Adversarial classification using signaling games with an application to phishing detection

Figueroa, N., L'Huillier, G., & Weber, R. (2017). Adversarial classification using signaling games with an application to phishing detection. Data mining and knowledge discovery, 31(1), 92-133. https://doi.org/10.1007/s10618-016-0459-9

Abstract

In adversarial classification, the interaction between classifiers and adversaries can be modeled as a game between two players. It is natural to model this interaction as a dynamic game of incomplete information, since the classifier does not know the exact intentions of the different types of adversaries (senders). For these games, equilibrium strategies can be approximated and used as input for classification models. In this paper we show how to model such interactions between players, as well as give directions on how to approximate their mixed strategies. We propose perceptron-like machine learning approximations as well as novel Adversary-Aware Online Support Vector Machines. Results in a real-world adversarial environment show that our approach is competitive with benchmark online learning algorithms, and provides important insights into the complex relations among players.

Keywords: Adversarial classification||Signaling games||Perfect Bayesian equilibrium||Incremental learning||Support vector machines||Cybercrime||Phishing filtering
Creado: Domingo, 22 de Noviembre, 2020