly identifies the true model tends to 1) can be achieved, a large variety of techniques such as extended Bayesian information criterion (EBIC) [11], modified BIC [12], corrected risk inflation criterion (RIC_c) [13] have been developed. As an alternative, people [6,14–16] often first sort variables in terms of their importance to the outcome and then employ a thresholding rule or a model selection criterion to make selection decisions. This kind of variable ranking and selection techniques are extremely useful in coping with ultra-high dimensional data. In view of its effectiveness and efficiency, we will follow the latter practice to perform variable selection in present work.

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

https://doi.org/10.1016/j.knosys.2018.04.016
0950-7051/

In recent years, the reproducibility of statistical findings has gained increasing attention of researchers [7,17,18]. As pointed out in [19], there are a number of causes for the irreproducibility. In the situations with high-dimensional data (especially with $p \gg n$), *variable selection uncertainty* is a main reason for the poor reproducibility. In fact, much evidence [3,17,18,20] has demonstrated that methods like subset selection and lasso can be highly *instable*. Here, "instable" implies that a small change to the data may lead to different outcomes. In interpretative modelling, instability is more undesirable since it is difficult for analysts to explain different findings. In addition, instable results are less reliable. Therefore, data analysts become less satisfactory with the output of a single model from a model selection process.

To resolve the above issues, *ensemble learning* [5,6,14,21–27] has exhibited its power in performing variable selection since averaging over a number of independent measures is often beneficial. Ensemble learning is well-known in machine learning fields for its good performance to construct a composite machine (also called *prediction ensemble*, abbreviated as PE subsequently) to make better prediction. The core idea of ensemble learning [28–31] is to enhance the performance of a single machine by constructing many base machines to complement each other. Motivated by the good behavior of PEs, Zhu and Chipman [14] extended it to the framework of variable selection and developed the first *variable selection ensemble* (VSE). The definition of a VSE will be provided in Section 2. In comparison with a single selector, a VSE holds the following advantages. First, it can greatly reduce false discovery rate and improve selection accuracy. For instance, genetic algorithm (GA) tends to select more variables than necessary (i.e., some noise variables are falsely included). Inspired by the bagging technique, [14] developed the parallel genetic algorithm (PGA) which significantly enhances the performance of GA. Second, a VSE can greatly reduce the risk to falsely select a model (i.e., miss important variables or wrongly include unimportant variables in the identified model). Last but not least, it can weaken the required assumptions for some methods like lasso to achieve model selection consistency [21]. It is well-known that the lasso [8] needs the so-called neighborhood stability condition (also known as irrepresentable condition) to achieve model selection consistency. The condition is usually very strong for the design matrix in a regression problem. In [21], the authors put forward a randomized lasso algorithm by first rescaling all input variables with *random weights* and then solving the standard lasso using the rescaled variables. In this way, the shrinkage imposed on each variable can be rescaled appropriately. Even though the idea is analogous to adaptive lasso [10], the reweighting method of [21] is random. By adopting stability selection, only a relaxed condition on the sparse eigenvalues of the design matrix can ensure that multiple runs of randomized lasso achieves selection consistency.

Because ST2E [6] has been confirmed to perform very well in many situations, we put our emphasis on the VSEs constructed by it. The main contribution of this article can be summarized as follows. First, ST2E is effectively extended to high-dimensional cases with the consideration that the existing literature [6,22] only studied it in small- or medium-scale problems. Second, a novel algorithm is proposed for building a smaller but more accurate VSE by applying the idea of *selective ensemble learning* (also known as ensemble pruning) to ST2E. In particular, a pruning step is executed by sorting the members of ST2E according to a criterion and fusing only a small proportion of top members. And thirdly, we investigate the accuracy-diversity patterns of the full and pruned ST2E ensembles to explore their differences. Compared with several other benchmark methods, the experiments conducted with simulated and real data show that pruned subensembles perform better in terms of variable ranking, variable selection and prediction in most situations. Moreover, the superiority is more significant in sparse high-dimensional models.

The remainder of the paper is organized as follows. Section 2 presents some related works of VSEs. In Section 3, the novel technique to prune ST2Es for variable ranking and selection is discussed in details. Sections 4 and 5 include some experiments conducted with simulated and real-world data to examine the performance of the proposed method, respectively. Finally, Section 6 offers the conclusions of the paper.

2. Related works

2.1. A brief introduction of variable selection ensembles (VSEs)

In [28], ensemble learning is defined as a process that uses a set of models with each being obtained by applying a learning process to a given problem. The set of models is integrated in some way to obtain the final prediction. Note that this is the definition for PEs. In fact, VSEs can also be put into this framework because they are brought forward by imitating the idea of PEs. A VSE includes a set of *variable selectors* whose outputs are fused to get the final results. Contrary to a PE in which each member is a predictive model, each individual in a VSE is a selector to identify which variables are important. Fig. 1 provides a generic diagram to illustrate the process to construct a VSE.

In general, we can build a VSE through two phases, that is, *ensemble generation* and *ensemble aggregation*. In ensemble generation, a series of accurate and diverse members (i.e., selectors in Fig. 1) are generated. Like PEs, the members of a VSE can be built by randomly sampling training data [21,23–27] or manipulating the *base learner* (i.e., a variable selection method) [6,14]. For example, stability selection [21] implements lasso on multiple subsamples which are drawn from the given data at random. De Bin et al. [26] made use of subsampling and bootstrapping to stabilize forward selection and backward elimination. As for the strategy to manipulate base learner, the core idea is to artificially inject some randomness and then to use the randomized algorithm to implement selection process. The results produced by each selector can usually be stored in a matrix, say, $\mathbf{E}$, of size $B \times p$ where B is the size of a VSE and p indicates the number of variables. Depending on the adopted base learner, each element $\mathbf{E}(b, j)$ $(b = 1, \ldots, B;\ j = 1, \ldots, p)$ often takes a binary value 1 or 0 (e.g., ST2E [6], stability selection [21], BSS [27]), or a real number from the interval [0,1] (e.g., PGA [14], PBoostGA [25]).

Based on $\mathbf{E}$, a simple averaging rule is commonly used as the aggregator to fuse the individuals into a VSE. Particularly, the average importance measure for variable j can be obtained as

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

According to $r_1, r_2, \ldots, r_p$, the p variables can then be sorted in descending order (i.e., from most important to least important). Subsequently, a further selection step [6] can be implemented if analysts would like to know which variables are important. To realize this process, one can utilize a thresholding rule like the mean rule. As an alternative, searching for the largest gap on the scree plot is also an effective technique.

2.2. Selective ensemble learning

In the study of PEs, a great number of *selective ensemble learning* (also known as ensemble pruning) [28,30,32–34] have been designed to improve prediction accuracy, to reduce storage need and to speed up prediction. Most methods work in the "overproduce