OpenML: A Collaborative Science Platform

Jan N. van Rijn[1], Bernd Bischl[2], Luis Torgo[3], Bo Gao[4], Venkatesh
Umaashankar[5], Simon Fischer[5], Patrick Winter[6], Bernd Wiswedel[6], Michael R.
Berthold[7], and Joaquin Vanschoren[1]

[1] Leiden University, Leiden, Netherlands, {jvrijn,joaquin}@liacs.nl
[2] TU Dortmund, Dortmund, Germany, bischl@statistik.tu-dortmund.de
[3] University of Porto, Porto, Portugal, ltorgo@inescporto.pt
[4] KU Leuven, Leuven, Belgium, bo.gao@cs.kuleuven.be
[5] Rapid-I GmbH, Dortmund, Germany, {venkatesh,fischer}@rapid-i.com
[6] KNIME.com AG, {patrick.winter,Bernd.Wiswedel}@knime.com
[7] University of Konstanz, Konstanz, Germany, Michael.Berthold@uni-konstanz.de

Abstract. We present OpenML, a novel open science platform that
provides easy access to machine learning data, software and results to
encourage further study and application. It organizes all submitted re-
sults online so they can be easily found and reused, and features a web
API which is being integrated in popular machine learning tools such as
Weka, KNIME, RapidMiner and R packages, so that experiments can be
shared easily.

Keywords: Experimental Methodology, Machine Learning, Databases,
Meta-Learning

1 Introduction

Research in machine learning and data mining can be speeded up tremendously
by moving empirical research results "out of people's heads and labs, onto the
network and into tools that help us structure and alter the information" [3].
The massive streams of experiments that are being executed to benchmark new
algorithms, test hypotheses or model new datasets have many more uses beyond
their original intent, but are often discarded or their details are lost over time.
In this paper, we present OpenML[8], an open science platform for machine learn-
ing research. OpenML is a website where researchers can share their datasets,
implementations and experiments in such a way that they can easily be found
and reused by others. It offers a web API through which new resources and re-
sults can be submitted, and is being integrated in a number of popular machine
learning and data mining platforms, such as Weka, RapidMiner, KNIME, and
data mining packages in R, so that new results can be submitted automatically.
Vice versa, it enables researchers to easily search for certain results (e.g., evalu-
ations on a certain dataset), to directly compare certain techniques against each
other, and to use all submitted data in advanced queries. An overview of the key
components of OpenML is provided in Figure 1.

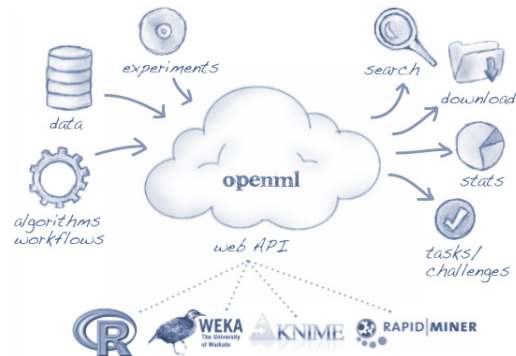[8] OpenML can be found at http://www.openml.org.

Fig. 1. Components of the OpenML platform

OpenML engenders a novel, collaborative approach to experimentation with important benefits. First, many questions about machine learning algorithms won't require the laborious setup of new experiments: they can be answered on the fly by querying the combined results of thousands of studies on all available datasets. OpenML also keeps track of experimentation details, ensuring that we can easily reproduce experiments later on, and confidently build upon earlier work [2]. Reusing experiments also allows us to run large-scale machine learning studies, yielding more generalizable results [1] with less effort. Finally, beyond the traditional publication of algorithms in journals, often in a highly summarized form, OpenML allows researchers to share all code and results that are possibly of interest to others, which may boost their visibility, speed up further research and applications, and engender new collaborations.

2 Sharing Experiments

To make experiments from different researchers comparable, OpenML uses *tasks*, 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 maximal predictive accuracy.* Somewhat similar to a data mining challenge, researchers are thus challenged to build algorithms that solve these tasks. Different tasks can be defined, e.g., parameter optimization, feature selection and clustering. They can be searched online, and will be automatically generated for newly submitted datasets. OpenML provides all necessary information to complete the task, such as a URL to download the input dataset, and what information should be submitted to the server. For some tasks, e.g., predictive tasks, it offers more structured input and output, such as exact train and test splits for cross-validation, and a submission format for all predictions. The server will then evaluate the predictions and store the scores for various evaluation metrics.

An attempt to solve a task is called a *run*, and includes the task itself, the algorithm or workflow (i.e., implementation) used, and a file detailing the obtained results. These are all submitted to the server, where new implementations

will be registered. Workflows are represented as a set of algorithms, and can be downloaded into the workbenches for detailed inspection. For each implementation, an overview page will be generated containing data about all tasks on which this algorithm was run. This will detail the performance of the algorithm over a potentially wide range of datasets, with various parameter settings. For each dataset, a similar page is created, containing a ranking of algorithms that were run on tasks with that dataset as input.

OpenML provides a RESTful API for downloading tasks and uploading datasets, implementations and results. This API is currently being integrated in various machine learning platforms such as Weka, R packages, RapidMiner and KNIME. For instance, in WEKA[9], OpenML is integrated as part of the Weka Experimenter. Given a task, it automatically downloads all associated input data, runs the experiment, and uploads the results to OpenML.

3   Searching OpenML

OpenML links various bits of information together in a single database. All results for different algorithms on the same tasks are stored in such a way that algorithms can directly be compared against each other (using various evaluation measures), and parameter settings are stored so that the impact of individual parameters can be tracked. Moreover, for all datasets, it calculates meta-data concerning their features (e.g., type, distinct values or mean and standard deviation) and their distributions, such as the number of features, instances, missing values, default accuracy, class entropy and landmarking results [4]. Likewise, for algorithms, it includes information about the (hyper)parameters and properties, such as whether the algorithm can handle nominal/numeric features, whether it can perform classification and/or regression and a bias-variance profile.

OpenML allows users to easily search for results of interest. First, it stores textual descriptions for datasets, algorithms and implementations so that they can be found through simple keyword searches, linked to overview pages that detail all related results. Second, runs can be searched to directly compare many algorithms over many datasets (e.g., for benchmarking). Furthermore, the database can be queried directly through an SQL editor, or through pre-defined advanced queries such as "Show the effect of a parameter P on algorithm A on dataset D" and "Draw the learning curve for algorithm A on dataset D".[10] The results of such queries are displayed as data tables, scatterplots or line plots, which can be downloaded directly.

4   Related Work

OpenML builds on previous work on *experiment databases* [5], but also enhances it by markedly facilitating the sharing of new experiments through the web API and by making results much easier to find and compare.

---

[9] A beta version can be downloaded from the OpenML website.
[10] See the 'Advanced' tab on `http://www.openml.org/search`.

In terms of sharing algorithms or workflows, it is somewhat similar to MyExperiment[11], an online repository where users can search and share workflows so that interesting workflows can be reused. However, MyExperiment offers little support for storing the *results* of workflows, or comparing workflows based on performance metrics.

On the other hand, MLComp[12] is a platform on which users can run their algorithms on known datasets (or vice versa). These runs are performed on the servers of MLComp, which saves the user resources. Although very useful, especially for comparing runtimes, OpenML differs from MLComp in two key aspects: First, OpenML allows users much more flexibility in running experiments: new tasks can easily be introduced for novel types of experiments and experiments can be run in any environment. It is also being integrated in data mining platforms that researchers already use in daily research. Finally, OpenML allows more advanced search and query capabilities that allow researchers to reuse results in many ways beyond direct comparisons, such as meta-learning studies [5].

5 Conclusions and Future Work

OpenML aims to engender an open, collaborative approach to machine learning research. Experiments can be shared in full detail, which generates a large amount of reproducible results available for everyone. Moreover, integration with popular data mining tools will make it very easy for researchers to share experiments with OpenML and the community at large.

Future work includes support for a broader range of task types, e.g., time series analyses, graph mining and text mining.

References

1. Hand, D.: Classifier technology and the illusion of progress. Statistical Science (Jan 2006)
2. Hirsh, H.: Data mining research: Current status and future opportunities. Statistical Analysis and Data Mining 1(2), 104–107 (Jan 2008)
3. Nielsen, M.: The future of science: Building a better collective memory. APS Physics 17(10) (2008)
4. Peng, Y., Flach, P., Soares, C., Brazdil, P.: Improved dataset characterisation for meta-learning. Lecture Notes in Computer Science 2534, 141–152 (Jan 2002)
5. Vanschoren, J., Blockeel, H., Pfahringer, B., Holmes, G.: Experiment databases. A new way to share, organize and learn from experiments. Machine Learning 87(2), 127–158 (2012), https://lirias.kuleuven.be/handle/123456789/297378

---

[11] http://www.myexperiment.org/

[12] http://www.mlcomp.org/