A New Data Corpus to Promote More Complete Autonomous Machine Learning Pipelines[*]

Richard Lippmann, Swaroop Vattam, Pooya Khorrami, and Cagri Dagli
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420
{lippmann, swaroop.vattam, pooya.khorrami, dagli@ll.mit.edu}

Abstract

Most past efforts on autonomous machine learning (AutoML) focused on selecting and tuning hyper-parameters of classification and regression algorithms using tabular datasets that have already been featurized. This automates only one stage in machine learning application pipelines. We describe the design, structure, schemas, and baselines created for a new dataset corpus that will enable the development and testing of AutoML tools that address other more time-consuming pipeline stages including: feature extraction from raw data, tidying or restructuring data that is sometimes in multiple files, and addressing common data quality issues. This corpus addresses a wide range of applications and many different types of machine learning problems. It includes more than 500 curated datasets with hand-generated baseline pipeline solutions, is being generated for the DARPA D3M program, includes online documentation, and is expected to be released to the machine learning community in the near future.

1 Introduction and Background

Fig 1 shows seven common stages present in machine learning pipelines that address real-world applications. Not all the tasks shown in each stage are present in all pipelines and, as shown by the arrows, each stage is not necessarily designed once, but may be modified and enhanced many times based on results from other stages. Although machine learning practitioners or data scientists perform most of these tasks, interactions with subject matter or domain experts are required for all stages to maintain good performance.

Most past competitions on Autonomous Machine Learning (AutoML) (e.g. [4]) have focused on stage five in Fig. 1 that involves selecting algorithms, tuning hyper parameters, training, and evaluating machine learning algorithms. One recent AutoML competition [10] has also begun to explore adaptation in stage seven. The vast majority of past work has used featurized, well structured, "tidy" data as described in [12]. Our experience, and that of many others (e.g. [6, 9, 12]), is that this stage typically consumes less than 20% of total time required to apply machine learning in new applications. The vast majority of time is typically spent understanding the problem, searching for and exploring data, labeling data, structuring and cleaning data, and selecting effective features. Realizing that automation of more complete pipeline stages is essential for widespread application of machine learning, our long-term goal is to create open datasets that can be used to develop and evaluate semi- or fully-autonomous versions of all pipeline stages. Our short-term goal is datasets that enable automation of pipeline stages four and five. The focus is to promote development of

[*] This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

open-source software that complements existing AutoML tools such as Auto-sklearn [3] and to extend data available from recent AutoML competitions [4] to cover more of the machine learning pipeline. All datasets have been developed and used under the DARPA D3M program [1, 7] and we plan to release the datasets in the near future as the program continues.

2 Dataset Components

Each dataset includes (1) A formatted directory structure and machine-readable schema that describes and contains the data, (2) a schema that describes the machine learning problem associated with the data, and (3) a runnable solution for the problem. Raw dataset files are stored as is, other datasets are provided as CSV files, and schemas are provided as JSON formatted files for machine readability and automatic verification. We provide coverage for many data types including image, video, audio, speech, text, graph, table, time series, and geospatial. Many types of machine learning problems are also supported including classification, regression, clustering, graph link prediction, graph vertex nomination, graph community detection, graph clustering, graph matching, time series classification, time series forecasting, collaborative filtering, and object boundary marking. The problem schema also supports 18 different performance metrics. Data and problem schemas have been developed over two years with feedback and testing during three formal evaluations and six major versions have been released. Extensive documentation, including schema specifications, is available online [8]. Runnable solutions are created by hand, often using scikit-learn components, but also using additional deep learning, graph, and other special libraries.

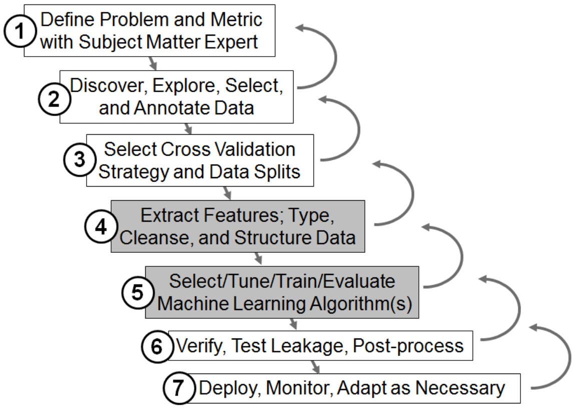
1 Define Problem and Metric with Subject Matter Expert
2 Discover, Explore, Select, and Annotate Data
3 Select Cross Validation Strategy and Data Splits
4 Extract Features; Type, Cleanse, and Structure Data
5 Select/Tune/Train/Evaluate Machine Learning Algorithm(s)
6 Verify, Test Leakage, Post-process
7 Deploy, Monitor, Adapt as Necessary

Figure 1: Seven steps commonly present in machine learning application pipelines.

3 Corpus Description

Level	Pipeline Steps Required	Description	Number
L0	5	Tabular, featurized, tidy	351
L1	4-5	Raw and tidy	216
L2	4-5	Raw or featurized and imperfect	19

A large set of diverse datasets was identified after prioritizing those where (1) a liberal open use license is provided, (2) a machine learning problem is defined for the data or is easy to define as in time series prediction, (3) truth labeling is available, and (4) a solution with baseline performance is already available either as a description or as running code. Three levels of datasets were created that required processing in the grayed-in stages four and five in Fig. 1. First, L0 datasets are tabular, featurized, tidy and fairly clean. Many of these datasets were selected from those available in OpenML [11] and the UCI Machine Learning Repository [2]. These datasets provide baseline performance for AutoML systems and require minimal data pre-processing. L1 datasets all include some type of raw data including different types of image, time series, graph, text, video, audio, geospatial, and tabular files. Note that raw data, including time series, log files, and event files, are often provided in tabular format, but that feature extraction is required on this tabular data to provide good performance. L1 datasets require feature extraction. Finally, L2 datasets are either raw or already featurized but are either poorly structured and not tidy or imperfect in other ways. L2 datasets sample five types of untidy data as defined in [12]. L2 datasets also sample eight data quality issues from those described in [5].

References

[1] Defense Advanced Research Projects Agency. Data-Driven Discovery of Models (D3M), June 2016. URL `https://www.fbo.gov/index?s=opportunity&mode=form&id=06049eb4ae38a68b7b8c54b94d9c7979&tab=core&_cview=1`.

[2] Dua Dheeru and Efi Karra Taniskidou. UCI Machine Learning Repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[3] M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, and F. Hutter. Practical Automated Machine Learning for the AutoML Challenge 2018. In *ICML 2018 AutoML Workshop*, July 2018.

[4] Isabelle Guyon, Lisheng Sun-Hosoya, Marc Boulle, Hugo Jair Escalante, Sergio Escalera, Zhengying Liu, Damir Jajetic, Bisakha Ray, Mehreen Saeed, Michele Sebag, Alexander Statnikov, Wei-Wei Tu, and Evelyne Viegas. *AUTOML: METHODS, SYSTEMS, CHALLENGES, Chapter 10. Analysis of the AutoML Challenge series 2015-2018*. Springer Series on Challenges in Machine Learning, 2018. URL `https://www.ml4aad.org/wp-content/uploads/2018/07/automl_book_draft_challenge.pdf`.

[5] Nuno Laranjeiro, Seyma Nur Soydemir, and Jorge Bernardino. A Survey on Data Quality: Classifying Poor Data. In *Dependable Computing (PRDC), 2015 IEEE 21st Pacific Rim International Symposium on*, pages 179–188. IEEE, 2015.

[6] Jimmy Lin and Dmitriy Ryaboy. Scaling Big Data Mining Infrastructure: The Twitter Experience. *ACM SIGKDD Explorations Newsletter*, 14:6–19, 04 2013.

[7] Richard Lippmann, William Campbell, and Joseph Campbell. An Overview of the DARPA Data Driven Discovery of Models (D3M) Program. In NIPS 2016 Workshop "Towards an Artificial Intelligence for Data Science", 2016. URL `http://workshops.inf.ed.ac.uk/nips2016-ai4datasci/papers/NIPS2016-AI4DataSci_paper_14.pdf`.

[8] Richard Lippmann, Swaroop Vattam, Pooya Khorrami, and Cagri Dagli. D3M-Schema, 9 2018. URL `https://github.com/mitll/d3m-schema/`.

[9] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden Technical Debt in Machine Learning Systems. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 2503–2511, Cambridge, MA, USA, 2015. MIT Press. URL `http://dl.acm.org/citation.cfm?id=2969442.2969519`.

[10] Wei-Wei Tu, Hugo Jair Escalante, Isabelle Guyon, Daniel L. Silver, Evelyne Viegas, Yuqiang Chen, Qiang Yang, Quanming Yao, Mengshuo Wang, Yuanyu Wan, and Hai Wan. NIPS 2018 Challenge: The 3rd AutoML Challenge: AutoML for Lifelong Machine Learning, December 2018. URL `https://www.4paradigm.com/competition/nips2018`.

[11] Jan N Van Rijn, Bernd Bischl, Luis Torgo, Bo Gao, Venkatesh Umaashankar, Simon Fischer, Patrick Winter, Bernd Wiswedel, Michael R Berthold, and Joaquin Vanschoren. OpenML: A Collaborative Science Platform. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 645–649. Springer, 2013.

[12] Hadley Wickham. Tidy Data. *The Journal of Statistical Software*, 59, 2014. URL `http://www.jstatsoft.org/v59/i10/`.