Regression Benchmark Datasets on OpenML

by Merlin Raabe and Philipp Probst

November 8, 2018

Abstract

Short description

We present a collection of regression datasets that is suitable for performing benchmarks. The datasets are chosen to provide different data scenarios including small and big datasets regarding observations and variables. The datasets are available on OpenML and tagged with XXX so that the download and usage of these datasets can be automatized. We provide an example with code where we download the datasets and make a little benchmark with some of the standard machine learning algorithms XXX.

1 Introduction

Machine learning has become to a widespread and frequently issued part of data science. The comunity of researchers keeps growing and there are many platforms that provide datasets and the possibility to share results which specific methods of machine learning achieved on them. Kaggle, PMLB, UCI and OpenML are probably the most common of them.

What the comunity is lacking are so called b̈enchmarking suitesẅhich are collections of datasets that are appropriate to test and compare the capabilities of methods of machine learning. Papers like `OpenML100` issue this topic and the authors created a benchmarking suite consisting of datasets which have been selected carefully. This benchmarking suite has been created for classification methods of machine learning.

texttttOpenML100 is one of more benchmarking suites for this kind of machine learning problem.

On the contrary for other kinds of machine learning problems like survival or regression there aren't that many suites. So it is also for problems of the range of regressionproblems. There are many datasets available on the platforms mentioned earlier but no benchmarking suites to compare results. That's why the authors of this paper colleccted datasets from those platforms and inserted them in a study called `PLACEHOLDER` on `OpenML`. At this stage it contains PLACEHOLDER datasets. For the selection of the datasets the authors defined hard criterias which are:

1. The dataset consists of at least 150 observations
2. The dataset provides at least 4 distinct features
3. The targetFeature consisits of at least 20 distinct numeric values
4. There are no missing values in the dataset
5. The dataset isn't a subset of or very similar to another dataset
6. The R-Squared calculated during a linear regression didn't reach 1

Several Methods of machine learning which are `K-nearest neighbors`, `Decision Tree`, `Random Forest`, `Elastic Net Regression` and `Support Vector Machine` have been executed on those datasets.

To compare them for each of them the average R-Squared and the average Kendalls Tau haven been calculated after training the methods with the use of 10-fold crossvalidation.

2 Literature Review

Here we cite some literature, e.g. Bischl et al. (2017).

Olson et al. (2017), Vanschoren et al. (2013)

3 Methods

4 The datasets

5 A little benchmark

«a, eval=TRUE, echo=TRUE»= a = 1 plot(a) @

6 Conclusion and Discussion

References

B. Bischl, G. Casalicchio, M. Feurer, F. Hutter, M. Lang, R. G. Mantovani, J. N. van Rijn, and J. Vanschoren. OpenML benchmarking suites and the OpenML100. *ArXiv preprint arXiv:1708.03731*, 2017. URL https://arxiv.org/abs/1708.03731.

R. S. Olson, W. La Cava, P. Orzechowski, R. J. Urbanowicz, and J. H. Moore. Pmlb: a large benchmark suite for machine learning evaluation and comparison. *BioData Mining*, 10(1):36, Dec 2017. ISSN 1756-0381. doi: 10.1186/s13040-017-0154-4. URL https://doi.org/10.1186/s13040-017-0154-4.

J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo. OpenML: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013.