Automated machine learning

Yevhen Kuzmovych

ČVUT - FIT

kuzmoyev@fit.cvut.cz

May 13, 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 [4][2][3]. In the framework of this project, simply obtainable meta-featres used in the StatLog project[2] will be combined with landmarks and relative landmarks described in *Sampling-Based Relative Landmarks: Systematically Test-Driving Algorithms Before Choosing*[4] to predict algorithms quality.

2 Methods

Output library will be able to extract meta-features from the given data and build predictive pipeline using train data. Algorithm will estimate performance of each used model, test them in the estimated order, choose the best one and fit to the train data in a limited time.

Whole process consists of the following parts.

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.
- **Dropping constant columns**.
- **Scaling**. Standard scaling of numerical data.

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

2.2 Meta-data collection

Meta-features of the given dataset are collected for prediction of the quality of the used models. Collected meta-features are described in table 1.

2.3 Models evaluation and selection

In model quality evaluation there are two primary characteristics: **accuracy** and **processing time**. Considering this characteristics, models quality is evaluated using so-called *Adjusted Ratio of Ratios*(ARR) that combines models accuracy and processing time to assess relative performance among other models. *In ARR the compromise between the two criteria is given by user in the form "the amount of accuracy I'm willing to trade for a 10 times speed-up is X%"*[4]. So for two given algorithms i and j on the data set d the ARR computed as follows:

$$ARR_{ij}^d = \frac{\frac{A_i^d}{A_j^d}}{1 + \log\left(\frac{T_i^d}{T_j^d}\right) * X}$$

where A_i^d is accuracy of the model i on the data set d and T_i^d is its processing time.

Accuracy in classification problems computed simply as a ratio of the number of correctly classified examples to the number of total examples: $A_i^d = \frac{C}{N}$. Accuracy for regression problems is computed as:

$$A_i^d = 1 - \frac{RMSE_i}{\max_j RMSE_j}$$

Now using computed ARRs, we can generate realtive landmarks for each of n models:

$$rl_i^d = \frac{\sum_{j \neq i} ARR_{ij}^d}{n-1}$$

Simple	
NExamples	Number of examples
NFeatures	Number of features
NBinary	Number of binary features
NCategorical	Number of categorical features
NNumerical	Number of numerical features
NExamplesWithNANs	Number of examples with missing values
NFeaturesWithNANs	Number of features with missing values
NClasses	Number of classes (in classification)
Statistical	
STDRatio	Geometric mean of columns standard deviations
CorrelationMean	Mean of columns correlation values
KurtosisMean	Mean of columns kurtosis values
SkewnessMean	Mean of columns skewness values
YImbalance	STD of number of classes/bins of output column
YStd	STD of output column (in regression)
Relative landmarks	
rl i	Relative landmark of the model i
rl ...	

Figure 1: Collected meta-features

which is used to select the best model and as a meta-features on the subset of the task of size 500(chosen arbitrarily).

3 Possible improvements

As preprocessing is the most important part in any data analysis task[1], it is the first part that should be improved in automated data analysis. In framework of this project this part did not get deserving attention because of the lack of authors time, but should be considered as the primary optimisation point of the implemented algorithm.

4 Outputs

5 Conclusion

References

[1] Chicco D. Ten quick tips for machine learning in computational biology. *BioData Mining*, 2017. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5721660/`.

[2] 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.

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

[4] 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.