The Journal of Open Source Software

MOAFS: A Massive Online Analysis library for feature selection in data streams

Matheus Bernardelli de Moraes[1] **and André Leon Sampaio Gradvohl**[1]

1 Faculty of Technology, University of Campinas

DOI: 10.21105/joss.01970

Software

- Review
- Repository
- Archive

Editor: Viviane Pons

Reviewers:

- @sptennak
- @DARSakthi

Submitted: 15 October 2019
Published: 23 January 2020

Summary

Each feature selection algorithm performs efficiently depending on different circumstances, such as data dimensionality (low, medium, high or ultra), speed rate, attribute type (nominal or numerical), number of classes, among others. Therefore, to perform different experiments on some of the most relevant feature selection algorithms proposed for data streams classification problems, the Massive Online Analysis Feature Selection (`MOAFS`) was created. `MOAFS` is a library for the Massive Online Analysis (MOA) (Bifet, Holmes, Kirkby, & Pfahringer, 2010) framework, and it is based on the MOAReduction (Ramírez-Gallego, Krawczyk, García, Woźniak, & Herrera, 2017) extension. It contains seven feature selection algorithms to be used as dimensionality reduction techniques in data streams classification problems, especially in the text-domain field, since they are not directly available on MOA. `MOAFS` includes incremental versions of information-based filter algorithms such as Information Gain and Gain Ratio (Quinlan, 1986), Symmetrical Uncertainty used by the Fast Correlation-Based Filter (Yu & Liu, 2003), Chi-squared (Pearson, 1992) and Cramers V-Test (Cramer, 1946), as well as wrapper algorithms such as Online Feature Selection (Wang, Zhao, Hoi, & Jin, 2014) and Extremal Feature Selection (Carvalho & Cohen, 2006). Therefore, `MOAFS` is a package for MOA to perform feature selection in data streams classification problems.

Statement of need

Data streams are continuous, potentially unbounded, high volume and high dimensional data, which turns impracticable its storage in traditional database mechanisms (Ramírez-Gallego et al., 2017). For its properties, data streams have to be processed and analyzed online. However, as it is potentially unbounded, data streams probabilistic distribution changes over time, the so-called Concept Drift phenomenon. This phenomenon turns the online data process and analysis completely dynamic. Using classification algorithms is one approach to learn from data streams, as it classifies data into different classes for future decisions. Nevertheless, data streams high dimensionality imposes a challenge on the classification process, since it increases both computational cost and time, as well as aggravate the concept drift impacts. To solve this problem, online feature selection algorithms have been proposed to reduce data streams dimensionality by removing irrelevant and redundant attributes from data streams. As shown in the figure below, feature selection methods affect the classification process in multiple forms. Using the appropriate method is an important step in the learning process, especially when handling the concept drift phenomenon.

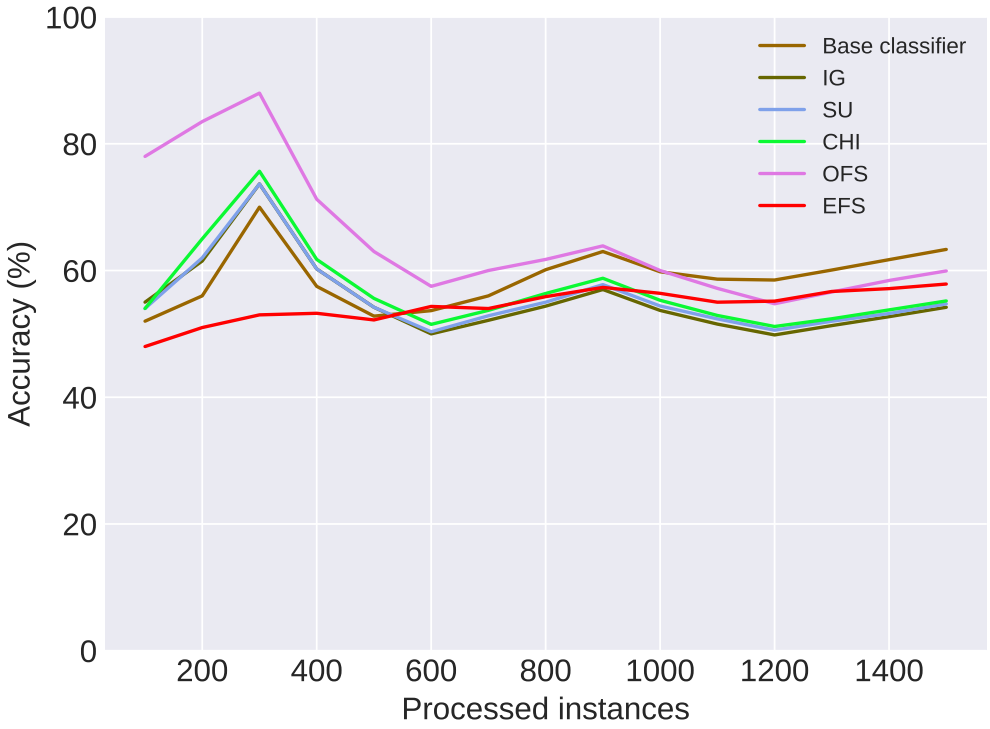

Figure 1: Accuracy over time.

Library design

Using a feature selection method is optional in `MOAFS`. If no method is defined, the base classifier (Naïve Bayes) performs classification in a standard approach, using all available features. On the other hand, if a feature selection algorithm is selected, it determines the most relevant features before the classification step. Then, the method transfers the selected subset of features to the base classifier which then performs classification using only the given subset of features. The figure below presents a flowchart example of this process. Therefore, with `MOAFS`, it is possible to verify which feature selection method is more suitable under a specific scenario.

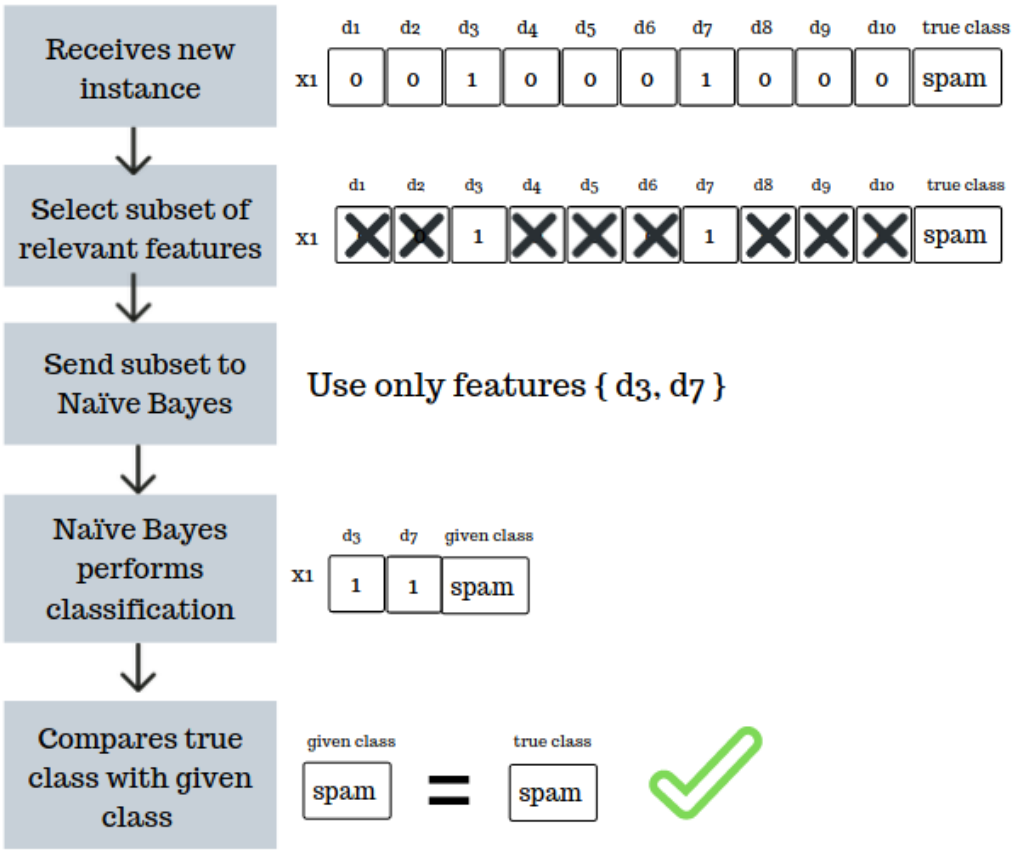

Figure 2: MOAFS flowchart example.

Acknowledgements

This work was supported by the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES).

References

Bifet, A., Holmes, G., Kirkby, R., & Pfahringer, B. (2010). MOA: massive online analysis. *Journal of Machine Learning Research*, *11*, 1601–1604.

Carvalho, V. R., & Cohen, W. W. (2006). Single-pass online learning: Performance, voting schemes and online feature selection. KDD '06, 548–553. doi:10.1145/1150402.1150466

Cramer, H. (1946).In *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

Pearson, K. (1992). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics: Methodology and distribution* (pp. 11–28). New York, NY: Springer New York. doi:10.1007/978-1-4612-4380-9_2

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, *1*(1), 81–106. doi:10.1023/A:1022643204877

Ramírez-Gallego, S., Krawczyk, B., García, S., Woźniak, M., & Herrera, F. (2017). A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing*, *239*, 39–57. doi:10.1016/j.neucom.2017.01.078

Wang, J., Zhao, P., Hoi, S. C. H., & Jin, R. (2014). Online feature selection and its applications. *IEEE Transactions on Knowledge and Data Engineering*, *26*(3), 698–710. doi:10.1109/TKDE.2013.32

Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In T. Fawcett & N. Mishra (Eds.), (Vol. 2, pp. 856–863). Presented at the Twentieth International Conference on Machine Learning, Washington: Scopus.