Automated machine learning

Yevhen Kuzmovych
ČVUT - FIT
kuzmoyev@fit.cvut.cz

May 12, 2018

1 Introduction

One of the inalienable parts of the data analyst work is selection of the appropriate predictive algorithm for the given task. In most cases, this part comes down to testing set of selected algorithms on the given dataset or its subset and selecting the one with the best performance. This process can be automated and improved with prediction of algorithm quality.

Assuming that each dataset has some hidden properties that could indicate tendency of some algorithms to perform better than the others, it should be possible to extract those properties and predict algoritms' quality based on them.

There were many attempts that tried to select appropriate meta-features [3][1][2]. In the framework of this project, simply obtainable meta-featres used in the StatLog project[1] will be combined with landmarks and relative landmarks described in *Sampling-Based Relative Landmarks: Systematically Test-Driving Algorithms Before Choosing*[3] to predict algorithms quality.

2 Methods

2.1 Preprocessing

For this project, only required preprocessing techniques were used:

- **Filling missing data** (because most of the tested models require complete data). NaNs are filled with the means in numerical coulmns and most frequent values in categorical.
- **Encoding categorical data**. Nominal data is encoded using *one hot encoding*. Ordinal features can be specified with the needed order, labels are then encoded with natural numbers.

Implementation is parameterized and can be easely extended with the other techniques.

3 Outputs

4 Conclusion

References

[1] R. D. KING, C. FENG, and A. SUTHERLAND. Statlog: Comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence*, 9(3):289–333, 1995.

[2] Kate A. Smith-Miles. Cross-disciplinary perspectives on meta-learning for algorithm selection. *ACM Comput. Surv.*, 41(1):6:1–6:25, January 2009.

[3] Carlos Soares, Johann Petrak, and Pavel Brazdil. Sampling-based relative landmarks: Systematically test-driving algorithms before choosing. In *Proceedings of the10th Portuguese Conference on Artificial Intelligence on Progress in Artificial Intelligence, Knowledge Extraction, Multi-agent Systems, Logic Programming and Constraint Solving*, EPIA '01, pages 88–95, London, UK, UK, 2001. Springer-Verlag.