Effect of Metalearning on Feature Selection Employment

Silvia Nunes das Dôres [1], Carlos Soares [2], Duncan Ruiz [1]

[1] Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brasil
[2] INESC TEC/Faculdade de Engenharia, Universidade do Porto, Porto, Portugal

Abstract. Feature Selection is important to improve learning performance, reduce computational complexity and decrease required storage. There are multiple methods for feature selection, with varying impact and computational cost. Therefore, choosing the right method for a given data set is important. In this paper, we analyze the advantages of metalearning for feature selection employment. This issue is relevant because a wrong decision may imply additional processing, when FS is unnecessarily applied, or in a loss of performance, when not used in a problem for which it is appropriate. Our results showed that, although there is an advantage in using metalearning, these gains are not yet sufficiently relevant, which opens the way for new research to be carried out in the area.

Keywords: feature selection, metalearning, metalearning evaluation

1 Introduction

Feature selection (FS) is a widely employed technique for reducing dimensionality [9]. It aims to select a subset of variables from the input data, that can efficiently describe these data while reducing effects from noise or irrelevant variables and still provide good prediction results, as well as lower computational cost and better model interpretability [5].

Recent works have approached the problem of FS method selection under the perspective of metalearning [1], showing interesting first results and paving the way for novel applications to be fully explored [11], [8], [4], [7], [6]. These works are based on the well-known no-free-lunch theorem [12] which assumes that there is no single FS algorithm that is always the best for all problems. In order to prevent the data miner to experiment with different methods using a trial and error approach, which can be time consuming and costly especially with very large data sets, metalearning is used for the automatic recommendation of the FS algorithm that is best suited for a given problem.

However, existing approaches on metalearning for FS method selection focus their evaluation at the meta-level, e.g. whether it is possible to guess which is the best FS method for a given data set. They ignore the magnitude of the gain/loss that the method yields in terms of base-level performance (e.g., what is the difference in classification performance of applying a learning algorithm with and without the selected FS method).

Thus, the purpose of this study is to extend the evaluation of metalearning approaches to FS method selection by taking into account the-base level performance.

We address two research questions. First, we investigate whether metalearning is useful for indicating when to use FS or not. The results show the advantages of using metalearning when compared to baselines as random choice and choice based on the majority class. In the second question, we analyze the gains in base-level performance associated with the use of metalearning. In this experiment, the results obtained by the metalearning strategy are very close to those obtained by an ideal meta-model (oracle), although they are also close to the recommendation based on the majority class.

This paper is organized as follows: Section 2 discusses related work. Section 3 shows our experimental analysis and results obtained. Section 4 presents our conclusions and future work.

2 Related Work

In the context of FS, the use of metalearning is relatively new and there are only a few studies that offer significant advances towards solving it [11], [8], [4], [7], [6].

Post et al. [7] presented a large scale experiment on the benefits of using FS for binary classification problems. They used 394 data sets, 12 classification algorithms and 1 FS method in two experiments: the first one investigates for which learning algorithms FS improves predictive performance, and the second experiment applies metalearning to recommend when to use FS for each learning algorithm. The first results pointed out the usefulness of FS for the different learning algorithms. In the second experiment the focus was the evaluation of different sets of meta-features under the behavior of metalearning strategy. The results show benefits of the use of metalearning, but the proposed strategy is not analyzed in terms of other recommendation strategies neither at the meta-level nor at the base-level, where the obtained performance gain is evaluated.

3 Metalearning for the Recommendation of Feature Selection Methods

In this paper, we intend to increase the quality of the FS employment in classification problems, extending the experiments carried out in [7] in order to evaluate the advantages of applying metalearning to decide when to employ FS, both at the meta-level and at the base-level. We address two research questions:

1. Is metalearning useful for indicating when to use FS or not?
2. What are the gains in base-level performance associated with the use of metalearning?

The experimental setup is detailed below.

3.1 Experimental Setup

As discussed earlier, we extend the experiments carried out in [7], since it is the most comprehensive study in the use of metalearning for FS method recommendation,

and also because it is the only one that uses metalearning in order to indicate when to use FS. For this purpose, we used the experimental results of the base-level provided by the authors in the OpenML platform to evaluate the performance gain of the proposed metalearning strategy. Further details about the data sets, algorithms, and meta-features can be obtained from OpenML [1].

Briefly, 394 data sets of binary classification problems were used, containing between 10 to 200,000 examples. One FS method (CfS-SubsetEval, a correlation-based method) and 12 machine learning algorithms (AdaBoost, HoeffdingTree, K-Nearest Neighbor, Logistic, Multilayer Perceptron, J48, JRip, Naive Bayes, Random Forest, REPTree, Stochastic Gradient Descent, Support Vector Machine) were applied to these data sets, using 10-fold-cross-validation. For each pair data set/learning algorithm the performance was evaluated with and without FS, based on the AUC measure, given the imbalanced distribution of classes in the data sets. At the meta-level, 40 meta-features were extracted from each data set (simple, statistical, information-theoretic, landmarkers and FS landmarkers). Every instance in the meta-database are two-fold cross validation runs on a algorithm, one run with FS and one run without FS, and the meta-target is a label indicating whether the run with FS had a better performance. As meta-learner was used the Random Forest algorithm (with 100 trees).

In this work, we employ a typical leave-one-out cross-validation procedure (LOO-CV) to evaluate the quality of the recommendation made by the metalearning strategy, both at the meta-level and at the base-level.

3.2 Meta-level Evaluation

For the first question, we carried out a set of experiments with the goal of providing a proof of concept in metalearning applied to FS recommendation. By testing whether the use of metalearning increases the quality of the decision between to use FS or not, we aim to show that using this strategy can avoid an *ad hoc* choice. We compare it with four distinct baselines: i) alwaysFS - a strategy that recommends the application of FS for all problems; ii) neverFS - a strategy that never recommends the use of FS; iii) random - a strategy that randomly selects whether to use FS or not; and iv) default - a strategy that selects the most frequent choice between using FS or not (majority class) in the data sets considered, for each algorithm. Table 1 shows the results of the experiment in terms of performance measured by AUC.

With these results, it is possible to see that metalearning shows a superior performance for all the learning algorithms. In ROC curves an ideal result is one that maximize the true positive rate while minimizing the false positive rate. As expected, the strategies alwaysFS, neverFS and default obtained an average performance, since they can not minimize the false positive rate. The random strategy has a balance between the true positive rate and false positive rate, which also leads to an average performance. Thus, metalearning stands out as being the best approach for decision making about the use of FS.

However, in a statistical validation we can see that, although there is a difference between the use of metalearning and the random strategy, this difference is not

[1] https://www.openml.org/s/15

Table 1. Performance at the Meta-Level

Algorithm	metalearning	alwaysFS	neverFS	random	default
AdaBoost	0.686	0.5	0.5	0.495	0.5
HoeffdingTree	0.661	0.5	0.5	0.504	0.5
K-NN	0.766	0.5	0.5	0.506	0.5
J48	0.761	0.5	0.5	0.522	0.5
Jrip	0.656	0.5	0.5	0.507	0.5
Logistic	0.719	0.5	0.5	0.498	0.5
MLP	0.716	0.5	0.5	0.537	0.5
NaiveBayes	0.728	0.5	0.5	0.492	0.5
REPTree	0.688	0.5	0.5	0.507	0.5
RandomForest	0.612	0.5	0.5	0.517	0.5
SGD	0.71	0.5	0.5	0.503	0.5
SVM	0.722	0.5	0.5	0.526	0.5
average	**0.702**	0.5	0.5	0.509	0.5

statistically significant. In contrast, the difference between using metalearning and alwaysFS, neverFS and default is statistically relevant. These tests were carried out using the methodology proposed by Demšar [3]: Friedman rank test with Nemenyi test for post-hoc multiple comparisons, and presented in Figure 1. Recommendation strategies are sorted by their average ranking (lower is better), and those connected by a horizontal line are statistically equivalent.

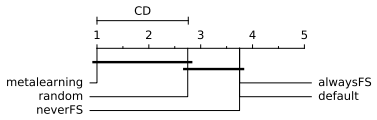

Fig. 1. Critical Difference diagram (with $\alpha = 0.05$) of the experiments at the meta-level.

3.3 Base-Level Evaluation

In this experiment, our goal is to evaluate the performance gain obtained through the use of metalearning for FS employment at base-level, i.e, directly on the performance of the learning algorithms. For this, metalearning was applied over the 12 meta-databases and, for each data set, the recommendation to use FS or not in a given data set was scored according to the performance obtained by the corresponding class at the base-level, in terms of AUC. As baselines, we use two approaches: i) default - a strategy that select the most frequent choice between using FS or not (majority class); and ii) oracle - an ideal meta-model that always selects the best option for each data set. Table 2 presents these results.

Results show a small advantage in the average performance of metalearning relative to the default strategy. Despite being small, the advantage is consistent for all algorithms. In addition, the results also show a small difference between the

Table 2. Performance Gain on Base-Level

Algorithm	oracle	metalearning	default
AdaBoost	0.873	0.868	0.867
HoeffdingTree	0.804	0.787	0.772
K-NN	0.814	0.796	0.795
J48	0.809	0.798	0.794
Jrip	0.805	0.796	0.796
Logistic	0.828	0.817	0.808
MLP	0.817	0.801	0.795
NaiveBayes	0.827	0.819	0.812
REPTree	0.802	0.783	0.776
RandomForest	0.885	0.881	0.877
SGD	0.769	0.763	0.756
SVM	0.767	0.762	0.756
average	0.817	0.806	0.800

performance of metalearning and the oracle, which shows a benefit in the application of the latter strategy of recommendation. The Critical Difference (CD) diagrams generated in the base-level experiments is presented in Figure 2.

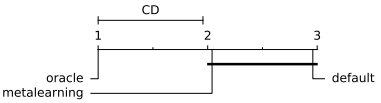

Fig. 2. Critical Difference diagram of the experiments at the base-level ($\alpha = 0.05$).

This diagram reflects the results presented in Table 2, indicating that there is no statistical difference between metalearning and the default recommendation (majority class) in terms of performance gain. In addition, we can also note that, although the average performance of metalearning is close the oracle, statistical analysis shows that this difference is significant.

However, when we take into account that the application of FS does not significantly improve the performance of the learning algorithms for most data sets [7], we can evaluate the gains obtained by the metalearning for problems in which the use of FS achieves greatest impact. Table 3 presents these results. For each learning algorithm, we considered the amount of data sets for which the application of FS get a gain above 5% in the final result and the percentage of metalearning hits on these significant sets.

Figure 3 presents details of these results for the K-NN and SGD algorithms. Since the instance-based algorithms are sensitive to the curse of dimensionality, it was expected that FS would bring performance gains to the K-NN, which was confirmed in the results. In general, the application of FS provided a performance increase greater than 5% in 154 data sets (about 39%), and among these, metalearning achieved a hit rate of 83.12%. In the case of the SGD algorithm, it was expected that its performance would not be affected by the number of features, so it did not

Table 3. Effect of metalearning on significant data sets

Algorithm	Frequency	Hit Rate
AdaBoost	15	26.70%
HoeffdingTree	127	60.62%
K-NN	154	83.12%
J48	63	58.73%
Jrip	42	19.04%
Logistic	99	56.56%
MLP	91	66.66%
Naive Bayes	114	70.17%
REPTree	119	53.78%
Random Forest	41	53.66%
SGD	72	55.55%
SVM	74	51.35%

have great advantage in the FS application. However, the results show that about 18% of the data sets benefited from the FS application, and that metalearning obtained a 55.5% hit rate on these data.

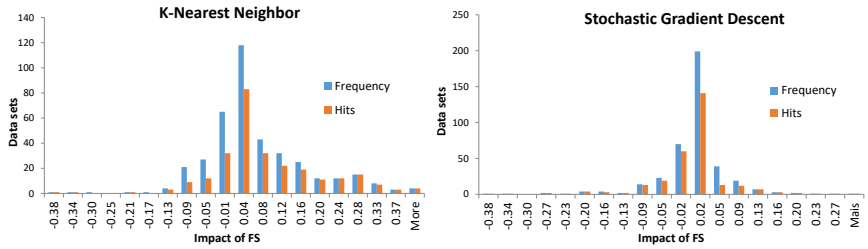

Fig. 3. Histogram of the hit rate hits of metalearning compared with the frequency with which FS impacts in the data sets.

4 Conclusions

In this work we evaluate the use of metalearning to indicate when to employ feature selection in conjunction with different classification algorithms. The results showed that both at the meta-level and at the base-level, metalearning outperforms the baselines analyzed, although in the statistical test the difference is not always significant. However, when we analyze the advantages of metalearning in the problems where FS causes the greatest impact in terms of performance, we can observe good results. Thus, suggestions for future works in this area indicate the possibility of extending the analysis to include new FS methods, in order to better generalize the results in relation to diversified inductive biases. In addition, we can develop new meta-features more appropriate for the problem.

Acknowledgments.

This work is financed by the ERDF European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013. The authors also would like to thank CAPES for the financial support through the project PDSE/88881.135787/2016-01.

References

1. Brazdil, P., Carrier, C.G., Soares, C., Vilalta, R.: Metalearning: Applications to data mining. Springer-Verlag Berlin Heidelberg (2008)
2. Chu, C., Hsu, A.L., Chou, K.H., Bandettini, P., Lin, C.: Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. 60, 59–70 (2012)
3. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research. 7, 1–30 (2006)
4. Filchenkov, A., Pendryak, A.: Datasets meta-feature description for recommending feature selection algorithm. In: Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference, pp. 11–18. IEEE, St. Petersburg (2015)
5. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of machine learning research. 3, 1157–1182 (2003)
6. Parmezan, A.R.S.,Lee, H.D.,Wu, F.C.: Metalearning for choosing feature selection algorithms in data mining: Proposal of a new framework. Expert Systems with Applications. 75, 1–24 (2017)
7. Post, M.J., van der Putten, P., van Rijn, J.N. Does Feature Selection Improve Classification? A Large Scale Experiment in OpenML. In: International Symposium on Intelligent Data Analysis, pp. 158–170. Springer, Stockholm (2016)
8. Shilbayeh, S., Vadera, S.: Feature selection in meta learning framework. In: Science and Information Conference, pp. 269–275. IEEE, London (2014)
9. Tang, J., Alelyani, S.,Liu, H.: Feature selection for classification: A review. Data Classification: Algorithms and Applications. 37, 1–33 (2014)
10. Vilalta, R. and Drissi, Y.: A perspective view and survey of meta-learning. Artificial Intelligence Review. 18, 77–95 (2002)
11. Wang, G.,Song, Q., Sun, H., Zhang, X., Xu, B., Zhou, Y.: A feature subset selection algorithm automatic recommendation method. Journal of Artificial Intelligence Research. 47, 1–34 (2013)
12. Wolpert, D. H., Macready, W.G.: No free lunch theorems for optimization. IEEE transactions on evolutionary computation. 37, 67–82 (1997)