OpenML: An Open Science Platform for Machine Learning

Jan N. van Rijn JVRIJN@LIACS.NL
Joaquin Vanschoren JOAQUIN@LIACS.NL
Leiden Institute of Advanced Computer Science, Leiden University, P.O. Box 9512, 2300 RA, Leiden, Netherlands

Keywords: Machine Learning, Databases, Meta-learning

Many machine learning studies have been conducted over the past few decades from which much knowledge has been obtained. However, due to space restrictions imposed on publications, these are only published in a highly summarized form (Vanschoren et al., 2012). It is scientifically important that the details of experiments are freely available to anyone for verifiability, reproducibility and generalizability. Therefore, we introduce a novel open science platform for machine learning research, called OpenML[1]. OpenML is a website where researchers can share all their datasets, algorithms and experiments, search for the results of others, and compare directly with the state of the art through controlled experimentation. Beyond the descriptions of algorithms in papers, OpenML allows researchers to share detailed experiments that are comparable with the results of other algorithms. Moreover, OpenML links all experimental results, and all meta-data of algorithms and datasets, for easy future analysis.

Users can define *tasks* which are well-described problems to be solved by a machine learning algorithm or workflow. A typical task would be: *Predict (target) attribute X of dataset Y with maximized predictive accuracy*. Other users are challenged to build algorithms that solve these tasks. The creation of tasks happens on the fly. Whenever a user is searching for tasks on which his algorithm can be run, the system automatically returns all tasks that are potentially of interest to the user. There exists excellent tools that facilitate controlled algorithm evaluation, such as MLComp[2] and Kaggle[3]. OpenML differs from these on key aspects: It is intended for sharing experiments and comparing research results, all information requisite for reproducing the experiments is openly available and the results are stored in a public, queriable database.

An attempt to solve a task is called a *run*. The server provides the input data and stores the output data for every algorithm. The algorithm is executed on the PC of the user. For some tasks, e.g., predictive tasks, it offers more structured input and output. For instance, a *supervised classification* task provides the folds with which a classifier can be trained and expects predictions for all input instances. The server evaluates the predictions and stores the scores for evaluation metrics. Also more general tasks can be defined, e.g., parameter optimization, feature selection and clustering.

We have developed a web API, which facilitates finding and downloading tasks and datasets and uploading implementations and results. This API will be integrated in various machine learning tools, like Weka, R, RapidMiner and KNIME. Given a task, the tools can automatically download all associated input data. Once executed, the result can be uploaded with just one click. For example, in the case of supervised classification tasks, the input consists of a dataset and the folds, and the result is a file containing the predictions.

For each algorithm in the database, an overview page will be generated containing data about all tasks on which this algorithm was run. This provides information about the performance of the algorithm over a potentially wide range of datasets, with various parameter settings. For each dataset a similar overview page is created, containing a ranking of algorithms that were run on tasks with the dataset as input.

[1] `http://www.openml.org/`, beta version
[2] `http://www.mlcomp.org/`
[3] `http://www.kaggle.com/`

Acknowledgments

This work is supported by grant 600.065.120.12N150 from the Dutch Fund for Scientific Research (NWO).

References

Vanschoren, J., Blockeel, H., Pfahringer, B., & Holmes, G. (2012). Experiment databases. A new way to share, organize and learn from experiments. *Machine Learning*, *87*, 127–158.