Computational Statistics and Data Analysis 94 (2016) 287–301

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Adaptive conditional feature screening

Lu Lin [a,*], Jing Sun [b]

[a] Shandong University Qilu Securities Institute for Financial Studies, Shandong University, Jinan, China
[b] School of Mathematics and Statistics Science, Ludong University, Yantai, China

ARTICLE INFO

Article history:
Received 20 March 2014
Received in revised form 31 August 2015
Accepted 1 September 2015
Available online 8 September 2015

Keywords:
High-dimensional data
Model free
Conditional feature screening
Adaptability
Marginal utility

ABSTRACT

When the correlation among the predictors is relatively strong and/or the model structures cannot be specified, the construction of adaptive feature screening remains a challenging issue. A general technique of conditional feature screening is proposed via combining a model-free feature screening with a predetermined set of predictors. The proposed centralization technique can remove the irrelevant part from the criterion of the model-free feature screening. Consequently, the new criterion can measure the marginal utilities of predictors conditional on the predetermined set of predictors. The conditional information about these predetermined predictors helps reducing the correlation among covariates and as a result the resulting method can reduce the false positive and the false negative rates in the variable selection procedure. Thus, our method is adaptive to both the correlation among the covariates and the model misspecification. The new procedures are computationally efficient and simple, and can be extended to other relevant methods.
© 2015 Elsevier B.V. All rights reserved.

1. Introduction

In some contemporary applications, such as biomedical imaging, functional magnetic resonance imaging, tomography, tumor classifications and finance, researchers are frequently confronted with high-dimensional variables and the models whose structure cannot be completely specified. In such situations, the number p of variables or parameters in the model can be much larger than the sample size n and only little information about the actual model structures is known in advance. When the correlation among the covariates is relatively strong and the model structures cannot be correctly specified, it is difficult to establish the dimension reduction methodologies that are adaptive to both the correlation among covariates and model misspecification. In this paper, we are going to try and address this issue.

It is known that when the dimension of predictor vector is much larger than the sample size, ranking and screening have been proved to be useful for dimension reduction under the situations where models are specified correctly and the true model structures are relatively simple, such as linear structure and generalized linear structure. This type of approaches is about feature screening or marginal utility screening. Fan and Lv (2008) first introduced sure independence screening (SIS) and iterated sure independence screening (ISIS) in the context of linear regression models; Fan et al. (2009) and Fan and Song (2010) extended the SIS and the ISIS to handle generalized linear models; Fan et al. (2011) developed the nonparametric independence screening (NIS) for nonparametric models with additive structure. For more related methodologies see Xue

* Corresponding author. Tel.: +86 531 88364791.
E-mail addresses: linlu@sdu.edu.cn (L. Lin), almasdu2010@gmail.com (J. Sun).

http://dx.doi.org/10.1016/j.csda.2015.09.002
0167-9473/© 2015 Elsevier B.V. All rights reserved.

and Zou (2011), Zhu et al. (2011), Li et al. (2012), Wang (2012), Zhao and Li (2012), Lin et al. (2013) and Chang et al. (2013), among others.

All the feature screening methods aforementioned are based on a common condition: the true model structures are specified accurately. Of course the performances of these methods critically depend on the belief that the models under study are equal to or at least are close to the underlying models. When the supposed structures are far from the underlying ones, however, their behaviors may become poor. To develop robust feature screening against model misspecification, Zhu et al. (2011) proposed a sure independent ranking and screening (SIRS). Their proposal can be available for a wide range of commonly used parametric and semiparametric models. Thus theirs could be thought of as a model-free method. Lin et al. (2013) proposed a nonparametric ranking feature screening (NRS) through local information flows of the predictors, by which the function-correlation between response and predictors can be captured successfully, without any model structure assumption. Li et al. (2012) proposed a distance correlation-based sure independence screening (DC-SIS). This is a model-free approach as well. Recently, He et al. (2013) introduced a quantile-adaptive model-free variable screening for high-dimensional heterogeneous data, such an approach allows the set of active variables to vary across quantile and thus make the variable selection more flexible to accommodate heterogeneity.

Moreover, as was mentioned in existing literature such as Fan and Lv (2008), Zhu et al. (2011) and Barut et al. (2012), the correlation among predictors heavily influence the marginal utility. When the correlation among the predictors is relatively high, simple feature screenings may result in false positives (i.e., the selected active predictors may be actually inactive) and false negatives (i.e., the true active predictors may be regarded as inactive predictors and then are removed from the models). Thus most of existing feature screening methods require the relevant conditions to restrict the correlation among predictors. However, it was proved by Hall and Li (1993), Fan and Lv (2008) that with growing dimensionality p, there always exist spurious correlations among predictors. Thus the correlation among predictors is an unevadable problem in statistical inference for all high-dimensional models. To adapt to the circumstances in which predictors may be relatively highly correlated, Cho and Fryzlewicz (2012) proposed a new criterion, for linear models, to measure the contribution of each predictor to response. Their method takes into account the correlations among predictors by projecting correlated predictors into the orthogonal spaces and then eliminating the correlations between the transformed variables. However, the projection method is difficultly applied in or cannot be extended to other models such as nonlinear models and nonparametric models.

In many applications, researchers know from previous investigations and experiences that certain predictors are responsible for the response. It was stated by Barut et al. (2012) that, with these known active predictors, conditioning can help reducing the correlation among the predictors. This is particularly the case when predictors share some common factors, as in many biological (e.g. treatment effects) and financial studies (e.g. market risk factors). Thus, it can be expected that conditioning could help improving the measure of marginal utility. But the conditional sure independence screening of Barut et al. (2012) strongly depends on the model structure assumption, generalized linear model, and needs to estimate the corresponding parameters. It is difficult to extend the method to the models with complex or unspecified structure.

As stated above, a strong correlation among the predictors seriously damages the quality of the existing feature screening methods. However, such a correlation can be easily predetermined. For example, the marginal correlation between any two predictors can be easily and efficiently estimated by sample correlation coefficient. It is an interesting issue to use such a predetermined correlation to reducing the correlation among the predictors and then to enhance the adaptability of the feature screening methods.

In this paper, a general technique is proposed for reducing correlation among the predictors and formulating conditional feature ranking. The key technique used here is to centralize the criterion of the existing model-free criterion, by which the irrelevant term that is related only to the predetermined set of predictors can be removed from the criterion of model-free screening. Consequently, the correlation between the centralized variable and the preselected variables is reduced significantly or eliminated completely, and the new criterion can measure the marginal utility of a predictor conditional on the known set of predictors. As stated above, the conditional information about the predetermined predictors helps reducing the correlation among predictors. It implies that the new method can reduce the false positive and the false negative rates in the variable selection process. This, together with the model-free property, ensures that our method is adaptive to both the correlation among the predictors and model misspecification, especially for the case of the number of the predetermined predictors being large. It is proved that with the number of predictors growing at an exponential rate of the sample size, the proposed procedure possesses consistency in ranking, which is both useful in its own right and can lead to consistency in selection. Moreover, unlike the conditional feature screening of Barut et al. (2012), the new criteria do not need to estimate any model parameter, the new procedures are computationally efficient and simple, and can be extended to other relevant methods.

The remainder of the paper is organized in the following way. In Section 2, the SIRS proposed by Zhu et al. (2011) is first reviewed to motivate the methodological development. Then, the SIRS is centralized so that the irrelevant term is removed from the original criterion and consequently, new conditional model-free feature screening is defined naturally. Furthermore, the consistent estimators for the new conditional model-free feature screening are proposed. In Section 3, for our method, the theoretical properties (including correlation reduction and ranking consistency) are investigated. Simulation studies, together with a two-stage procedure, are presented in Section 4, and the technical proofs are postponed to Appendix.

Download English Version:

https://daneshyari.com/en/article/6869448

Download Persian Version:

https://daneshyari.com/article/6869448

Daneshyari.com