SOUR: an Outliers Detection Algorithm in Learning to Rank (Abstract)

Federico Marcuzzi[1,*], Claudio Lucchese[1] and Salvatore Orlando[1]

[1]*Università Ca' Foscari Venezia, Via Torino, 155, 30170 Mestre, Venezia VE, Italy*

Keywords
Information Retrieval, Learning to Rank, Machine Learning

Outlier data points are known to affect negatively the learning process of regression or classification models, yet their impact in the learning-to-rank scenario has not been thoroughly investigated so far. In this talk we present our effort to solve this research problem. The full version of this work will appear at ICTIR 2022 [1]. We designed SOUR, a learning-to-rank method that detects and removes outliers before building an effective ranking model. We limit our analysis to gradient boosting decision trees, but our algorithm can be easily adapted to handle different learning strategy, such as artificial Neural Network. SOUR searches for outlier instances that are consistently incorrectly ranked in several consecutive iterations of the learning process. We performed an extensive evaluation analysis on three publicly available datasets and we empirically demonstrated that *i)* removing a limited number of outlier data instances before re-training a new model, provides statistically significant improvements in term of effectiveness *ii)* SOUR outperforms state-of-the-art de-noising and outlier detection methods such as [2]. Finally, we investigated how the removal of the outliers affects the ensemble structure and we found that the ensemble leaves were purer when trained without the presence of the outliers.

References

[1] F. Marcuzzi, C. Lucchese, S. Orlando, Filtering out outliers in learning to rank, in: Proceedings of the 2022 12th International Conference on the Theory of Information Retrieval, ICTIR 2022, Madrid, Spain, July 11-12, 2022, 2022.

[2] X. Wu, Q. Liu, J. Qin, Y. Yu, Peerrank: Robust learning to rank with peer loss over noisy labels, IEEE Access 10 (2022) 6830–6841. URL: https://doi.org/10.1109/ACCESS.2022.3142096. doi:10.1109/ACCESS.2022.3142096.

IIR 2022: 12th edition of Italian Information Retrieval Workshop, June 29-30, 2022, Milan, Italy
*Corresponding author.
federico.marcuzzi@unive.it (F. Marcuzzi); claudio.lucchese@unive.it (C. Lucchese); orlando@unive.it (S. Orlando)
0000-0002-8141-8294 (F. Marcuzzi)