université de BORDEAUX

Résumé d'article

MONTIEL Diane
PORTEJOIE Sarah
POURTIER Jacques

Enseignants
M. BOUSSICAULT Adrien
M. HOFER Ludovic
M. NARBEL Philippe

1 Adversarial Machine Learning in Network Intrusion Detection Systems

Le travail de l'article que nous avons choisi découle d'un besoin d'améliorer la robustesse des systèmes de machine learning. Jusqu'à présent, seule la performance des modèles générés par le machine learning importait. Il a été démontré que ces modèles étaient vulnérables aux attaques contradictoires.

L'étude de cet article porte sur le comportement d'un système de détection d'intrusion de réseau (NIDS) face à des exemples contradictoires comme données.
L'approche traditionnelle pour réaliser un NIDS repose sur la mise en place de règles et de la définition de ce que doit être un flux normal de paquets et une intrusion. Le problème de cette approche se trouve dans la difficulté à détecter et de se protéger de ces nouvelles intrusions non répertoriées précédemment. L'utilisation de machine learning permet ainsi de faciliter l'automatisation de tout ces processus.

Les modèles de machines learning étant donc devenus indispensables pour les NIDS, les chercheurs ont alors simulé une attaque boîte blanche pour évaluer différents modèles de machine learning. Cela leur a permis de comparer les performances sur un éventail de systèmes possibles. Pour générer des exemples de problèmes contradictoires, les chercheurs ont utilisé trois différentes méthodes de perturbations :

— PSO : Particle Swarm Optimization
— GA : Genetic Algorithm
— GAN : Generative Adversarial Network

Le résultat de ces méthodes ont été comparés avec la méthode de simulation de Monte-Carlo.

Ces algorithmes ont été utilisés sur des jeux de données connus qui sont `NSL-KDD`, et `UNSW-NB15`. `NSL-KDD` est une version altérée du dataset `KDD99`. Ces 2 datasets contiennent le trafic d'un réseau composé d'un mix entre trafic normal et infecté. Seuls les vecteurs étiquetés comme malicieux ont été retenu pour l'évaluation des méthodes. Ces deux datasets sont disponibles publiquement.

L'expérience a été entièrement développée en Python grâce aux bibliothèques Scikit-learn et Keras. Cette dernière n'ayant été utile qu'à l'implémentation du GAN, nécessitant un réseau de neurone.

Les modèles de machine learning testés durant cette étude sont aux nombre de onze : le machine à vecteur de support, l'arbre de décision, la classification naïve Bayésienne, les K plus proches voisins, la forêt d'arbre décisionnels, le Perceptron multicouche, le Gradient Boosting, la régression logistique, l'analyse discriminante linéaire, l'analyse du discriminant quadratique ainsi que le bootstrap aggregating.

En observant les résultats, on peut voir que pour le dataset `NSL-KDD` la méthode de perturbation trompant le mieux les modèles est le PSO, à un taux de 99,99 %.
Pour le `UNSW-NB15` il s'agit du GA à un taux de 100 %.
On peut rajouter qu'il n'y a eu aucun réglage des hyperparamètres pour les classifieurs dans l'intérêt que ceux-ci s'adaptent à n'importe quelles métriques possibles.
Les chercheurs en ont conclu qu'aucun modèle n'est immunisé au machine learning contradictoire. Cependant, en vue des résultats obtenus, ils déconseillent d'utiliser les méthodes SVM et DT au sein de NIDS. Ils cherchent dorénavant à étudier le fonctionnement interne des modèles de machine learning utilisés.