Génération de trafic adverse malveillant, guidée par l'explicabilité

Gregory Blanc, Houda Jmila

Thématiques — Cybersécurité, Explicabilité de l'IA, Génération d'attaques adverses.

Contexte

L'utilisation des algorithmes d'apprentissage automatique (en anglais, *Machine Learning* ou *ML*) pour la détection d'intrusions est une technique répandue [1]. Cependant, ces dernières années, les recherches ont montré la vulnérabilité de ces algorithmes à ce que l'on appelle les "attaques contradictoires" (*adversarial attacks* [1]), où un échantillon d'une classe est légèrement perturbé afin d'être classifié dans une autre classe. Dans le cadre du contournement d'un détecteur d'intrusion (*Intrusion detection System* ou *IDS*), un échantillon de trafic malveillant peut ainsi se faire passer pour du trafic bénin, en induisant en erreur l'IDS [2].

Heureusement, la génération d'un trafic contradictoire de manière réaliste n'est pas une tâche évidente. En effet, la majorité de l'état de l'art s'est restreint à la génération de vecteurs de caractéristiques (*features*) représentant le trafic adverse, plutôt que du trafic réel, c'est à dire qui peut véritablement circuler sur le réseau informatique [3]. Par conséquent, la principale limite de ces méthodes est que les valeurs générées pour les features peuvent ne pas correspondre à ce qui est observé dans un trafic réel.

Récemment, Vitorino *et al.* [3] ont proposé de remédier à cette problèmatique en définissant des intervalles de validité pour les features et générer des valeurs de features dans ces dits intervalles, afin de s'assurer du réalisme de l'échantillon généré [2].

D'un autre côté, l'explicabilité des algorithmes de l'IA permet de déterminer les features les plus importantes pour la prise de décision de l'IDS [4]. Nous pourrions ainsi générer du trafic contradictoire en se limitant à la manipulation/perturbation des features les plus importantes, au lieu de perturber toutes les features d'un échantillon. Cela réduirait le coût et le temps de génération du trafic contradictoire. C'est ce que nous proposons d'explorer dans ce sujet de recherche. Deux références inspirantes sont les travaux de Sun *et al.* [5] et de Rosenberg *et al.* [6].

Déroulement

1. Etat de l'art : se familiariser avec les notions de :
   — la détection des anomalies à l'aide de l'IA,
   — les attaques contradictoires,
   — l'explicabilité de l'IA (XAI).

1. Aussi intitulées *attaques par exemples contradictoires* `https://www.cnil.fr/fr/definition/attaque-par-exemples-contradictoires-adversarial-examples-attack`

2. Les auteurs partagent d'ailleurs l'outil de génération du trafic qu'ils ont conçu (A2PM). `https://github.com/vitorinojoao/a2pm`

2. Réalisation :
   — Outils :
      — Pour la génération du trafic adverse : prendre la main sur l'outil A2PM [7].
      — Pour l'explicabilité : prendre la main sur un ou plusieurs outils d'explicabilité (e.g. SHAP [8])
      — Base de données (choix entre CICIDS [9], BotIoT [10], UNSW [11] etc.)
   — Feuille de route :
      (a) Sur la base de données choisie, déterminer les features les plus importants à l'aide de la méthode d'explicabilité.
      (b) Mener une réfelexion sur la définition des "intervalles de réalisme" des features.
      (c) Génération du trafic adverse en perturbant les features dans les intervalles définis précédemment, à l'aide de l'outil [7].
   — Output attendus :
      — Rapport
      — Code source de l'implémentation

Références

[1] Hongyu Liu et Bo Lang : Machine learning and deep learning methods for intrusion detection systems : A survey. *applied sciences*, 9(20):4396, 2019.

[2] Houda Jmila et Mohamed Ibn Khedher : Adversarial machine learning for network intrusion detection : A comparative study. *Computer Networks*, page 109073, 2022.

[3] João Vitorino, Nuno Oliveira et Isabel Praça : Adaptative perturbation patterns : Realistic adversarial learning for robust nids. *arXiv preprint arXiv :2203.04234*, 2022.

[4] Subash Neupane, Jesse Ables, William Anderson, Sudip Mittal, Shahram Rahimi, Ioana Banicescu et Maria Seale : Explainable intrusion detection systems (x-ids) : A survey of current methods, challenges, and opportunities. *arXiv preprint arXiv :2207.06236*, 2022.

[5] Ruoxi Sun, Wei Wang, Tian Dong, Shaofeng Li, Minhui Xue, Gareth Tyson, Haojin Zhu, Mingyu Guo et Surya Nepal : Measuring vulnerabilities of malware detectors with explainability-guided evasion attacks. *arXiv e-prints*, pages arXiv–2111, 2021.

[6] Ishai Rosenberg, Shai Meir, Jonathan Berrebi, Ilay Gordon, Guillaume Sicard et Eli Omid David : Generating end-to-end adversarial examples for malware classifiers using explainability. *In 2020 international joint conference on neural networks (IJCNN)*, pages 1–10. IEEE, 2020.

[7] Adaptative perturbation pattern method. `https://github.com/vitorinojoao/a2pm`.

[8] Shap documentation. `https://shap.readthedocs.io/en/latest/index.html`.

[9] Iman Sharafaldin, Arash Habibi Lashkari et Ali A Ghorbani : Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, 1:108–116, 2018.

[10] Nickolaos Koroniotis, Nour Moustafa, Elena Sitnikova et Benjamin Turnbull : Towards the development of realistic botnet dataset in the internet of things for network forensic analytics : Bot-iot dataset. *Future Generation Computer Systems*, 100:779–796, 2019.

[11] Nour Moustafa et Jill Slay : Unsw-nb15 : a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). *In 2015 military communications and information systems conference (MilCIS)*, pages 1–6. IEEE, 2015.