4ZD-01

クリーンラベルバックドア攻撃に関する基礎的検討

目黒 諒 [†1]　加藤 広野 [†2]　成定 真太郎 [†2]　披田野 清良 [†2]　福島 和英 [†2]
内林 俊洋 [†3]　樋地 正浩 [†4]　菅沼 拓夫 [†5,1]

[†1] 東北大学工学部　[†2] 株式会社 KDDI 総合研究所
[†3] 九州大学情報基盤研究開発センター　[†4] 東北大学会計大学院
[†5] 東北大学サイバーサイエンスセンター

1　はじめに

グラフニューラルネットワーク（GNN）は，攻撃者が意図的に混入した毒データで学習データを汚染し，毒データのみを特定のクラスに分類されるように仕向けるバックドア攻撃に対して脆弱であることが分かっている [1]．バックドア攻撃には，学習データに混入する毒データの正解ラベルを変更するラベルフリップによる攻撃と，ラベルを変更しないクリーンラベルによる攻撃があるが，グラフ分類を対象とした GNN に対するバックドア攻撃の研究ではラベルフリップによる攻撃の提案がほとんどであり，クリーンラベルでの攻撃について有効性は十分に明らかにされていない．

既に提案されているクリーンラベルバックドア攻撃 [2] においては，グラフ内の辺の接続情報を操作してトリガー (訓練データの毒データと入力する毒データを関連付ける印) を作成し，全ての毒データに対して共通の固定トリガーを付与している．そのため，毒データが検知されやすいという問題がある．

本研究では，各グラフデータごとに異なるトリガーを付与するクリーンラベルバックドア攻撃を提案する．本提案においては攻撃対象の GNN モデルとは異なる代理モデルにおける勾配を利用し，各グラフの構造を考慮して，トリガーを付与する．また，攻撃者が攻撃対象のモデルに関する知識を必要としないブラックボックスでの攻撃を想定する．

本発表では，提案方式のアルゴリズムを示し，既存研究での固定トリガーを付与するクリーンラベルバックドア攻撃との比較実験を行って，提案方式を評価する．

2　提案方式

本論文では，ブラックボックス環境下で，各グラフごとに異なるトリガーを付与するクリーンラベル攻撃を提案する．提案方式では，各グラフに適したトリガーを付与するために，TRAP 攻撃 [3] をもとに，隣接行列に関する損失関数の勾配を用いて各辺に対するスコアを算出する．そして，スコアを元に辺情報の改変を行う．TRAP 攻撃は，スコアの算出に，攻撃対象の学習データと同じものを利用しており，かつラベルフリップによる攻撃である．一方，提案方式では，TRAP 攻撃と類似した手順でスコアを算出するが，攻撃対象の学習データとは異なるデータで訓練された代理モデルを利用する．代理モデルは，隠れ層のニューロン数のみが異なるモデルを利用する．Algorithm1 は提案方式において，毒データを一つ生成する際のアルゴリズムである．学習データに混入する毒データは，Algorithm1 に，target ラベルを持つグラフデータを与えて作成する．一方，モデルに入力として与える毒データは，Algorithm1 に，source ラベルを持つグラフデータを与えて作成する．

Algorithm 1: 提案方式の毒データ生成

```
Input: one graph data : graph
Output: one graph data which has trigger

Function make_one_poison:
    # calculate gradient
    out ← surrogate_model(graph)
    loss ← loss_function(out,1-graph.label
      )
    loss.backward()
    grad ← graph.adj.grad
    # sort edges based on scores
    S ← grad * (2 * graph.adj - 1)
    sorted_edges ← argsort(-S)
    # invert M edges
    for m in range(M):
        invert sorted_edges[m] information

    return Data(x=graph.x, y=graph.y, adj=
      new_adj)
```

Fundamental Study on Clean Label Backdoor Attacks
Ryo MEGURO[†1], Hiroya KATO[†2], Shintaro NARISADA[†2], Seira HIDANO[†2], Kazuhide FUKUSHIMA[†2], Toshihiro UCHIBAYASHI[†3], Masahiro HIJI[†4], and Takuo SUGANUMA[†5,1]
[†1]School of Engineering, Tohoku University
[†2]KDDI Research
[†3]Research Institute for Information Technology, Kyushu University
[†4]Tohoku University Accounting School
[†5]Cyberscience Center, Tohoku University

表 1: target ラベルが 0 の場合の結果

Dataset attack	model	ASR(%) | CAD(%)				
		proteins	COX2	AIDS	BZR	DHFR
固定トリガー	GCN	71.22 | 2.10	99.52 | 0.00	8.25 | 0.07	72.77 | 0.12	26.05 | 1.18
	GIN	99.88 | 2.15	100.00 | -0.10	93.75 | 0.37	100.00 | 0.00	96.05 | -0.06
提案方式	GCN	33.55 | 2.95	100.00 | -0.31	0.50 | 0.42	84.50 | 0.36	83.55 | 3.09
	GIN	90.77 | 3.36	100.00 | 0.00	91.90 | 0.27	98.88 | 0.12	68.94 | -0.19

表 2: target ラベルが 1 の場合の結果

Dataset attack	model	ASR(%) | CAD(%)				
		proteins	COX2	AIDS	BZR	DHFR
固定トリガー	GCN	54.68 | -0.04	64.46 | 0.10	4.25 | 0.05	39.75 | -1.34	85.25 | -1.77
	GIN	100.00 | -1.07	85.53 | 0.00	99.12 | 0.02	98.50 | 0.48	94.57 | 3.28
提案方式	GCN	0.54 | -0.08	47.87 | -7.62	0.00 | -0.12	60.74 | 0.24	99.32 | 0.85
	GIN	94.68 | 0.80	29.14 | 0.00	78.25 | 0.00	67.25 | -0.24	86.77 | 0.78

3 実験と評価

既存の固定トリガーによるクリーンラベルバックドア攻撃 [2] との比較実験を行う．

3.1 条件

毒データの個数はデータセット全体のグラフの個数の 10% とする．また，トリガーサイズに関しては, 固定トリガーを用いる手法においては, 平均ノード数の 20% とする．提案方式では，固定トリガーのトリガーサイズに合わせて，$0.2n \times (0.2n-1)/2$ 個の辺を改変することとする．学習データとテストデータと毒データの割合の比は，7:2:1 である．

3.2 モデルとデータセット

GCN [4], GIN [5] により評価を行う．データセットとしては, 無向グラフ, ノード特徴がある, ラベルが 2 値，という条件を満たす proteins, COX2, AIDS, BZR, DHFR [6] を用いた．

3.3 評価指標

攻撃の有効性及び回避性を評価する指標として，ASR と CAD を用いる．ASR は，毒データ入りの学習データセットで訓練されたモデルに対して，トリガー入りのグラフデータを与えたときに target ラベルとして誤分類される割合を表す．CAD は，毒データが無い学習データセットで訓練されたモデルの分類精度と，毒データ入りの学習データセットで訓練されたモデルの分類精度の差である．

3.4 結果と評価

表 1，表 2 にそれぞれ，target ラベルが 0 のときと，target ラベルを 1 のときの結果を示す．

ASR に関して, 固定トリガーの攻撃が優位なデータセットが複数あった．これは，提案方式が代理モデルの精度に依存しており，代理モデルの精度が低い場合，M 個の辺を適切に反転できないためと考えられる．

防御の観点から見ると，固定のトリガーを利用することは，毒データの検知がされやすいという問題がある．提案方式のトリガーは固定では無く，グラフごとに異なるように作成したため，毒データの検知から逃れるという観点において優れていると言える．

4 おわりに

グラフ分類を対象とした GNN に対するクリーンラベルバックドア攻撃を提案し，既存研究である固定トリガーによるクリーンラベルバックドア攻撃との比較を行った．提案方式では，スコアを計算する際の損失の算出にラベルの情報を用いているが，今後は，潜在空間の特徴表現，隣接行列，ノード特徴も利用することで ASR の向上が期待できる．

参考文献

[1] Zaixi Zhang, Jinyuan Jia, Binghui Wang, and Neil Zhenqiang Gong. Backdoor attacks to graph neural networks. In *Proceedings of the 26th ACM Symposium on Access Control Models and Technologies*, pages 15–26, 2021.

[2] Jing Xu and Stjepan Picek. Poster: Clean-label backdoor attack on graph neural networks. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 3491–3493, 2022.

[3] Shuiqiao Yang, Bao Gia Doan, Paul Montague, Olivier De Vel, Tamas Abraham, Seyit Camtepe, Damith C Ranasinghe, and Salil S Kanhere. Transferable graph backdoor attack. In *Proceedings of the 25th International Symposium on Research in Attacks, Intrusions and Defenses*, pages 321–332, 2022.

[4] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[5] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[6] Ryan A. Rossi and Nesreen K. Ahmed. The network data repository with interactive graph analytics and visualization. In *AAAI*, 2015.