VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

PROGRAMŲ SISTEMŲ STUDIJŲ PROGRAMA

Bakalauro baigiamasis darbas

Varžymosi principais grįstų atakų aptikimas naudojant paaiškinamo dirbtinio intelekto metodą kenkėjiškų programų kontekste

Defense Against Adversarial Malware Obfuscation Attacks Using Explainable Artificial Intelligence

Liudas Kasperavičius

Darbo vadovas : prof. dr. Olga Kurasova

Recenzentas : assoc. prof. Linas Petkevičius

Vilnius
2025

Turinys

Terminų žodynas

Maišymo funkcija. Tai funkcija $f : \{0,1\}^* \to \{0,1\}^m$. Naudojama, kai iš begalinės įvesčių erdvės norima gauti fiksuoto dydžio (m) išvestį 3

Pėdsakas. Programos struktūros ir požymių santrauka, beveik unikaliai identifikuojanti programą (pvz., maišymo funkcija) 5

Sprendimų priėmimo riba *(angl. Decision Boundary)*. Paprasčiausiems ML modeliams tai yra kreivė plokštumoje. Sudėtingesniems – daugiadimensiniams modeliams – daugdara *(angl. manifold)* 5

Varžymosi principais pagrįstos atakos *(angl. Adversarial Attacks)*. Tai atakos, pritaikytos „apgauti“ ML klasifikatorius 5

Santrumpos

AE. varžymosi principais pagrįstomis atakomis obfuskuoti kenkėjiško kodo pavyzdžiai 5

DI. dirbtinis intelektas 5

ML. mašininis mokymasis 3, 5

Įvadas

Pastaraisiais metais kenkėjiškas kodas ir programos kuriamos itin sparčiai ($\sim$450000 kenkėjiškų programų per dieną 2024 m. AV-TEST[1] duomenimis). Kenkėjiško kodo aptikimo programos, kurios tradiciškai remiasi programų pėdsakais, nespėja atnaujinti pėdsakų duomenų bazių pakankamai greitai. Dėl to DI, tiksliau mašininio mokymosi (ML), naudojimas kenkėjiškų programų ar kenkėjiško kodo aptikimo srityje tapo itin populiarus [DCB+21]. Tačiau ML modeliai, nors ir geba aptikti kenkėjiškas programas iš naujų, dar nematytų, duomenų, yra pažeidžiami varžymosi principais pagrįstoms atakoms [CSD19; HT17; RSR+18; ZHZ+22]. Šių atakų principas yra ML modelio – klasifikatoriaus – sprendimų priėmimo ribos radimas – žinant šią ribą pakanka pakeisti kenkėjiškos programos veikimą taip, kad ML modelis priimtų sprendimą klasifikuoti ją kaip nekenksmingą [DCB+21]. Žinoma, rasti šią ribą nėra trivialus uždavinys. Mokslinėje literatūroje išskiriami 3 ribos paieškos atvejai [FWL+19]:

1. **Baltos dėžės** atvejis: kenkėjiško kodo kūrėjas turi visą informaciją apie ML modelį, t. y. modelio architektūrą, svorius, hiperparametrus.
2. **Juodos dėžės su pasitikėjimo įverčiu** atvejis: kenkėjiško kodo kūrėjas gali tik testuoti modelį – t. y. pateikti programą ir gauti atsakymą. Atsakymo forma – klasifikacija ir tikimybė, kad klasifikacija yra teisinga (pasitikėjimo įvertis).
3. **Juodos dėžės** atvejis: kenkėjiško kodo kūrėjas gali tik testuoti modelį. Atsakymo forma yra tik klasifikacija.

Akivaizdu, jog „juodos dėžės“ atvejis yra sudėtingiausias, bet ir labiausiai atitinka realias sąlygas [AKF+18]. Todėl šiame darbe nagrinėjami modeliai, gebantys generuoti varžymosi principais pagrįstų atakų obfuskuotus kenkėjiško kodo pavyzdžius (AE) „juodos dėžės“ atvejams.

Tikslas – nustatyti labiausiai tinkantį karkasą varžymosi principais pagrįstoms atakoms „juodos dėžės“ atvejais.

Uždaviniai:

1. Apžvelgti kenkėjiško kodo obfuskacijos metodus.
2. Nustatyti kriterijus varžymosi principais grįstų atakų karkasams ir juos įvertinti.
3. Atlikti eksperimentinį tyrimą su vienu iš įvertintų karkasų ir patikrinti vertinimo rezultatus.

[1] https://www.av-test.org/en/statistics/malware

Literatūra ir šaltiniai

[AKF+18] H. S. Anderson, A. Kharkar, B. Filar, D. Evans, P. Roth. *Learning to Evade Static PE Machine Learning Malware Models via Reinforcement Learning*. 2018. https://doi.org/10.48550/arXiv.1801.08917. (Žiūrėta 2024-09-30).

[CSD19] R. L. Castro, C. Schmitt, G. Dreo. „AIMED: Evolving Malware with Genetic Programming to Evade Detection“. Iš: *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. 2019, puslapiai 240–247. https://doi.org/10.1109/TrustCom/BigDataSE.2019.00040. (Žiūrėta 2024-09-23).

[DCB+21] L. Demetrio, S. E. Coull, B. Biggio, G. Lagorio, A. Armando, F. Roli. „Adversarial EXEmples: A Survey and Experimental Evaluation of Practical Attacks on Machine Learning for Windows Malware Detection“. Iš: *ACM Trans. Priv. Secur.* 24.4 (2021), 27:1–27:31. ISSN: 2471-2566. https://doi.org/10.1145/3473039. (Žiūrėta 2024-09-30).

[FWL+19] Z. Fang, J. Wang, B. Li, S. Wu, Y. Zhou, H. Huang. „Evading Anti-Malware Engines With Deep Reinforcement Learning“. Iš: *IEEE Access* 7 (2019), puslapiai 48867–48879. ISSN: 2169-3536. https://doi.org/10.1109/ACCESS.2019.2908033. (Žiūrėta 2024-09-18).

[HT17] W. Hu, Y. Tan. *Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN*. 2017. (Žiūrėta 2024-09-18).

[RSR+18] I. Rosenberg, A. Shabtai, L. Rokach, Y. Elovici. „Generic Black-Box End-to-End Attack Against State of the Art API Call Based Malware Classifiers“. Iš: *Research in Attacks, Intrusions, and Defenses*. Sudarė M. Bailey, T. Holz, M. Stamatogiannakis, S. Ioannidis. Cham: Springer International Publishing, 2018, puslapiai 490–510. ISBN: 978-3-030-00470-5. https://doi.org/10.1007/978-3-030-00470-5_23.

[ZHZ+22] F. Zhong, P. Hu, G. Zhang, H. Li, X. Cheng. „Reinforcement Learning Based Adversarial Malware Example Generation against Black-Box Detectors“. Iš: *Computers & Security* 121 (2022), puslapis 102869. ISSN: 0167-4048. https://doi.org/10.1016/j.cose.2022.102869. (Žiūrėta 2024-09-14).