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

GAN構造を用いたポイズニング検知
Detecting Poisoning Attacks using GAN Structure

清水 俊也 *
Toshiya Shimizu

キーワード 機械学習, 敵対的機械学習, ポイズニング, GAN

あらまし

背景と課題 機械学習のセキュリティ問題の一つであるポイズニング攻撃は, 訓練データ中に少量の異常データを挿入する（もしくは訓練データそのものを変化させる）ことにより, モデルを汚染する攻撃をさす. ユーザからデータを集め逐次的に学習するシステムなどでは, データがモデルに及ぼす影響を予測し, 悪影響を及ぼすデータを訓練時に排除することが必要となる. ポイズニング攻撃への対策手法は, 予めデータに対する異常検知器を作り, 訓練データを収集した際に検知器を用いて異常データを排除する手法と, 収集したデータを利用して実際にモデルの訓練を行うことによりデータがモデルに与える影響を評価する手法の二系統が主流である. 前者は検知にかかる時間が短いが, 検知精度が低く, 適応的攻撃にも脆弱なことが知られており, 後者は検知精度が良い分, 訓練を要するため検知時間が長いことが課題となっている.

提案手法 本稿では, モデル運用時に素早く検知が行える検知器事前生成型の手法に着目し, 検知精度の向上させつつ, 適応的攻撃へのロバストネスを得られる検知手法を提案する. 提案手法の特徴は以下の通りである.

- 検知精度を向上するために, モデルの説明可能性 (XAI) を特徴量抽出手法として用いる. Influence Function[1] などの XAI 技術はデータのモデルへの影響を定量化できるため, ポイズニング検知手法と非常に相性がよい.
- 適応的攻撃への対策として, 検知器の学習に敵対的生成ネットワーク構造（GAN）を用いる. GAN 構造を用いてモデルを汚染する可能性のあるデータを生成する学習器と検知器を対立させることにより, 様々な異常データを効率的に学習できる.

提案手法を MNIST, CIFAR10, CIFAR100 などの画像データセットに適用したところ, 従来の検知器を事前に生成する系統の検知手法（[2] など）に比べ, データセットに依存するものの, 平均しておおよそ（百分率における値として）10%以上検知精度が向上した. . 特に適応的攻撃に関しては, 防御手法を知っている完全知識の攻撃者の仮定のもとでも, 60%以上の検知効率を達成することができた.

まとめ 今回, ポイズニングデータ検知のために, XAI と敵対的構造を用いた検知器の提案を行い, 適応的攻撃を含む攻撃の検知精度をあげることができた. 今後は, テーブルデータ等の, 入力データのドメインに制限がある場合のために提案手法の GAN 構造に改良をする, もっとも検知に適している XAI の理論的解析を行うなど, さらに研究を進めていきたい.

参考文献

[1] K. W. Pang, L. Percy, “Understanding Black-box Predictions via Influence Functions”, Proceedings of the 34th International Conference on Machine Learning(ICML), PMLR 70:1885-1894, 2017.

[2] J. Steinhardt, P.W. Koh, and P. Liang, “Certi ed Defenses for Data Poisoning Attacks”, Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS’17)”, pp. 3520-3532.

* 富士通株式会社, 神奈川県川崎市中原区上小田中 4 丁目 1-1 Fujitsu Ltd., 1-1, Kamikodanaka 4-chome, Nakahara-ku, Kawasaki shimizu.toshiya@fujitsu.com