SCIS 2022 2022 Symposium on Cryptography and Information Security
Osaka, Japan & Online, Jan. 18 – 21, 2022
The Institute of Electronics, Information and Communication Engineers

機械学習ベースマルウェア検知モデルに対する clean-label バックドア攻撃とその対策について

Clean-label Backdoor Attack on Machine Learning-based Malware Detection Models and Countermeasures

鄭 万嘉 * 面 和成 *
Wanjia Zheng Kazumasa Omote

キーワード ポインティング攻撃，clean-label バックドア攻撃，機械学習，マルウェア検知，ブラックボックス

あらまし

近年機械学習技術が活用されている中，AI システムのセキュリティについても注目されつつある．画像認識システムはもちろん，機械学習ベースのマルウェア検知システムにおけるセキュリティ関連の研究も多く行われている．従来より機械学習モデルの学習段階で加工されたポイズンニングデータを攻撃者が混入し，機械学習モデルを誤動作させるポイズンニング攻撃と呼ばれる攻撃手法が存在するが，そのうち，通常ではモデルが正常に動作されており，特定のバックドアデータに対してのみ誤動作が起るというバックドア攻撃は脅威である．最近，画像認識領域において clean-label バックドア攻撃と呼ばれる攻撃手法が多く研究されており [2, 3, 4]，モデル学習時に混入させるポイインティングデータを加工する際，データのラベルを変更しないのが特徴的である．

しかし，マルウェア検知システムに対する研究 [1] はまだ少ないため，我々はマルウェア検知システムにおける clean-label バックドア攻撃に注目し，より現実的でリスクが高い機械学習モデルの知識を必要としないブラックボックスの前提で，新たに clean-label バックドア攻撃を提案する．そして，提案攻撃手法に対する実験評価も行い，バックドアデータを仕掛ける際の攻撃成功率が最大 82.05%を達成し，提案攻撃手法の有効性を示す．

さらに，同様な攻撃条件において，Giorgio らの研究 [1] で提案された SHAP に基づく攻撃手法と比較し，提案手法では同レベルの攻撃成功率を維持しながら，攻撃時間を短縮できたことを実験的に示す．

最後，本研究では初めて次元圧縮技術による clean-label バックドア攻撃の防御効果について実験評価を行い，機械学習モデルの学習段階に主成分分析（PCA）と線形判別分析（LDA）を組み込み，バックドア攻撃に最大 76.70%の攻撃成功率削減の効果があることを示す．

参考文献

[1] Giorgio S., Jim M., Scott C., Alina O.(2021). “Explanation-Guided Backdoor Poisoning Attacks Against Malware Classifiers“. 30th USENIX Security Symposium. 1487-1504.

[2] Shafahi, A., Huang, W., Najibi, M., Suciu, O., Studer, C., Dumitras, T., Goldstein, T. (2018). “Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks“. NIPS 2018. 6106–6116.

[3] Peri, N., Gupta, N., Huang, W., Fowl, L., Zhu, C., Feizi, S., Goldstein, T., Dickerson, John. (2020). “Deep k-NN Defense Against Clean-Label Data Poisoning Attacks“. Computer Vision–ECCV 2020 Workshops. 55-70.

[4] Turner, A., Dimitris T., Aleksander M.(2019). “Clean-Label Backdoor Attacks“. https://people.csail.mit.edu/madry/lab.

* 筑波大学, 〒 305-8577 茨城県つくば市天王台 1-1-1, University of Tsukuba, University of Tsukuba,1-1-1 Tennoudai, Tsukuba, Ibaraki, 305-8573, Japan.