Evento	Salão UFRGS 2020: SIC - XXXII SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2020
Local	Virtual
Título	Improving the Identification of Safe Behaviors in Machine Learning Applications
Autor	ALINE WEBER
Orientador	BRUNO CASTRO DA SILVA

Improving the Identification of Safe Behaviors in Machine Learning Applications

Aline Weber (Prof. Bruno Castro da Silva)
Federal University of Rio Grande do Sul

Machine learning algorithms have been used in many real-life applications, but often produce dangerous solutions—e.g., solutions that discriminate. Recent work[1] introduced a framework for designing safe machine learning algorithms. It allows users to specify arbitrary definitions of undesirable behaviors. This framework has been used, for instance, to construct a safe regression algorithm capable of predicting student performance while avoiding gender-based discrimination. The safe regression algorithm has two mechanisms: Candidate Selection and Safety Test. The former identifies a candidate solution that optimizes the primary objective (i.e., to accurately predict grades) and that is likely to be considered safe. The latter uses statistical analyses to verify that a candidate solution is indeed safe. The originally-proposed method for Candidate Selection, however, is often over-confident that a solution will pass the Safety Test. We propose a novel method to identify candidate solutions that are likely to be considered safe. We analyze high-confidence bounds on a random variable that represents the success rate of the safety test procedure, given an approximation of the dataset distribution from which the training data was drawn. Our approach eliminates the requirement of manually setting a key hyperparameter of the original method. We show that our approach achieves performance superior to the originally-proposed algorithm, thereby more frequently identifying safe solutions to real-life machine learning problems.

[1] THOMAS, Philip S.; SILVA, Bruno Castro da; BARTO, Andrew G.; GIGUERE, Stephen; BRUN, Yuriy; BRUNSKILL, Emma. Preventing undesirable behavior of intelligent machines. **Science**, [S.L.], v. 366, n. 6468, p. 999-1004, 21 nov. 2019.