Computational Statistics and Data Analysis 71 (2014) 725–742

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Using random subspace method for prediction and variable importance assessment in linear regression

Jan Mielniczuk [a,b,*], Paweł Teisseyre [a]

[a] Institute of Computer Science, Polish Academy of Sciences, Poland

[b] Warsaw University of Technology, Faculty of Mathematics and Information Science, Poland

ARTICLE INFO

Article history:
Received 16 December 2011
Received in revised form 24 September 2012
Accepted 28 September 2012
Available online 9 October 2012

Keywords:
Random subspace method
High-dimensional model selection
Prediction
Variable importance
Positive selection rate
False discovery rate

ABSTRACT

A random subset method (RSM) with a new weighting scheme is proposed and investigated for linear regression with a large number of features. Weights of variables are defined as averages of squared values of pertaining t-statistics over fitted models with randomly chosen features. It is argued that such weighting is advisable as it incorporates two factors: a measure of importance of the variable within the considered model and a measure of goodness-of-fit of the model itself. Asymptotic weights assigned by such a scheme are determined as well as assumptions under which the method leads to consistent choice of significant variables in the model. Numerical experiments indicate that the proposed method behaves promisingly when its prediction errors are compared with errors of penalty-based methods such as the lasso and it has much smaller false discovery rate than the other methods considered.

1. Introduction

Prediction problem with a high dimensional feature space is one of the most challenging tasks of contemporary applied statistics. There is a growing number of domains nowadays that produce data with a large number of features, while the number of observations is limited. Examples include microarray datasets that measure genes activity, Quantitative Trait Loci (QTL) data, drug design datasets, high-resolution images and high-frequency financial data among others. For examples and discussion, see e.g., Donoho (2000). An important and intensively studied line of research is focused on regularization, or penalty-based methods (cf. e.g., Tibshirani, 1996; Zou and Hastie, 2005). Another important approach is a method of dimensionality reduction based on the so called sure independence screening proposed by Fan and Lv (2008). Recently, Bühlmann et al. (2010) have introduced a novel, computationally feasible method relying on a certain hierarchical testing algorithm. There are also approaches using information criteria modified to the high-dimensional setup; see e.g., Frommlet et al. (2012). In this paper, we propose a different approach based on the random subset method (RSM).

In the RSM a random subset m of features having cardinality $|m|$ smaller than a number of potentially useful regressors M is chosen and the problem is solved with the reduced feature space of the selected predictors. Features under consideration are assigned weights based on their performance in the constructed solution. The selection of a random subset of features and model fitting is executed B times and a cumulative weight of a feature is calculated based on its relevance in all models where it is used. The cumulative weights (or scores) thus correspond to relative importance of variables in the considered problem. The variables are then ordered according to the assigned weights. The ordering is essential in a construction of a final model, which can be e.g. based on a predetermined number of the most significant predictors or obtained by a selection

* Correspondence to: Jana Kazimierza 5, 01–248, Warsaw, Poland. Tel.: +48 22 3800500; fax: +48 22 3800510.
E-mail addresses: miel@ipipan.waw.pl (J. Mielniczuk), teisseyrep@ipipan.waw.pl (P. Teisseyre).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.09.018

method applied to the hierarchical list of models given by the ordering. By choosing $|m|$ much smaller than M the problem of overfitting is circumvented. Note that in an extreme case when $|m| = 1$ the scores correspond to the individual performance of variables.

The procedure was proposed by Ho (1998) for classifying objects and independently by Breiman (2001) for the case when a considered prediction method was either a classification or a regression tree. Breiman's approach leads to a construction of a random forest. There, a score of a feature corresponds to the difference of prediction errors averaged over trees which used this feature and its analogue for which values of the variable are randomly permuted. For the important developments, see also Lai et al. (2006) and Draminski et al. (2008).

The RSM method belongs to the category of wrappers in the sense that feature selection is 'wrapped around' building a prediction method i.e. it is inherent part of its construction. Here, all variables are ranked first based on their averaged performance in small fitted models and then selection of variables is performed for ensuing hierarchical family of models with the use of cross-validation or independent test sample. Different group, called filters, includes methods for which the feature selection method is not related to the construction of a prediction tool. Such methods can perform ranking and variable selection simultaneously; for a representative example see e.g., Stoppiglia et al. (2003). It should be stressed however, that since ranking of variables in the RSM is based on fitting small linear models the method does not impose any conditions on the number of candidate variables M. In classification and regression it is proved to be an effective way to avoid pitfall of curse of dimensionality in situations when the number of features M is comparable or even significantly larger than the sample size n.

A related problem, which is also addressed by the RSM, is assignment of weights to the variables in such a way that their magnitudes would correspond to variables' usefulness in the prediction. Note that this problem is different from the explanation problem in which we try to determine significant variables in the 'true' model, that is a model which fits the data well. A variable may be important for prediction although it does not belong to the set of significant features even in the ideal case when the data conform to a certain model like a linear model (1). The problem of assigning scores to features which reflect their importance in prediction when the number of features is small compared to the sample size is also an important line of research; cf. Grömping (2007) for comprehensive review. Here, important development is the method proposed by Lindemann, Merenda and Gold (`lmg`); see Lindemann et al. (1980) and also Chevan and Sutherland (1991). In this approach, the score of variable x equals an average over all permutations r of $\Delta R^2_{x,r}$, where $\Delta R^2_{x,r}$ is an increase of coefficient of determination R^2 due to adding x to the list of active variables ordered by r (see in Section 4 for a formal definition). The method was further developed by Feldman (1999), who considered data dependent weights with their magnitude corresponding to the goodness-of-fit of the ordering. Note however, that for large M both approaches become extremely computationally intensive and they break down in the case of linear model fitting when M is larger than n. We also mention the weight assignment method based on Multivariate Adaptive Regression Splines (MARS) discussed in Section 4.2.

The aim of the paper is twofold. First, for a linear regression we introduce a new scheme of assigning scores to variables. In our approach variables in a randomly chosen subset are assigned weights equal the squared values of the respective t-statistics in the pertaining fitted model. We argue in Section 2 that this is an intuitively sound choice of weights as Eq. (3) indicates that the square of t-statistic is a product of two factors, one of which corresponds to the importance of variable within the model and the second to the importance of the model itself. Second, we investigate models based on the ordered features according to the proposed weighting and study their prediction strength by means of simulations. We establish some formal properties of the proposed scheme, namely we determine the form of asymptotic ordering of the variables when the subset is fixed (Theorem 1) and we establish the asymptotic form of weights assigned by the RSM (Theorems 2 and 3). In the case of fixed subset m and random regressors the ordering is asymptotically equivalent to that given by the multiple correlation coefficients of y and variables in m with a single consecutive variable dropped. This is an extension of Zheng and Loh (1995) result, who have shown that in the case when m contains all relevant variables the obtained ordering is such that the relevant variables precede all spurious ones.

The prediction accuracy of the RSM based approach is compared with that of the lasso by means of numerical experiments and its performance appears to be promising, especially when there are many potential strongly dependent regressors. It turns out that in the considered examples false discovery rate for the RSM is much smaller than for the other methods whereas positive selection rate for all of them is comparable and close to 1. Also, we compared the pertaining method of weight assignment with Breiman's measures and a weight assignment based on MARS.

The paper is structured as follows. Properties of the t-based ordering for a fixed subset of regressors are studied in Section 2 along with the examples in which the explicit conditions are given for it to be the correct one. Section 3 introduces the random subspace method and states the results for considered weighting scheme and Section 4 summarizes the outcomes of numerical experiments. Proofs of the results are relegated to the Appendix.

We define now a formal setup of the paper. Assume that observed data have the form $(\mathbf{Y}, \mathbf{X})$, where $\mathbf{Y} = \mathbf{Y}_n$ is an $n \times 1$ vector of n responses which variability we would like to explain and $\mathbf{X} = \mathbf{X}_n$ is an $n \times M$ design matrix consisting of vectors of M potential regressors collected from n objects. Responses are related to regressors by means of the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is an unobservable vector of errors, assumed to have $N(0, \sigma^2\mathbf{I})$ distribution. Vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)'$ is an unknown vector of parameters. We consider two scenarios: the case of deterministic and random $\mathbf{X}$. In the latter case rows of $\mathbf{X}_n$ constitute n independent realizations of M-dimensional random variable $\mathbf{x}$ and a vector $\mathbf{Y}$