How to down-weight observations in robust regression: A metalearning study

Jan Kalina[1], Zbyněk Pitra[2]

Abstract. Metalearning is becoming an increasingly important methodology for extracting knowledge from a data base of available training data sets to a new (independent) data set. The concept of metalearning is becoming popular in statistical learning and there is an increasing number of metalearning applications also in the analysis of economic data sets. Still, not much attention has been paid to its limitations and disadvantages. For this purpose, we use various linear regression estimators (including highly robust ones) over a set of 30 data sets with economic background and perform a metalearning study over them as well as over the same data sets after an artificial contamination.
We focus on comparing the prediction performance of the least weighted squares estimator with various weighting schemes. A broader spectrum of classification methods is applied and a support vector machine turns out to yield the best results. While results of a leave-1-out cross validation are very different from results of autovalidation, we realize that metalearning is highly unstable and its results should be interpreted with care. We also focus on discussing all possible limitations of the metalearning methodology in general.

Keywords: metalearning, robust statistics, linear regression, outliers.

JEL classification: C14
AMS classification: 68T37

1 Introduction

Metalearning can be characterized as a perspective methodology for extracting knowledge from a data base of training data sets, allowing to apply the knowledge to new independent (validation) data sets. It can be described as learning to learn over metaknowledge, i.e. knowledge about whole data sets which serve as a prior knowledge rather than measured values contained in these data sets. Metalearning represents an approach to machine learning (i.e. automated statistical learning) popular in recent computer science and data mining [2], starting to penetrate also to economic applications [1] suitable also (but not limited to) high-dimensional economic data [14].

The currently most renowned works on metalearning principles [2, 13] recommend metalearning especially for those domains, in which a theoretical knowledge would be hard to acquire. However, besides appealing properties of metalearning, we must also realize its limitations, mainly its instability and sensitivity to data contamination. However, a truly critical evaluation of metalearning seems to be still missing. It is mainly the fully automatic character of the metalearning process which hinders a profound interpretation of the results. The community of computer scientists finds however heuristics and black-box procedures more appealing, i.e. does not incline to detailed interpretations anyway.

Metalearning principles have been already applied to search for the best robust regression method in [11]. Our new approach here is to use metalearning to predict the most suitable weighting scheme for the highly robust least weighted squares estimator of [15]. In addition, we accompany the study with a larger comparison of different classification methods from both multivariate statistics and machine learning. First, we recall principles and methods of robust regression in Section 2. We also propose several new weighting schemes for the least weighted squares estimator. Our metalearning study is described in Section 3 and the results are presented in Section 4. A discussion follows in Section 5. The final Section 6 discusses the instability of metalearning in general.

2 Robust regression

Throughout this paper, the standard linear regression model

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

[1]Institute of Information Theory and Automation of the Czech Academy of Sciences, Pod Vodárenskou věží 4, Praha 8 & Institute of Computer Science of the Czech Academy of Sciences, Pod Vodárenskou věží 2, Praha 8, Czech Republic, kalina@utia.cas.cz
[2]Institute of Computer Science of the Czech Academy of Sciences, Pod Vodárenskou věží 2, Praha 8, Czech Republic, pitra@cs.cas.cz

is considered with n values of p regressors and a continuous response Y under the presence of random errors $e_1, \dots, e_n$. The least squares estimator, which is the notoriously known standard estimation tool for the linear regression model, is however too vulnerable to the presence of outlying measurements (outliers) [7].

Within the framework of robust statistic, a variety of robust estimators was proposed as an alternative to the least squares. These robust tools are more suitable estimation techniques for data contaminated by outliers and thus gradually become important for the analysis of economic data [6]. M-estimators are the most commonly used robust methods, but do not possess a high breakdown point which has become of one fundamental robustness measures [4]. Therefore, highly robust estimators were proposed as an alternative, which is more reliable under a more severe data contamination.

The least weighted squares (LWS) estimator of [15] represents a (possibly highly) robust estimator defined as

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

where $u_i(b)$ is a residual corresponding to the i-th observation for a given b,

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

are values arranged in ascending order and h is a given trimming constant. Its properties were investigated by [3] or [9]. The estimator is consistent and asymptotically normal for $n \to \infty$. If suitable weights are chosen, the estimator is efficient under normal (non-contaminated) models and at the same time robust under contamination. This property seems to perform well on real data sets in various applications [8, 10].

Nevertheless, a systematic comparisons of the performance of robust regression estimator is missing in spite of intensive attempts [12], while proving theoretical results is tedious and only of limited application, because the results often depend on unknown quantities. Thus, there remains a lack of comparisons of robust estimators on real data and particularly it remains unknown how to choose weights for the LWS for a given data set. Indeed, insufficient comparisons can be described as one of main reasons why robust methods have not much penetrated to real applications. Therefore, metalearning seems as a reasonable tool for the task of predicting the most suitable method for a particular (new) data set or under particular conditions.

3 Description of the study

We proposed and performed a metalearning study with the aim to compare various linear regression estimators and to find a classification rule allowing to predict the best one for a given (new) data set. Based on a data base of training data sets, the aim is also to detect the most relevant criteria for determining the most suitable weights.

The primary learning task is to fit various linear regression estimators for each of the given data sets. The best estimator is found using a specified characteristic of a goodness of fit. The subsequent metalearning part has the aim to learn a classification rule allowing to predict the best regression method for a new data set not present in the training data base. Its input data are only selected features of individual data sets together with the result of the primary learning, which typically has the form of the index of the best method for each of the training data sets.

In general, the user of metalearning must specify a list of essential components (parameters) [13], which will be now described together with our choices for the particular metalearning study of Section 4.

3.1 Primary learning

Metalearning should always use real data sets, because any random generation of data is performed in a too specific (i.e. non-representative, biased) way. We use 30 data sets, while 24 of them were used already by [11] and we include 6 additional data sets with a clearly economic motivation, which come from the online UCI Repository. The list of all these 30 publicly available data sets is presented in Table 2.

In each of the data sets, we consider the model (1) and use one of the following five estimators, where (4) and (5) are novel proposals of this paper:

1. Least squares.
2. LWS with data-dependent adaptive weights of [3].
3. LWS with linear weights
$$w_i = \frac{2(n+1-i)}{n(n+1)}, \quad i = 1, \dots, n, \tag{4}$$
4. LWS with trimmed linear weights. Let the true level of contamination equal $\varepsilon \cdot 100\ \%$ for a known $\varepsilon \in [0, 1/2)$.

In this paper, we take $h = \lfloor 3n/4 \rfloor$, where $\lceil x \rceil = \min\{n \in \mathbb{N};\ n \geq x\}$. The weights equal

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

where $\mathbb{1}[.]$ denotes an indicator function.

5. LWS with weights generated by the (strictly decreasing) logistic function

$$w_i = \left(1 + \exp\left\{\frac{i - n - 1}{n}\right\}\right)^{-1}, \quad i = 1, \ldots, n. \tag{6}$$

In addition, each choise of weights is standardized to fulfil $\sum_{i=1}^{n} w_i = 1$. For the prediction measure, we use the most standard choice, i.e. the (prediction) mean square error (MSE). In the primary learning task, we find the best method for each data set. This is done using MSE in a leave-1-out cross validation, which represents a standard attempt for an independent validation. There is 1 randomly chosen observation left out, the estimator is computed and used to predict the value for the observation being left out. This is repeated for all n observations and the resulting values of the MSE are averaged. Then, the output of the primary learning is the knowledge (i.e. factor variable, index) of the best method for each of the data sets.

3.2 Metalearning

The subsequent metalearning task exploits 10 features for each data set and the factor variable of Table 1 denoting the index of the best method. We use the following features.

1. The number of observations n,
2. The number of regressors p (excluding the intercept),
3. The ratio n/p,
4. Normality of residuals, evaluated as the p-value of the Shapiro-Wilk test,
5. Skewness,
6. Kurtosis,
7. Coefficient of determination R^2,
8. Percentage of outliers estimated by the LTS (by the subjective method of Section 4 of [9]),
9. Heteroscedasticity of residuals evaluated as the p-value of the Whites test,
10. Condition number of the matrix $(X^T X)^{-1}$.

For the subsequent metalearning task, which is a task of classification to 5 groups, we exploit various classification methods. For those implemented in R software, we use default settings of all their parameters. This is true for support vector machines (SVM), k-nearest neighbors, a classification tree, and others. We use several less known methods including a regularized version of linear discriminant analysis (LDA) denoted as SCRDA of [5] or a robust version of LDA denoted as quadratic MWCD classification [8]. We also investigate the effect of dimensionality reduction, particularly in the form of replacing the data by 3 principal components obtained by the principal component analysis (PCA).

4 Results

We used the R software for all the computations. Table 1 shows the best method for each of the data sets using the notation $1, \ldots, 5$ according to list of methods in Section 3.1. Using MSE as a measure of prediction performance, both autovalidation and leave-1-out cross validation were used, where autovalidation means predicting for observations present in the training data, while cross validation is a standard attempt for an independent validation. It seems from the autovalidation study that the least squares estimator is the most successful among the five possibilities, while the LWS with data-dependent weights of [3] is the most successful according to the cross validation.

Further, the training 30 data sets are classified to one of the 5 groups using the 10 features. The results for various classifiers are overviewed in Table 2, namely as classification accuracy of the metalearning classification task evaluated in a leave-1-out cross validation study. The classification accuracy is defined as the number of correctly classified cases divided by the total number of cases (i.e. percentage of correct results). If all 10 features are used, the best result is 0.40 obtained with an SVM classifier with a Gaussian kernel.

5 Discussion

Let us now discuss the performance of robust methods withing the metalearning study aiming to predict the best weighting scheme for the highly robust LWS estimator in linear regression.

The LWS estimator turns out to be quite often more suitable than least squares, which is a novel argument in favor of the method. It is an important task to decide, for which there have been no recommendations for a practical data analysis [15]. The data-depent weights of [3] turn out to be the winner of the comparisons. However, it is not the best option for all data sets and the metalearning study allows to predict the most suitable weighting scheme also for a new (independent) data set.

It cannot be easily interpreted which features are the most relevant for predicting the most suitable weighting scheme. The reason is the black-box character of the SVM, which turns out to be best classifier here. Some other classifiers, which are quite commonly used in metalearning, yield relatively weaker results. This is also the case of the k-nearest neighbors, which seem to perhaps the most common method in the metalearning task. Further, our study presents also a unique comparison of SCRDA and MWCD-LDA with standard LDA. Both these approaches are able to outperform LDA, which is especially for the relatively recent MWCD-LDA a new argument in its favor.

The effect of reducing the dimensionality, which is often performed with metalearning, is clearly leading to a too drastic loss of information. Here, the principal component analysis does not allow to construct a reliable subsequent classification rule.

Finally, a big different between results of autovalidation and cross validation reveal the instability of metalearning, because omitting a single data set from the training data base leads to huge differences in the performance of the resulting classification rule. Therefore, we include a final Section 6 discussing the instability of metalearning in general and its reasons.

6 Instability of metalearning

Besides presenting results of a particular metalearning study across 30 real publicly available data sets, our study reveals also limitations of metalearning and motivates a possible future critical evaluation of the metalearning process. Possible factors contributing to the sensitivity and/or instability of metalearning include:

- The choice of data sets. We use here a rather wide spectrum of data sets with different characteristics from different research tasks, while metalearning is perhaps more suitable only for more homogeneous data (e.g. with analogous dimensionality) or for data from a specific narrow domain. Apparently, it remains difficult (and unreliable) to perform any extrapolation for a very different (outlying) data set.
- The prediction measure. In our case, MSE is very vulnerable to outliers.
- The number of methods. If their number is larger than very small, we have the experience that learning the classification rule becomes much more complicated and less reliable.
- The classification methods for the metalearning task depend on their own parameters or selected approach, which is another source of uncertainty and thus instability.
- Solving the metalearning method by classification tools increases the vulnerability as well, because only the best regression estimator is chosen ignoring information about the performance of other estimators.
- The process of metalearning itself is too automatic so the influence of outliers is propagated throughout the process and the user cannot manually perform an outlier detection or deletion.

A less sensitive (more robust) approach to metalearning remains to represent an important topic for future research. It is clear that such alternative methodology requires a more complex effort than just a robustification of each of its individual steps (e.g. using a robust prediction measure, a robust classifier or a robust dimensionality reduction). One idea seems to consider a robust version of the mean square error. Robustifying the final classification may be performed by means of using ensemble classifiers or perhaps regression methodology. This would require replacing the right column from Table 1 by additional knowledge, e.g. the performance of all 5 estimators for each data set. It is also important to investigate the effect of data contamination as well as dimensionality reduction on the sensitivity of metalearning.

Acknowledgements

The research was supported by grants 17-07384S and 17-01251S of the Czech Science Foundation.

References

[1] Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). MetaFraud: A meta-learning framework for detecting financial fraud. *MIS Quarterly*, 36, 1293–1327.

[2] Brazdil, P., Giraud-Carrier, C., Soares, C., & Vilalta, E. (2009). *Metalearning: Applications to data mining.* Berlin: Springer.

[3] Čížek, P. (2011). Semiparametrically weighted robust estimation of regression models. *Computational Statistics & Data Analysis*, 55 (1), 774–788.

[4] Davies, L. & Gather, U. (2005). Breakdown and groups. *Annals of Statistics*, 33, 977–1035.

[5] Guo, Y., Hastie, T., & Tibshirani, R. (2007). Regularized discriminant analysis and its application in microarrays. *Biostatistics*, 8 (1), 86–100.

[6] Hubert, M., Rousseeuw, P.J., & van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statistical Science*, 23, 92–119.

[7] Jurečková, J., Sen, P.K., & Picek, J. (2012). *Methodology in robust and nonparametric statistics.* Boca Raton: CRC Press.

[8] Kalina, J. (2012): Highly robust statistical methods in medical image analysis. *Biocybernetics and Biomedical Engineering*, 32 (2), 3–16.

[9] Kalina, J. (2015). Three contributions to robust regression diagnostics. *Journal of the Applied Mathematics, Statistics and Informatics*, 11 (2), 69–78.

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

[11] Kalina, J. & Peštová, B. (2017). Robust regression estimators: A comparison of prediction performance. In *Conference Proceedings MME 2017, 35th International Conference Mathematical Methods in Economics* (pp. 307–312). Hradec Králové: University of Hradec Králové.

[12] Mašíček, L. (2004). Optimality of the least weighted squares estimator. *Kybernetika*, 40, 715–734.

[13] 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.

[14] Varian, H.R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28, 3–28.

[15] Víšek, J.Á. (2011). Consistency of the least weighted squares under heteroscedasticity. *Kybernetika*, 47, 179–206.

[16] Yohai, V.J. & Zamar, R.H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, 83, 406–413.

Classification method	Classification accuracy
Classification tree	0.27
Logistic regression	0.33
LDA	0.33
SCRDA	0.37
Quadratic MWCD classification	0.37
Multilayer perceptron	0.30
k-NN ($k = 3$)	0.27
SVM (linear)	0.37
SVM (Gaussian kernel)	0.40
PCA $\Longrightarrow$ LDA	0.23
PCA $\Longrightarrow$ Logistic regression	0.23

Table 1 Results of metalearning evaluated as the classification accuracy in a leave-1-out cross validation study.

Index	Data set	The best method found by autovalidation	LOOCV
1	Aircraft	1	2
2	Amazon access samples	2	5
3	Ammonia	1	2
4	Auto MPG	1	2
5	Cirrhosis	3	3
6	Coleman	1	1
7	Communities and crime	2	5
8	Delivery	1	2
9	Education	1	2
10	Electricity	1	3
11	Employment	1	2
12	Energy efficiency	2	5
13	Facebook metrics	2	5
14	Furniture 1	1	1
15	Furniture 2	1	3
16	GDP growth	5	2
17	Houseprices	2	2
18	Housing	1	3
19	Imports	1	4
20	Insurance company benchmark	4	4
21	Istanbul stock exchange	2	5
22	Kootenay	3	3
23	Livestock	1	3
24	Machine	1	2
25	Murders	1	2
26	NOx	1	2
27	Octane	1	1
28	Pasture	1	4
29	Pension	2	3
30	Petrol	1	2

Table 2 Results of primary learning in a leave-1-out cross-validation study (LOOCV).

36th International Conference

Mathematical Methods in Economics

September 12th – 14th, 2018, Jindřichův Hradec, Czech Republic

Conference Proceedings

Published by:
MatfyzPress,
Publishing House of the Faculty of Mathematics and Physics Charles University
Sokolovská 83, 186 75 Praha 8, Czech Republic
as the 565. publication.
Printed by Reprostředisko MFF UK.

Printed in Prague — September 2018

Organized by:
Faculty of Management, University of Economics, Prague

Under auspices of:
Czech Society for Operations Research
Czech Econometric Society

Credits:
Editors: Lucie Váchová, Václav Kratochvíl
LaTeXeditor: Václav Kratochvíl
Cover design: Jiří Přibil
using LaTeX's 'confproc' package, version 0.8 by V. Verfaille

ISBN: 978-80-7378-371-6 (printed version)
978-80-7378-372-3 (electronic version)