Effective, Efficient, and Robust Learning Algorithms for Ranking and Classification

Federico Marcuzzi
Università Ca' Foscari di Venezia
Italy
federico.marcuzzi@unive.it

Abstract

Over the past decade, machine learning has gained significant traction and is now deployed across diverse domains, including information systems, finance, healthcare, cybersecurity, autonomous driving, and more. As machine learning finds applications in various sensitive scenarios, the demand for models that exhibit accuracy and robustness during the operational phase has grown exponentially. One crucial factor that profoundly shapes the quality of machine learning models revolves around the training data they rely upon and the input data encountered at the operational phase. Therefore, the development of data-aware algorithms is of paramount importance in achieving high-quality machine-learning models. This thesis contributes to this overarching objective by delving into the development of data-aware algorithms, emphasizing the importance of this awareness during both the training and operational phases of machine learning models. The research presented in this thesis focuses on two primary domains. The first domain is information retrieval, with a particular emphasis on enhancing both the efficiency of learning-to-rank learning algorithms and the effectiveness of the learned models in solving ranking tasks. The thesis includes three works in this domain: Marcuzzi et al. [2022] provides a novel algorithm to detect and remove *consistent-outliers* documents from the training data. In Marcuzzi et al. [2023], we designed a new learning algorithm that handles the problem of gradient incoherencies affecting LambdaRank-based algorithms. Finally, in Lucchese et al. [2023], we designed a new sampling function for the *Selective Gradient Boosting* algorithm to exploit the most useful low-ranked non-relevant document. The second domain is adversarial machine learning, which focuses on increasing the robustness of binary classifiers against adversarial inputs encountered at the operational phase. Furthermore, the research in this domain focuses on providing certifiable models to efficiently assess robustness against adversarial machine learning attacks. In this regard, in Calzavara et al. [2021], we designed a novel robust learning algorithm to train ensembles of decision trees robust to evasion attacks along with its polynomial robustness-certification algorithm designed to compute a robustness lower bound. Finally, in Calzavara et al. [2022], we provided a new evaluation metric named *Resilience* to better access the security of machine learning models.

Awarded by: Università Ca' Foscari di Venezia, Venice, Italy **on** 19 April 2024.
Supervised by: Claudio Lucchese.
Available at: https://federicomarcuzzi.github.io/resources/thesis_phd.pdf.

Selected Publications

Stefano Calzavara, Claudio Lucchese, Federico Marcuzzi, and Salvatore Orlando. Feature partitioning for robust tree ensembles and their certification in adversarial scenarios. *EURASIP J. Inf. Secur.*, 2021(1):12, 2021. doi: 10.1186/S13635-021-00127-0. URL https://doi.org/10.1186/s13635-021-00127-0.

Stefano Calzavara, Lorenzo Cazzaro, Claudio Lucchese, Federico Marcuzzi, and Salvatore Orlando. Beyond robustness: Resilience verification of tree-based classifiers. *Comput. Secur.*, 121:102843, 2022. doi: 10.1016/J.COSE.2022.102843. URL https://doi.org/10.1016/j.cose.2022.102843.

Claudio Lucchese, Federico Marcuzzi, and Salvatore Orlando. On the effect of low-ranked documents: A new sampling function for selective gradient boosting. In Jiman Hong, Maart Lanperne, Juw Won Park, Tomás Cerný, and Hossain Shahriar, editors, *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing, SAC 2023, Tallinn, Estonia, March 27-31, 2023*, pages 646–652. ACM, 2023. doi: 10.1145/3555776.3577597. URL https://doi.org/10.1145/3555776.3577597.

Federico Marcuzzi, Claudio Lucchese, and Salvatore Orlando. Filtering out outliers in learning to rank. In Fabio Crestani, Gabriella Pasi, and Éric Gaussier, editors, *ICTIR '22: The 2022 ACM SIGIR International Conference on the Theory of Information Retrieval, Madrid, Spain, July 11 - 12, 2022*, pages 214–222. ACM, 2022. doi: 10.1145/3539813.3545127. URL https://doi.org/10.1145/3539813.3545127.

Federico Marcuzzi, Claudio Lucchese, and Salvatore Orlando. Lambdarank gradients are incoherent. In Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos, editors, *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pages 1777–1786. ACM, 2023. doi: 10.1145/3583780.3614948. URL https://doi.org/10.1145/3583780.3614948.