AutoML Benchmark

An open source extensible framework to evaluate a wide range of AutoML tools on tabular data

We present a benchmark framework to evaluate AutoML tools on a wide range of tabular datasets. The benchmark is completely open source and features an extensible design which makes it easy to add new AutoML tools or datasets. The framework is easy to use as it abstracts away the different interfaces to the automl tools, data loading and evaluation.

Benchmarking AutoML tools is difficult, since configuring and using them correctly is both laborious and a source of errors. We want to provide researchers and practictioners a reliable way to easily compare results and track progress.

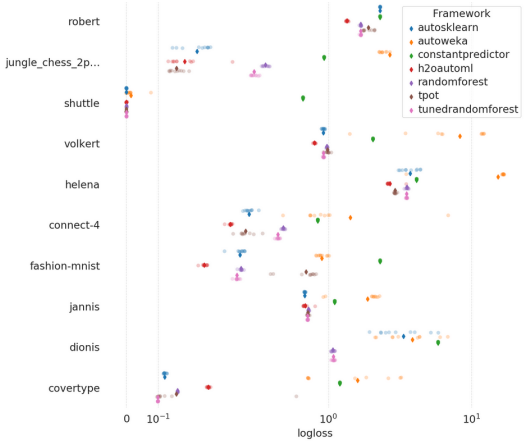

Figure 1. Results for 10-fold cross-validation on a 4 hour budget on various multi-class classification problems. Results obtained for the 2019 paper, none of the frameworks consistently outperformed a Random Forest within 4 hours.

13 Integrated Tools Already Integrated

Together with the AutoML authors we already integrated: AutoGluon, auto-sklearn, Auto-WEKA, autoxgboost, FLAML, GAMA, H2O, LightAutoML, MLNet, ML-Plan, mljarsupervised, mlr3automl and TPOT.

```
> python runbenchmark.py autosklearn openml/t/7952 -f 0
> python runbenchmark.py TPOT openml/s/269 4h8c -m aws
```

Figure 2. Two example invocations of the benchmark tool. The first line evaluates auto-sklearn on the first fold of the 'adult' dataset from OpenML. The second line evalutes TPOT on the whole automl benchmark with a 4 hour time budget and resource contraints on AWS.

Current Work

- Run new experiments for up-to-date results for classification and regression
- Provide a tool for interactive data analysis
- Perform a statistical analysis

Future Work

- Include other problem types (e.g. semi-supervised)
- Include other data types (e.g. string)
- Compare AutoML results to human performance

https://github.com/openml/automlbenchmark

Gijsbers, P., LeDell, E., Thomas, J., Poirier, S., Bischl, B., & Vanschoren, J. (2019). An open source automl benchmark. *arXiv preprint arXiv:1907.00909*