The Journal of Open Source Software

datacleanbot: an automated data cleaning tool

Ji Zhang[1]

1 Eindhoven University of Technology

DOI: 10.21105/joss.01608

Software

- Review
- Repository
- Archive

Submitted: 16 July 2019
Published: 08 August 2019

Summary

Data in real life almost never come in a clean way, and poor data quality may severely affect the effectiveness of learning algorithms (Sessions & Valtorta, 2006). Consequently, raw data need to be cleaned before being able to proceed with training or running machine learning models.

`datacleanbot` is a Python package which can offer automated, data-driven support to help users clean data effectively and smoothly. Given a random parsed raw dataset representing a supervised learning problem, `datacleanbot` is capable of automatically identifying the potential issues and reporting the results and recommendations to the end-user in an effective way. To be noticed, `datacleanbot` is aimed for supervised learning tasks and data need to be parsed as numeric format beforehand.

`datacleanbot` is equipped with the following capabilities:

- Present an overview report of the given dataset
  - The most important features
  - Statistical information (e.g., mean, max, min)
  - **Data types of features**
- Clean common data problems in the raw dataset
  - Duplicated records
  - Inconsistent column names
  - **Missing values**
  - **Outliers**

The three aspects `datacleanbot` meaningfully automates are marked in bold.

Data types of features: `datacleanbot` detects both basic data types (bool, date, float, integer and string) and statistical data types (real-valued, positive real-valued, count and categorical). `datacleanbot` detects basic data types by applying some logical rules. For statistical data types, `datacleanbot` utlizes the Bayesian model abda (Vergari et al., 2018).

Missing values: `datacleanbot` identifies characters 'n/a', 'na', '–', '?' as missing. Users can add extra characters to be identified as missing. After the missing values being detected, `datacleanbot` presents the missing data in effective visualizations with the help of missingno (Bilogur, 2018). Afterwards, `datacleanbot` recommends the appropriate approach to delete or impute missing values according to the missing mechanism of the given dataset.

Outliers: A meta-learner is trained beforehand to predict the optimal outlier detection algorithm for the given dataset. Then outliers are reported to users in various visualizations. Users can choose whether or not to drop the outliers.

`datacleanbot` has a strong connection to OpenML(Vanschoren, Rijn, Bischl, & Torgo, 2013), a platform where people can easily share data, experiments and machine learning models. Users can easily acquire data from OpenML and clean these data with the assistance of `datacleanbot`.

Acknowledgements

Many thanks to Dr. Joaquin Vanschoren for his dedication and guidance throughout this project in my master study.

References

Bilogur, A. (2018). Missingno: A missing data visualization suite. *The Journal of Open Source Software*, *3*(22), 547. doi:10.21105/joss.00547

Sessions, V., & Valtorta, M. (2006). The effects of data quality on machine learning algorithms. In *ICIQ*.

Vanschoren, J., Rijn, J. N. van, Bischl, B., & Torgo, L. (2013). OpenML: Networked science in machine learning. *SIGKDD Explorations*, *15*(2), 49–60. doi:10.1145/2641190.2641198

Vergari, A., Molina, A., Peharz, R., Ghahramani, Z., Kersting, K., & Valera, I. (2018). Automatic bayesian density analysis. *arXiv preprint arXiv:1807.09306*.