A ROBUSTIFIED METALEARNING PROCEDURE FOR REGRESSION ESTIMATORS

Jan Kalina − Aleš Neoral

Abstract

Metalearning represents a useful methodology for selecting and recommending a suitable algorithm or method for a new dataset exploiting a database of training datasets. While metalearning is potentially beneficial for the analysis of economic data, we must be aware of its instability and sensitivity to outlying measurements (outliers) as well as measurement errors. The aim of this paper is to robustify the metalearning process. First, we prepare some useful theoretical tools exploiting the idea of implicit weighting, inspired by the least weighted squares estimator. These include a robust coefficient of determination, a robust version of mean square error, and a simple rule for outlier detection in linear regression.

We perform a metalearning study for recommending the best linear regression estimator for a new dataset (not included in the training database). The prediction of the optimal estimator is learned over a set of 20 real datasets with economic motivation, while the least squares are compared with several (highly) robust estimators. We investigate the effect of variable selection on the metalearning results. If the training as well as validation data are considered after a proper robust variable selection, the metalearning performance is improved remarkably, especially if a robust prediction error is used.

Key words: model choice, computational statistics, robustness, variable selection

JEL Code: C60, C14, C38

Introduction

Metalearning can be characterized as a useful tool for selecting a suitable (optimal) algorithm or method for a new dataset exploiting a database of training datasets. Within metalearning, the knowledge acquired over a training database serves as a prior information which can be incorporated to analyzing new datasets (Smith-Miles et al., 2014). Metalearning may be also alternatively denoted as methodology for optimal selection of algorithms (Mersmann et al., 2015).

The concept of metalearning has established its position in the machine learning community (particularly in the field of automated statistical learning) and has found applications in various tasks of optimization, computer science, and data mining and also (but with a smaller intensity) in econometrics. The design of a metalearning study requries the user to carefully choose five basic categories of input, which were denoted as P, A, F, Y and S by Smith-Miles et al. (2014):

[P] Problem (i.e. datasets),

[A] Algorithms (i.e. methods),

[F] Features (also denoted as metadata),

[Y] Prediction measure,

[S] Selection mapping (i.e. metalearning method).

This paper is interested in comparing estimators of parameters in the linear regression. Because the least squares estimator in the linear regression is very well known to suffer from the presence of outlying measurements (outliers), various robust estimators have been proposed as more resistant alternatives (Jurečková et al., 2019). A metalearning study for selecting the best one among various robust estimators was presented already by Peštová & Kalina (2018). There, however, a rather standard (non-robust) approach to metalearning was performed. The aim of the current paper is to increase the robustness of this metalearning procedure by means of prior variable selection and at the same time some newly proposed robust statistical tools. Recommending an estimator will be based on comparing a selected set of features (including robust ones) computed over a new dataset with those computed over training datasets.

Section 2 of the paper recalls the least weighted squares (LWS) estimator, which is one of promising estimator with a high robustness (if suitable weights are used). Robust measures of prediction error, including a novel version based on implicit weighting, are proposed in Section 3. The metalearning study is described in Section 4 and its results are presented in Section 5.

1 Least weighted squares estimator

This section recalls the LWS estimator proposed by Víšek (2011). The standard linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i, \quad i = 1, \ldots, n, \tag{1}$$

is considered, where $Y_1, \dots, Y_n$ are values of a continuous response variable and $e_1, \dots, e_n$ are random errors (disturbances) with a common value of $var\ e_i = \sigma^2$, where $\sigma > 0$. The task is to estimate the regression parameters $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$. The classical least squares estimator denoted as b_{LS} is very well known to be too vulnerable to the presence of outlying measurements (outliers) in the data (Jurečková et al., 2019).

The least trimmed squares (LTS) estimator of β investigated e.g. by Rousseeuw and van Driessen (2006) is currently the most commonly used robust regression estimator with a high breakdown point. Its weighted version is the LWS estimator, which assigns implicitly given weights to individual observations. The formal definition of the LWS requires the user to specify a sequence of magnitudes of weights

$$w_1 \geq w_2 \geq \cdots \geq w_n, \quad \sum_{i=1}^{n} w_i = 1, \tag{2}$$

which are assigned to individual observations only after some permutation.

In (1), let us denote squared residuals corresponding to a given estimator b of β as $u_i(b)$. We will consider arranged values of squared residuals

$$u^2_{(1)}(b) \leq u^2_{(2)}(b) \leq \cdots \leq u^2_{(n)}(b), \tag{3}$$

which allow to express the definition of the LWS estimator b_{LWS} in the form

$$\arg \min_{b \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} w_i u^2_{(i)}(b). \tag{4}$$

The LWS estimator has appealing properties (see the discussion in Kalina (2013)) and the estimator also turns out to perform well on real data (Kalina & Schlenker, 2015).

While the weight selection influences the LWS estimate, we recommend to assign a zero weight to some (at least small) percentage of observations, which ensures a high robustness. The LWS estimator remains consistent for any non-increasing weights (Víšek, 2011).

Let us now define novel weights for the LWS denoted as trimmed linear weights. We assume now the true level of contamination to be equal to $\varepsilon \cdot 100$ % with $\varepsilon \in [0, 1/2)$. We define $h = \lceil (1 - \varepsilon)n \rceil$, where $\lceil x \rceil = min\{n \in \mathbb{N}; n \geq x\}$, and use the notation I for indicator function. The trimmed linear weights are now defined as

$$w_i = \frac{h-i+1}{h} I[i \leq h], \quad i = 1, \dots, n. \tag{5}$$

Already for n exceeding (about) 20, the computation of the LWS estimator must exploit an approximate algorithm obtained as a weighted extension of the FAST-LTS algorithm of Rousseeuw & van Driessen (2006).

1.1 Outlier detection

We propose a novel approach for estimating the number of outliers in a dataset assuming the model (1). We compute the LWS estimator and to avoid confusion, its residuals will be denoted as $u_1^{LWS}, \dots, u_n^{LWS}$. We also need $\hat{\sigma}_{LWS}$, which denotes the estimator of σ obtained in (1) by the LWS estimator; using a constant $\gamma > 0$ evaluated in Víšek (2010), we use

$$\hat{\sigma}_{LWS}^2 = \frac{1}{n\gamma}\sum_{i=1}^{n} \widetilde{w}_i u_i^2, \tag{6}$$

which ensures $\hat{\sigma}_{LWS}^2$ to be a consistent estimator of σ^2 for the particular choice of weights. The optimal weights found by the LWS in (1), i.e. after the optimal permutation, are denoted as $\widetilde{w}_1, \dots, \widetilde{w}_n$. We propose the following simple rule for outlier detection in (1). An observation with index i $(i = 1, \dots, n)$ is considered to be outlying, if and only if

$$\frac{|u_i^{LWS}|}{\hat{\sigma}_{LWS}} \geq 2.5. \tag{7}$$

1.2 Weighted coefficient of determination

Further, we propose a novel weighted coefficient of determination, again based on the idea of implicit weighting. The standard coefficient of determination R^2 is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} u_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \tag{8}$$

and it is natural to propose its robust (implicitly weighted) version (say R_W^2) as

$$R_W^2 = 1 - \frac{\sum_{i=1}^{n} \widetilde{w}_i\, u_i^2}{\sum_{i=1}^{n} \widetilde{w}_i (Y_i - \bar{Y}_W)^2} \tag{9}$$

with weights $\widetilde{w}_1, \dots, \widetilde{w}_n$ equal to the optimal weights found by the LWS in (1), i.e. after the optimal permutation. We may interpret R_W^2 as a generalization of the implicitly weighted robust correlation coefficient of Kalina & Schlenker (2015) or an implicitly weighted analogy of the robust R^2 of Renaud & Victoria-Feser (2010). The definition (9) is meaningful, because the sums of squares can be decomposed as in the classical case, namely as in the form

$$\sum_{i=1}^{n} \widetilde{w}_i (Y_i - \bar{Y}_W)^2 = \sum_{i=1}^{n} \widetilde{w}_i \left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n} \widetilde{w}_i \left(\hat{Y}_i - \bar{Y}_W\right)^2, \quad (10)$$

where $\hat{Y}_i$ denotes the fitted value of the i-th observation, $\bar{Y}_W$ denotes the weighted mean $\bar{Y}_W = \sum_{i=1}^{n} \widetilde{w}_i Y_i$, and $u_i = Y_i - \hat{Y}_i$ for $i = 1, \dots, n$ are the residuals.

2 Robust measures of prediction error

In the metalearning study of Section 4, we use three different measures of prediction error for a given dataset. The notoriously popular measure of prediction error for the model (1) is the mean square error (MSE) defined as

$$MSE = \frac{1}{n} \sum_{i=1}^{n} r_i^2, \quad (11)$$

where $r_i = Y_i - \hat{Y}_i$ are prediction errors and $\hat{Y}_i$ denotes the fitted value of the i-th observation for $i = 1, \dots, n$. The standard MSE however suffers from the presence of outliers in the data.

A possible robust alternative is the trimmed mean square error (TMSE) defined as

$$TMSE(\alpha) = \frac{1}{h} \sum_{i=1}^{h} r_{(i)}^2, \quad (12)$$

where h is integer part of αn, $\alpha \in [0.5,1)$ is a fixed constant (ensuring $n/2 \leq h \leq n$), and squared prediction errors are arranged as $r_{(1)}^2 \leq \cdots \leq r_{(n)}^2$. We propose now a novel robust prediction error measure denoted as the weighted mean square error (WMSE) and defined as

$$WMSE = \sum_{i=1}^{n} w_i r_{(i)}^2 \quad (13)$$

with some non-increasing weights (2). This represents an analogue of the LWS estimator and the trimmed linear weights (5) may be a reasonably robust choice also here.

3 Description of the study

We perform a metalearning study using the 20 datasets previously analyzed by Peštová & Kalina (2018). We use R software package together with libraries rrcov, rda and e1071. We use three different measures of prediction of Section 3, namely MSE, TMSE$(3/4)$ and WMSE with trimmed linear weights. The primary learning is computed twice, namely over raw data (without a dimensionality reduction) and after a variable selection. The subsequently performed metalearning procedure is computed again twice, namely over raw data and also after a variable selection.

We use the following estimators of parameters in model (1):

- Least squares,
- Huber's M-estimator (Jurečková et al., 2019),
- LTS with h equal to $\lfloor 3n/4 \rfloor$, where $\lfloor x \rfloor$ denotes the integer part of $x \in \mathbb{R}$,
- LWS with trimmed linear weights (5).

We use the following classifiers for the metalearning task:

- Linear discriminant analysis (LDA),
- Regularized discriminant analysis (RDA) of Friedman (1989),
- Shrunken centroid regularized discriminant analysis (SCRDA) of Guo et al. (2012),
- A linear SVM (support vector machine) classifier.

A regularized version of LDA may namely improve robustness compared to the LDA, because it is known that a suitable regularization improves (local) robustness properties.

3.1 List of features

We use the following set of 9 features for each dataset.

(1) The number of observations *n*,
(2) The number of regressors *p* (excluding the intercept),
(3) Normality of residuals, evaluated as the *p*-value of the Shapiro-Wilk test,
(4) Skewness,
(5) Kurtosis,
(6) Heteroscedasticity of residuals evaluated as the *p*-value of the White's test,
(7) Condition number of the matrix $(X^T X)^{-1}$,
(8) Percentage of outliers, as estimated by a novel procedure of Section 4.2 below,
(9) Weighted coefficient of determination R_W^2 proposed in Section 2.2.

3.2 Variable selection

As described above, the metalearning computations are performed on raw data and also on data after a supervised variable selection, which will be presented in this section. We perform the Minimum Redundancy Maximum Relevance (MRMR) variable selection studied by Ding & Peng (2005). This popular approach requires to measure relevance of a set of variables for the classification task, i.e. to evaluate the contribution of a given variable to the classification task. Also it is necessary to use a measure of redundancy of a set of variables. While variable selection is often performed by means of hypothesis testing in practice, most often by *t*-tests, we must remark that such approach ignores the multivariate structure of data as well as the

problem of repeated testing, if each of the tests is (of course incorrectly) performed on the 5% significance level.

The procedure selects gradually one variable after another and these form a set denoted as *S*. Let the index corresponding to the group label be denoted as *H*. First, we need to evaluate for each variable (say *Z*) its relevance for the classification task, which will be evaluated as the *F*-statistic denoted as $F(Z,H)$. The very first selected variable maximizes this relevance among all variables. Further, we need to evaluate for each variable *Z*, which is not present in *S* yet, the value of

$$\frac{1}{|S|}\sum_{i\in S^*} F(X_i,H) - \lambda \sum_{i,j\in S^*} |r(X_i,X_j)|, \tag{14}$$

where S^* is the set *S* after adding the variable *Z* and *r* is the (Pearson) correlation coefficient. The parameter $\lambda \in (0,1)$ is found by leave-one-out cross validation. Such variable is selected to the set *S*, which maximizes (5). The selection of variables according to (5) is repeated and new variables are added to the set of selected variables until a given stopping rule is fulfilled. We require the selected variables to explain at least 90 % of the total variability of the data.

4 Results of the metalearning study

The results of primary learning heavily depend on the choice of the prediction error measure, as indicated in Table 1. The least squares estimator turns out to be the best most often, only if the standard MSE is used. Robust estimators are more suitable if a robust prediction error is considered; the LTS estimator is the most successful method for TMSE; and finally the LWS estimator is the most succesful method for WMSE. The results of the subsequently performed metalearning are presented in Table 2. Our comparison reveals the robust metalearning to be more sucessful compared to a standard approach. Thus, the new results are much improved compared to those of Peštová & Kalina (2018).

Conclusion

Metalearning has the ability to recommend a suitable algorithm for a given dataset, based on prior knowledge learned over training datasets. It is however known to suffer from vulnerability to outliers. At the same time, vulnerability to outliers is highly intertwined with the instability of metalearning; the relationship between robustness and stability was discussed by Breiman (2001) or Shawe-Taylor & Cristianini (2004).

A theoretical novelty of this paper is a proposal of several tools accompanying the LWS estimator. They include a robust coefficient of determination, a robust version of MSE, and a procedure for outlier detection. Each of these new implicitly weighted methods could be used also independently (i.e. not necessarily within metalearning). The computation remains intensive however also if this reliable algorithm is used.

An experimental study presented in this paper has a unique aim to perform the metalearning in a robustified way. The main factors contributing to the improvements are using a robust prediction error measure and performing a variable selection of the data. The MRMR variable selection is helpful also if some features do not suffer from the presence of redundant variables (which is the case e.g. of the coefficient of determination). In addition, it seems as an adequate aim for future research to investigate metalearning (exploiting again dimensionality reduction) for available real high-dimensional data with the number of variables in the order of thousands.

Acknowledgment

The work was supported by the project 17-07384S of the Czech Science Foundation.

References

Breiman, L. (2001): Random forests. *Machine Learning*, 45, 5-32.

Ding, C. & Peng, H. (2005): Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3, 185-205.

Friedman, J.H. (1989): Regularized discriminant analysis. *Journal of the American Statistical Association*, 84, 165-175.

Guo, Y., Hastie, T. & Tibshirani, R. (2012): *rda: Shrunken centroids regularized discriminant analysis*. R package version 1.0.2. https://CRAN.R-project.org/package=rda.

Jurečková, J., Picek, J. & Schindler, M. (2019): *Robust statistical methods with R.* 2nd edn. CRC Press, Boca Raton.

Kalina, J. (2013): Highly robust methods in data mining. *Serbian Journal of Management*, 8, 9-24.

Kalina, J. & Schlenker, A. (2015): A robust supervised variable selection for noisy high-dimensional data. *BioMed Research International*, 2015, Article 320385.

Mersmann, O., Preuss, M., Trautmann, H., Bischl, B. & Weihs, C. (2015): Analyzing the BBOB results by means of benchmarking concepts. *Evolutionary Computation*, 23, 161-185.

Peštová, B. & Kalina, J. (2018): Robust metalearning: Comparing robust regression using a robust prediction error. *Proceedings of the 12th International Days of Statistics and Economics (MSED 2018)*. Melandrium, Slaný, 1367-1376.

Renaud, O. & Victoria-Feser, M.-P. (2010): A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference*, 140, 1852-1862.

Rousseeuw, P.J. & van Driessen, K. (2006): Computing LTS regression for large data sets. *Data Mining and Knowlege Discovery*, 12, 29-45.

Shawe-Taylor, J. & Cristianini, N. (2004): *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press.

Smith-Miles, K., Baatar, D., Wreford, B., & Lewis, R. (2014): Towards objective measures of algorithm performance across instance space. *Computers & Operations Research*, 45, 12-24.

Víšek, J.Á. (2010): Robust error-term-scale estimate. In: *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis*. Institute of Mathematical Statistics Collections, vol. 7, pp. 254-267.

Víšek, J.Á. (2011): Consistency of the least weighted squares under heteroscedasticity. *Kybernetika*, 47, 179-206.

Tab. 1: Results of primary learning evaluated as the percentage of datasets, for which the given estimator is the best. The primary learning was performed using three different measures of prediction error.

	MSE	TMSE(3/4)	WMSE
	Primary learning performed over raw data		
Least squares	0.35	0.10	0.10
M-estimator	0.10	0.05	0.05
LTS	0.35	0.70	0.30
LWS	0.20	0.15	0.55
	Primary learning performed after variable selection		
Least squares	0.15	0.10	0.10
M-estimator	0.20	0.00	0.00
LTS	0.40	0.80	0.25
LWS	0.25	0.10	0.65

Source: own computation

Tab. 2: Results of the metalearning study evaluated as the ratio of correctly classified cases in a leave-one-out cross validation study.

	MSE	TMSE (3/4)	WMSE
	Metalearning performed on raw data		
LDA	0.35	0.45	0.45
RDA	0.35	0.45	0.45
SCRDA	0.40	0.50	0.50
SVM	0.40	0.50	0.50
	Metalearning performed after a prior variable selection		
LDA	0.65	0.70	0.70
RDA	0.65	0.75	0.70
SCRDA	0.70	0.80	0.80
SVM	0.75	0.75	0.80

Source: own computation

Contact

Jan Kalina
The Czech Academy of Sciences, Institute of Information Theory and Automation
Pod Vodárenskou věží 4, 182 08 Praha 8, Czech Republic
& The Czech Academy of Sciences, Institute of Computer Science
Pod Vodárenskou věží 2, 182 07, Praha 8, Czech Republic
kalina@cs.cas.cz

Aleš Neoral
Faculty of Nuclear Sciences and Physical Engineering
Czech Technical University in Prague
Břehová 7, 115 19 Praha 1, Czech Republic
neoraale@fjfi.cvut.cz