ROBUST METALEARNING: COMPARING ROBUST REGRESSION USING A ROBUST PREDICTION ERROR

Barbora Peštová − Jan Kalina

Abstract

The aim of this paper is to construct a classification rule for predicting the best regression estimator for a new data set based on a database of 20 training data sets. Various estimators considered here include some popular methods of robust statistics. The methodology used for constructing the classification rule can be described as metalearning. Nevertheless, standard approaches of metalearning should be robustified if working with data sets contaminated by outlying measurements (outliers). Therefore, our contribution can be also described as robustification of the metalearning process by using a robust prediction error. In addition to performing the metalearning study by means of both standard and robust approaches, we search for a detailed interpretation in two particular situations. The results of detailed investigation show that the knowledge obtained by a metalearning approach standing on standard principles is prone to great variability and instability, which makes it hard to believe that the results are not just a consequence of a mere chance. Such aspect of metalearning seems not to have been previously analyzed in literature.

Key words: metalearning, robust regression, outliers, robust prediction error

JEL Code: C14, C63, C21

Introduction

Metalearning is a popular methodology for the task to learn knowledge over a training database and apply it to new independent data sets. In other words, the training data sets are discarded and only a set of their features (called metadata) is retained and used. The training data sets serve as a prior which can be incorporated to analyzing new data sets. We refer to Vilalta et al. (2004) or Suh (2012) for a very detailed overview of metalearning. Practical issues of metalearning have been discussed in the machine learning community (particularly in the field of automated statistical learning) and applied to various tasks of optimization, computer science, and data mining (Kordík et al., 2010).

While metalearning originated in the seminal paper (Rice, 1976) and its principles and appealing properties have been repeatedly appraised (Smith-Miles, 2014), its truly critical evaluation seems to be still missing. It si the fully automatic character of the metalearning process (without a detailed interpretation) which hinders a profound interpretation of the results, which would be standard in the statistical community. The field of computer science however finds heuristics and black box procedures more appealing.

The aim of this paper is to extend the metalearning study of Kalina & Peštová (2017) to a robust version exploiting also a robust measure of prediction error. The results are presented with a detailed interpretation in two particular situations. While an experienced practitioner would trust the results without a deeper analysis, we show that the seemingly optimistic result is obtained by a mere chance, while failing in separating noise from signal. Actually, the result down-weights the role of signal and builds conclusions from noise.

Section 1 of this paper recalls principles of robust statistical estimation in the linear reression model. A study comparing the prediction performance of several common linear regression estimators (standard and robust) over 20 real data sets is presented in Section 2. Results of our study are included in Section 3. However, some results are demonstrated as controversial in a detailed analysis in Sections 4 and 5.

1 Robust regression

It is well known that robust estimators of parameters are more suitable compared to the least squares, if observations in the standard linear regression model

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

are contaminated by outliers. Here, a continuous response Y is assumed together with the total number of p regressors (independent variables) and random errors $e_1, \ldots, e_n$. In this paper, we work with several robust alternatives of the least squares. All of them are reliable for data more or less contaminated by outliers for theoretical reasons as well as according to empirical evidence (Kalina, 2012; Jurečková et al., 2013).

Particularly, we work with Huber‘s or Hampel‘s M-estimators which are the most commonly used robust methods (Hampel et al., 1986). Because they do not possess a high breakdown point which has become of one fundamental robustness measures, we work also with the least trimmed squares (LTS) estimator of (Rousseeuw & Leroy, 1987), which is perhaps the most popular example of highly robust estimators.

The novelty of this paper is exploiting a robust evaluation of the prediction error in the search for the most suitable robust estimator. We attempt to perform a systematic comparison of the performance of various estimators using a trimmed version of the mean square error, similarly with Beck et al. (2006). Our robust metalearning approach extracts information from a database of training data sets with a larger or smaller level of contamination.

3 Design of the metalearning 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 perform a robust prediction of the best one for a given (new) data set. It remains also an open question which features (variables) are the most relevant criteria for determining the most suitable robust method.

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 robust 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 database. Its input data are only selected features of individual data sets together with the result of the primary learning, i.e. an index of the best method for each of the training data sets.

The user of metalearning must specify a list of essential components (parameters), which have been systematically described by Smith-Miles et al. (2014) and denoted (abbreviated) as P, A, F, Y, and S. Components designated as Problem (P), Algorithms (A) and Performance (Y, i.e. prediction measure) are used in the task of primary learning (base learning), while Features (F) and Selection mapping (S, i.e. metalearning method) in the subsequent metalearning. Their meaning and our specific choices will be now described.

3.1 Primary learning

The data sets (P) in metalearning should be always real data sets, because any random generation of data is performed in a too specific (i.e. non-representative, biased) way. We work with 20 publicly available data sets coming from Kalina & Peštová (2017); here however we limit the database to data sets which have at least roughly comparable sizes. All these data sets have an available documentation, which reveals that linear regression is a suitable model and that the data have undergone standard pre-processing (cleaning, transformations of variables). Only continuous regressors are used and all observations with any missing value are deleted. The standard linear model (1) is considered for each data set.

Algorithms (A) are used as four different regression estimators:

- Least squares,
- Huber's M-estimator (Hampel et al., 1986),
- Hampel's M-estimator (Hampel et al., 1986),
- Least trimmed squares (LTS) with two choices of the trimming parameter h, where smaller h (closer to 0.5) is suitable for a larger data contamination by outliers (Rousseeuw & Leroy, 1987; Víšek, 2006). We use h as the integer part of $3n/4$ or integer part of $n/2$; we refer to Kalina (2015) for a deeper discussion of choosing suitable h for the LTS.

Except for the least squares, all these estimators are robust. However, robust statistics distinguishes between local and global robustness (resistance, insensitivity). While Huber's and Hampel's M-estimators are robust in the local sense, only the LTS estimator is highly robust in the global sense, i.e. to outliers. We may refer to Hampel et al. (1986) or Víšek (2006) for a deeper explanation of the concepts, which are to a large extent contradictory.

Prediction measure (Y) is considered in the form of the mean square error (MSE) or its robust counterpart known as the trimmed mean square error (TMSE), defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n} u_i^2, \quad TMSE(\alpha) = \frac{1}{n}\sum_{i=1}^{k} u_{(i)}^2, \tag{3}$$

where prediction errors are denoted as $u_i = Y_i - \hat{Y}_i$ (for $i = 1, \dots, n$), $\hat{Y}_i$ are fitted value of the i-th observation (in each of the data sets), k is integer part of αn, $\alpha \in [0.5,1)$ is a fixed constant (ensuring $n/2 \leq k \leq n$), and $u_{(1)}^2(b) \leq u_{(2)}^2(b) \leq \cdots \leq u_{(n)}^2(b)$ are arranged values.

In the primary learning task, we find the best method for each data set using MSE or TMSE with a specified α in a leave-one-out cross validation, which represents a standard attempt for an independent validation. 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

We use 10 features of data sets (F). These include all the 9 features used by Kalina & Peštová (2017) and additionally we use the Donoho-Stahel outlyingness measure of all the regressors as the last feature. Metalearning method (S) is used as one of the 3 following classifiers:

- A linear support vector machine (SVM) classifier,
- Linear discriminant analysis (LDA),
- k-nearest neighbors for $k = 3$.

4 Results

For the primary learning, Table 2 shows the best method for each of the data sets. The best method is evaluated by means of MSE in columns (a) to (c), TMSE(0.9) in columns (d) to (f), and TMSE(0.5) in columns (g) to (i). Here, we also exploit the effect of merging some estimators to reduce the number of considered groups. Thus, the classification is considered to

- 5 groups (in columns a, d, g),
- 3 groups (in columns b, e, h) obtained by merging Huber's and Hampel's estimator together and merging the LTS estimators with both values of h together,
- 2 groups (in columns c, f, i) obtained by merging the least squares with M-estimators together and merging the LTS estimators with both values of h together.

We consider the merging useful, because it allows to interpret if a highly robust method is desirable or not, or if the least squares estimator is suitable or not, while the approach with 5 groups further specifies also a more delicate classification.

The subsequent metalearning task starts with classifying the 20 training data sets to one of the 5 groups using the 10 selected features. The results of metalearning are overviewed in Table 1, namely as classification performances of the classification rules learned within the metalearning tasks. There, the performance is evaluated as a classification correctness in a leave-one-out cross validation study, which is performed as a common attempt for an independent validation.

The first row of the table presents results with all 10 features for the standard prediction error MSE and also for its two robust versions. The presented value is the best among the 3 considered classifiers, which were listed under (S) above. The best result with 10 features and all 5 groups is 0.47 obtained with TMSE(0.9). We also exploit the effect of merging some estimators to reduce the number of considered groups of estimators. Thus, we consider a classification task not only to 5 groups, but also to 3 or 2 groups constructed after merging the estimators as described in the introduction of Section 3.

If all 10 features are used, the best result is 0.67 with TMSE(0.9) and only 2 groups. Further, we investigated the classification with only a subset of the features. We investigated the classification performance for all possible subsets of the 10 features. We can see that the classification performance is able to increase (even remarkably) if some of redundant features are ignored. We can see that the best result is 0.86 obtained with TMSE(0.5) or TMSE(0.5) but with only 2 features. These will be interpreted in the next section. Let us now also state that the best classifier is sometimes LDA and sometimes SVM, but the k-nearest neighbors classifiers suffer from the most dramatic loss of performance, although the method is very common (perhaps the most common) in the metalearning task.

5 A more detailed interpretation I

We pay closer attention to interpretation of the results presented in Section 4. We simplify the situation as much as possible and have a closer look at two particular results of Table 1. We use TMSE(α) with $\alpha = 0.5$ and classification to 2 groups with 1 variable. Then, the classification performance is equal to 0.71 with LDA.

The result is obtained with best single variable, which is the p-value of the Breusch-Pagan heteroscedasticity test, which will be denoted as p_{BP}. The classification rule of LDA can be interpreted this way: the LTS is the best method for the particular data set if and only if p_{BP}-value is smaller than 0.4. We point out that a p-value in general is not constructed to be compared with the value 0.4. Actually, its distribution under homoscedasticity is uniform over (0,1). In addition, p_{BP} also depends on the size of the data set and is also sensitive to violations of normality. The relatively high classification accuracy is attained thanks to a non-balanced situation, as there are 13 data sets (i.e. 62 %) for which the LTS is the best method. 85 % of them have $p_{BP} < 0.4$. Out of the remaining 8 data sets, 4 have $p_{BP} < 0.4$, because they are very small (although not ideally homoscedastic) and the other half has $p_{BP} > 0.4$, which are not so small but are nicely normal. Roughly speaking, the result is attained only by a mere chance but is not contraintuitive, because for data not extremely nice in terms of p_{BP}, the LTS is predicted as the best method.

6 A more detailed interpretation II

We use TMSE(α) with $\alpha = 0.9$ and classification to 2 groups with 1 variable. Then, the classification performance is equal to 0.86 with LDA. The result is obtained with best single variable, which is the p-value of the Shapiro-Wilk normality test denoted as p_{SW}. LDA predicts the LTS to be the best method for a particular data set if and only if $p_{SW} > 0.695$. This is contraintuitive, because contaminated data sets should have actually rather smalle values of p_{SW}. Still, the result is correct as we now explain. There are 8 data sets (i.e. 38 %) for which the LTS is the best method. Out of them, 3 data sets have $p_{SW} < 0.695$, because they are contaminated by outliers, and 5 have $p_{SW} > 0.695$, because they are very small and the test has a very low power. Out of the remaining data sets, 92 % has $p_{SW} < 0.695$; these data sets are not very contaminated but not extremely nicely normal.

This section does not discredit metalearning, but reveals its single dangerous aspect. If there are more features and many of them are only noise, the same effect may happen that the noise prevail over the signal. Then, the classification rule of metalearning may come in an extreme situation to a contraintuitive result if not closely investigated and interpreted.

Conclusion

Metalearning can be characterized as a popular tool in various tasks, however typically used in a form which is vulnerable to outliers. Here, we attempt to perform the metalearning task of extracting information from training data sets by means of a robust measure of prediction error. The acquired knowledge is further applied on new data sets. While the results allow to predict the most suitable robust regression estimator for a given data set, a detailed analysis illustrates rather controversial results, which have not been described in references.

The metalearning study performed in the context of robust estimators in linear regression brings new useful knowledge about suitability of robust estimators for various data sets. Although a lot of effort was invested to a study of properties or even optimality of robust regression estimators, there seem no theoretical results allowing to predict which of the considered estimators is the most suitable for a particular data set. Nor it is clear which are the main criteria for predicting the most suitable estimator. Therefore, this paper resorts to a metalearning study, which is performed in the spirit described by various references (Smith-Miles et al., 2014). Particularly, we performed a metalearning study investigating the prediction performance of robust regression estimators and to compare them on 20 carefully selected real data sets with different properties and coming from different fields. Our approach stands is actually unique in the context of robust linear regression.

First, a direct comparison of the prediction performance of individual estimators shows that there is no single method uniformly better than all remaining ones. For MSE, the least squares estimator performs as the most reliable estimator, while M-estimators are the winner for TMSE(0.9) and the LTS is the best estimator if TMSE(0.5) is used. For the last case, the very robust measure prefers a highly robust method. To be more precise, a measure robust to severe outliers prefers an estimator again robust to severe outliers, while major disadvantages of the estimator (low efficiency or local sensitivity) are not revealed. The LTS performs in a very different way from the remaining estimators, i.e. the differences between least squares and M-estimators are relatively small. Robust versions of the prediction error also allow actually to improve the classification performance compared to MSE.

On the whole, several ideas allowed to improve remarkably the classification performance obtained with 10 features, 5 groups and MSE, which equals to 0.10. Improvement was acquired with reducing the number of features, reducing the number of groups (estimators), and using a robust counterpart of MSE. The best classification performance is improved to 0.86, which seems already very reliable and a standard metalearning procedure would be finished with announcing this seemingly promising result.

Further, a unique interpretation of metalearning results is presented in the paper. Our detailed analysis in two special cases shows that the were caused by a mere chance, are overly optimistic and do not make sense. Such contra-intuitive result appears here in a rather extreme situation with the metalearning based on a single best feature. In practice, one would definitely use more feature than one, but such a phenomenon may occur also in the multivariate case. To support this claim, it is sufficient to accompany the single variable with other variables which are completely random. Relying of results of the fully automated metalearning process without any attempt for a critical evaluation of results may thus be considered a weak (but common) strategy not having been reported in literature.

Acknowledgment

The work was supported by the project 17-01251S (Metalearning for extraction of rules with numerical consequents) of the Czech Science Foundation.

References

Beck, A., Ben-Tal, A., & Eldar, Y.C. (2006): Robust mean-squared error estimation of multiple signals in linear systems affected by model and noise uncertainties. *Mathematical Programming*, 107, 155-187.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., & Stahel, W.A. (1986): *Robust statistics: The approach based on influence functions*. New York: Wiley.

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

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

Kalina, J. (2015): Three contributions to robust regression diagnostics. *Journal of Applied Mathematics, Statistics and Informatics*, 11 (2), 69-78.

Kalina, J. & Peštová, B. (2017): Robust regression estimators: A comparison of prediction performance. *Proceedings Mathematical Methods in Economics (MME 2017)*, University of Hradec Králové, 307-312.

Kordík, P., Koutník, J., Drchal, J., Kovářík, O., Čepek, M., & Šnorek, M. (2010): Meta-learning approach to neural network optimization. *Neural Networks,* 23, 568-582.

Rice, J.R. (1976): The algorithm selection problem. *Advances in Computers*, 15, 65-118.

Maechler, M., Rousseeuw, P.J., Croux, C., Todorov, V., Ruckstuhl, A., Salibián-Barrera, M., Verbeke, T., Koller, M., Conceicao, E.L.T., & di Palma, M.A. (2016): *robustbase: Basic robust statistics.* R package version 0.92-7.

Rousseeuw, P.J., & Leroy, A.M. (1987): *Robust regression and outlier detection*. New York: Wiley.

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.

Suh, S.C. (2012): *Practical data mining applications*. Sudbury: Jones & Bartlett Learning.

Vilalta, R., Giraud-Carrier, C., Brazdil, P., & Soares, C.: Using meta-learning to support data mining (2004): *International Journal of Computer Science & Applications,* 1, 31-45.

Víšek, J.Á. (2006): The least trimmed squares. Part I: Consistency. *Kybernetika*, 42, 1-36.

Tab. 1: Results of metalearning evaluated as the ratio of correctly classified cases in a leave-one-out cross validation study.

Number of features	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
10	0.10	0.43	0.43	0.43	0.57	0.67	0.33	0.62	0.62
9	0.33	0.52	0.57	0.52	0.71	0.71	0.33	0.62	0.76
8	0.43	0.52	0.67	0.48	0.76	0.76	0.33	0.71	0.86
7	0.48	0.62	0.67	0.52	0.67	0.81	0.28	0.67	0.86
6	0.48	0.67	0.67	0.48	0.76	0.86	0.33	0.76	0.76
5	0.48	0.71	0.67	0.43	0.67	0.81	0.28	0.76	0.81
4	0.48	0.71	0.67	0.33	0.67	0.81	0.33	0.67	0.81
3	0.48	0.71	0.76	0.43	0.67	0.86	0.38	0.71	0.86
2	0.48	0.71	0.71	0.43	0.76	0.86	0.38	0.71	0.86
1	0.48	0.57	0.71	0.33	0.67	0.81	0.38	0.71	0.71

Source: own computation

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

Data set	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
1 Aircraft	1	1	1	5	3	2	1	1	1
2 Ammonia	5	3	2	5	3	2	4	3	2
3 Cirrhosis	1	1	1	1	1	1	2	2	1
4 Coleman	5	3	2	3	2	1	2	2	1
5 Delivery	4	3	2	3	2	1	2	2	1
6 Education	2	2	1	5	3	2	5	3	2
7 Electricity	5	3	2	2	2	1	5	3	2
8 Employment	2	2	1	2	2	1	2	2	1
9 Houseprices	1	1	1	1	1	1	1	1	1
10 Imports	1	1	1	5	3	2	4	3	2
11 Kootenay	4	3	1	5	3	2	5	3	2
12 Livestock	3	2	2	4	3	2	5	3	2
13 Machine	1	1	1	3	2	1	5	3	2
14 Murders	2	2	1	3	2	1	1	1	1
15 Octane	1	1	1	1	1	1	3	2	1
16 Pasture	1	3	1	3	2	1	5	3	2
17 Pension	4	1	2	4	3	2	4	3	2
18 Petrol	1	1	1	5	3	2	5	3	2
19 Stars	1	1	1	2	2	1	4	3	2
20 Wood	2	2	2	2	2	1	4	3	2

Source: own computation

Contact

Barbora Peštová
The Czech Academy of Sciences, Institute of Computer Science
Pod Vodárenskou věží 2, 182 07, Praha 8, Czech Republic
pestova@cs.cas.cz

Jan Kalina
The Czech Academy of Sciences, Institute of Computer Science
Pod Vodárenskou věží 2, 182 07, Praha 8, Czech Republic
kalina@cs.cas.cz