Reinforcement Learning for Data Cleaning and Data Preparation

Laure Berti-Équille
ESPACE-DEV/IRD, UMR 228, IRD/UM/UG/UR, Montpellier, France
Aix Marseille Université, Université de Toulon, CNRS, LIS, DIAMS, Marseille, France
laure.berti@ird.fr

ABSTRACT

Data cleaning and data preparation have been long-standing challenges in data science to avoid incorrect results, biases, and misleading conclusions obtained from "dirty" data. For a given dataset and a given analytics task, a plethora of data preprocessing techniques and alternative data cleaning strategies are available, but they may lead to dramatically different outputs with unequal ML model quality performances. For adequate data preparation, the users generally do not know how to start or which methods to use. Most current work focus either on proposing new data cleaning algorithms –often specific to certain types of data glitches considered in isolation and generally with no "pipeline vision" of the whole data preprocessing sequence– or on developing automated machine learning approaches (AutoML) that can optimize the hyper-parameters of the given ML model but that often rely on by-default preprocessing methods. We argue that more efforts should be devoted to proposing a principled data preparation approach to help and learn from the users for selecting the optimal sequence of data curation tasks and obtain the best quality performance of the final result. In this abstract, we present Learn2Clean[1], a method based on Q-Learning, a model-free reinforcement learning technique that selects, for a given dataset, a given ML model, and a quality performance metric, the optimal sequence of tasks for preprocessing the data such that the quality metric is maximized. Learn2Clean has been presented in The Web Conf 2019 [1] and we will discuss Learn2Clean enhancements for semi-automated data preparation guided by the user.

Motivations. In a 2017 survey conducted by the data science community Kaggle, "dirty data" is the common answer for 49.4% members responding to the questionnaire, when asked about the biggest barriers faced in data science[2]. As mitigation, data preparation often accounts for about 80% of the work of data scientists. As scientists, they would rather be deriving new knowledge and insights. The paradox is that without principled and adequate data preparation, those new insights are suspect at best. Data preparation and data cleaning are known to be very challenging tasks that require detection and elimination of a variety of complex data quality problems, such as duplicate records, inconsistent, missing, or outlying values. A wide range of methods for statistical analysis [9], constraint mining and checking [2], entity matching [5, 7], and machine learning [6, 8, 10] are used nowadays for data quality checks, data cleaning, and data repairing [4]. However, most existing systems suffer from important limitations. First, data preprocessing include many steps (e.g., normalization, transformation, encoding, discretization, imputation, deduplication, pattern or rule enforcement, feature selection/engineering, etc.) but current systems generally provide support for a limited number of steps in this pipeline. Second, for a given data preprocessing step, current systems generally rely on a few by-default methods. The users have to try and test the methods in order to select the most appropriate ones in a tedious and time-consuming process; to extend the system with another and eventually more adequate method, the users will have to write code or include other libraries, e.g., imputing missing values can be based on median, mean, most frequent values as default methods, but multiple imputations by chained equations (MICE) may be more accurate, but it is generally not included. Third, the systems do not recommend the adequate preprocessing methods for a given dataset and ML task, nor the execution sequence (or ordering) of the methods, e.g., imputation of missing values can be done after or before deduplication, and this may lead to a dramatically different preprocessed dataset. While some AutoML systems [3] can find the best model and configuration of hyper-parameters, the optimization has not been yet applied to the entire data preprocessing pipeline which remains fixed and based on by-default methods and orderings. Fourth, the users do not know which preprocessing methods can be applied to optimize the final results downstream. A solution would require executing all possible methods for each step of preprocessing, as well as all possible combinations and orderings of the methods. Finally, they may not have generic solutions from previous datasets or data cleaning tasks from which a system could learn a model to automate data preparation for any dataset specifics. For these reasons, we believe that data preparation is a novel application of reinforcement learning where model-free methods may be suitable for exploring data preparation space.

Human-In-the-Loop Computational Model. Learn2Clean is based on Q-learning which is a model-free reinforcement method that learns the transition probability $T(s_1|(s_0, a))$ from the pair of current state s_0 and action a to the next state s_1. In the context of data preparation, the transition probabilities and the dynamics of the system are not given a priori, and the space of possible states and actions is very large. Learn2Clean learns through trial-and-error experience in an unsupervised way. It explores from state to state until it reaches the goal (i.e., to maximize the user-defined quality metric). Each exploration is equivalent to one training session in which the system explores the data curation graph possibilities, receives the reward (if any) until it reaches the goal state. We have recently enhanced Learn2Clean approach to leverage user's input so that it can determine the next preprocessing action in conformance with the user's input and without waiting until the end of the episode (i.e., the entire cleaning pipeline). This makes the system agile and very performant compared to other data preparation strategies as it can progressively guide and be guided by the user during data preprocessing.

[1] https://github.com/LaureBerti/Learn2Clean
[2] https://www.kaggle.com/surveys/2017

REFERENCES

[1] L. Berti-Équille. Learn2Clean: Optimizing the Sequence of Tasks for Web Data Preparation. In *Proc. of the The Web Conf 2019*, 2019.

[2] W. Fan. Data quality: From theory to practice. *SIGMOD Record*, 44(3):7–18, 2015.

[3] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems 28*, pages 2962–2970. 2015.

[4] I. F. Ilyas and X. Chu. Trends in cleaning relational data: Consistency and deduplication. *Foundations and Trends in Databases*, 5(4):281–393, 2015.

[5] P. Konda, S. Das, P. S. G. C., A. Doan, A. Ardalan, J. R. Ballard, H. Li, F. Panahi, H. Zhang, J. F. Naughton, S. Prasad, G. Krishnan, R. Deep, and V. Raghavendra. Magellan: Toward building entity matching management systems. *PVLDB*, 9(12):1197–1208, 2016.

[6] S. Krishnan, M. J. Franklin, K. Goldberg, and E. Wu. Boostclean: Automated error detection and repair for machine learning. *CoRR*, abs/1711.01299, 2017.

[7] G. Navarro. Approximate string matching. In *Encyclopedia of Algorithms*, pages 102–106. 2016.

[8] T. Rekatsinas, X. Chu, I. F. Ilyas, and C. Ré. Holoclean: Holistic data repairs with probabilistic inference. *PVLDB*, 10(11):1190–1201, 2017.

[9] J. Schafer. *Analysis of incomplete multivariate data*. Chapman & Hall, 1997.

[10] M. Yakout, L. Berti-Équille, and A. K. Elmagarmid. Don't be scared: use scalable automatic repairing with maximal likelihood and bounded changes. In *Proc. of the ACM SIGMOD*, pages 553–564, 2013.