表形式データ学習モデルに対する外部公開統計情報を用いたモデル複製攻撃

宍戸 克成[1,a)]　清水 俊也[1]　樋口 裕二[1]　森川 郁也[1]　矢嶋 純[1]　辰巳 将崇[2]　岩花 一輝[2]
矢内 直人[2]

概要：機械学習に対する攻撃の一種である複製攻撃は，提供されているモデルに対するクエリとその出力の振る舞いから，攻撃者がそのモデルと同等のモデルを複製する攻撃である．CVPR 2021 で提案されたデータフリーモデル複製攻撃は，攻撃者の一切の背景知識を仮定することなく，MNIST は 1M クエリ，CIFAR-10 は 10M クエリ程度でモデルの複製に成功している．本稿では，表形式データを学習したモデルに対して，外部に公開されている統計情報を利用する仮定により現実的なデータフリーモデル複製攻撃の可能性を示す．ランダムクエリベースの手法と遺伝的アルゴリズムベースの手法を提案し，表形式データにおいて攻撃者が統計情報をもとに生成したランダムなデータをクエリするだけで数千クエリから数万クエリで同等のモデルを複製できる知見を得た．またランダムクエリベースの手法と遺伝的アルゴリズムアルゴリズムベースの手法の向き不向きについて議論する．

キーワード：モデル複製攻撃，データフリー，表形式，機械学習

Tabular Data Model Extraction by Public and External Statistics

Katsunari Shishido[1,a)]　Toshiya Shimizu[1]　Yuji Higuchi[1]　Ikuya Morikawa[1]　Jun Yajima[1]
Masataka Tatsumi[2]　Kazuki Iwahana[2]　Naoto Yanai[2]

Abstract: Model extraction attack is a kind of machine learning attack, where the attacker is able to steal provided model. In CVPR 2021, Data-Free Model Extraction Attack that does not require any dataset and samples was proposed and succeeded to steal an image trained model by a few million queries. In this paper, we have studied a realistic data-free tabular data model extraction attack by public and external statistics and show that there is a risk of stealing trained model by a few thousand queries.

Keywords: Model Extraction, Data-Free, Tabular data, Machine Learning

1. 序論

1.1 背景

計算機能力の向上やデータ数の爆発的な増加に伴い，深層学習などの機械学習を用いたデータ解析やデータ分析が盛んに行われている．機械学習モデルの構築において学習データの収集と訓練処理は高い負荷を要する作業であり，例えば Yang ら [1] のモデルでは学習に 61,000 ドルから 250,000 ドルも消費している[*1]．インターネットサービス事業者は，金銭と時間的なコストをかけて訓練したモデルを，画像に映った人物や物品を認識して自動的に分類するサービスや不適切な画像を検知して削除する機能などに利用している．また自動車メーカーは，自動運転技術の開発に必要不可欠な標識や信号の認識に深層学習を用いてい

[1] 富士通株式会社
Fujitsu Limited
[2] 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University
[a)] k.shishido@fujitsu.com

[*1] https://syncedreview.com/2019/06/27/the-staggering-cost-oftraining-sota-ai-models/

る．このように，深層学習などの機械学習は自動化技術の普及に必要不可欠な価値の高いものとなっている．

一方で，機械学習の仕様や特性を用いた攻撃が注目されている．機械学習に対する攻撃として，推定結果を意図的に誤らせる敵対的サンプルやモデル汚染攻撃，訓練データやモデルの情報を盗む訓練データ推定やモデル複製攻撃がある．多くの攻撃は未だ研究レベルに留まっているが，現実世界で発生した攻撃も存在している．「マルウェア検知の回避」は，脆弱性として JVN レポート[*2]が挙げられている．また Black Hat USA 2020 にて Herbert-Voss は，現実世界で発生した攻撃として，「暗号通貨取引ボットを誘導・推薦エンジンを騙して自社を有利にする・オンライン詐欺検知の回避」[2] を紹介している．これらの攻撃の共通点は，攻撃者にとって金銭的なメリットが得やすいことである．つまり，攻撃者にとって金銭的なメリットがある攻撃やユースケースから研究を進めることが重要と考える．

そこで本研究では，攻撃者にとって攻撃のメリットがあると考えられる金融分野や医療分野に注目し，その分野で扱われることが多い「表形式データ」を学習したモデルに対するモデル複製攻撃に注目する．モデル複製攻撃は，クラウドやエッジといった環境に配置されている訓練済みモデル（被害者モデル）の任意のデータの入出力の組を利用して，同等の精度を持つモデルを不正に複製するものである．先にもあげたように，機械学習モデルの価値の向上を受けて，攻撃者が自らデータ収集や訓練処理に必要なコストよりも低いコストで被害者モデルと同等のモデルを複製できれば，攻撃者にとって大きなメリットとなる．複製したモデルが得られると，攻撃者は利用料金を一切支払うことなくモデルから推論結果を得られたり，類似サービスを提供できる．それらに加え，敵対的サンプルの踏み台 [3] に利用できることが報告されている．このように，モデルを複製されることによる直接的な損害だけでなく，そのモデルに対するさらなる攻撃にもつながることから，モデル複製攻撃はモデル所有者にとって極めて重要な問題である．

1.2 モデル複製攻撃の変遷

2016 年に Tramèr らがモデル複製攻撃 [4] を報告して以降，攻撃者が事前に持つべき知識量を減らす方向の研究がされている．Tramèr らの手法では，攻撃者は被害者モデルの訓練に利用された訓練データを 2 割ほど保有する必要があった．2019 年に Jutti らが提案した手法 (PRADA)[5] では，攻撃者は各クラス数十サンプルの訓練データのみ保有していればよく，それらのサンプルに対する敵対的サンプルを生成することでモデル複製を実現している．また同年に Orekondy らが提案した手法 (Knockoff Nets)[6] では，攻撃者は訓練データを一切保有する必要がなく，代替データセットさえ用意すればモデル複製が可能なことを示している．しかしながら，代替データセットは訓練データに類似したデータを含んでいる必要があるため，性質の良い代替データセットを準備するコストが高いことを指摘されており，攻撃者が一切のデータを保有することなくモデルを複製するデータフリーな手法 [7], [8] が近年報告されている．データフリーな手法は，攻撃者が一切のデータを保有する必要がない一方で，敵対的ネットワークを用いてデータを生成してモデルを複製するため，必要なクエリ量は既存手法のおよそ 1 万倍になることが示されている．

攻撃者が持つ知識量を減らす方向性以外にも，訓練時のハイパーパラメータを抽出する攻撃 [9] や画像処理 [10]・言語処理 [11], [12] といったアプリケーションへの攻撃も示されている．

様々な方向性で研究が行われているが，攻撃者にメリットがあると考えられる金融・医療分野において使われることの多い「表形式データ」に注目した研究は我々が知る限り行われていない．

1.3 貢献

本研究では，表形式データを学習したモデルに対する新たなモデル複製攻撃を示した．新たなモデル複製攻撃を示す上で，新たな攻撃クラス「補助情報を用いた攻撃」を導入した．提案した攻撃クラスは，訓練データの平均・分散や最小値・最大値を指している．詳細は 3 節に述べるが，既存のデータフリーな手法よりも現実的な仮定といえる．

本研究は，「補助情報を用いた攻撃」として「ランダムクエリを用いた攻撃・遺伝的アルゴリズムを用いた攻撃」を提案し，4 つの表形式データを用いた実験を通して有効性を示した．補助情報を用いることで，分類が比較的簡単な問題では，ランダムクエリを用いた攻撃でさえ被害者モデルの 70–80%程度モデルの複製ができることがわかった．また，ランダムクエリが有効に働かないケースにおいて，遺伝的アルゴリズムを用いた攻撃の有効性も示唆された．

2. 準備

本節では予備知識として機械学習およびモデル抽出攻撃の概念について説明する．

2.1 機械学習

$\mathbb{N}$ を自然数の集合, $\mathbb{R}$ を実数の集合, $\mathcal{C}$ をラベルの集合とする．機械学習モデルは任意の $m, n \in \mathbb{N}$ において, m 次元の特徴量を持つ $x \in \mathbb{R}^m$ を入力とし，n 個のラベルそれぞれに関する確率を出力する関数 M として定義される．より正確には，$M(x)$ は $i \in [1, n]$ における各ラベルを持つクラス $c_i \in \mathcal{C}$ に x が属する確率 p_i をそれぞれ出力する．このとき，機械学習モデル M は任意のデータとラベルの組 $(x, c) \in \mathbb{R}^m \times \mathcal{C}$ を与えられることで M の内部パラメータ

[*2] https://jvn.jp/vu/JVNVU98738756/

θ を鍛える学習と，パラメータ θ を持つモデル M_θ が未知の入力 x' を与えられることで上述した確率を出力する推論を処理として持つ．

2.2 モデル複製

モデル複製攻撃における攻撃者 $\mathcal{A}$ の目的は，攻撃対象となる機械学習モデル（被害者モデルと呼称）$M_\mathcal{V}$ を模倣する複製モデル $M_\mathcal{A}$ を得ることにある．モデル複製攻撃では，攻撃者は任意のデータを被害者モデルに入力できる Black-box 攻撃を前提として議論され，入出力関係から被害者モデルを複製する．

2.2.1 モデル複製の評価指標

複製したモデルと被害者モデルの類似度指標に使われる「Accuracy」の定義を以下に述べる．

定義 1 (Accuracy)**.** テストデータセット S と複製モデル $M_\mathcal{A}$ に対し，Accuracy は以下で定義される．

$$accuracy = \frac{1}{|S|} \cdot \sum_{(x,y)\in S} I(M_\mathcal{A}(x), y),$$

但し，$I(a,b) = 1$ if $a = b$, otherwise $I(a,b) = 0$ である．あるデータセット S として，テストデータセット D_{TEST} が使用されることが多い．つまり，Accuracy はテストデータセットに対する精度であり，Fidelity は真のラベルを $M_\mathcal{V}(x)$ とした場合の精度にあたる．

2.2.2 攻撃者クラス

攻撃者が保有するデータ量の違いを表す言葉とその意味合いについて説明する．これまでに提案された手法は攻撃者の持つ知識量によって分類できる．下記に示すように，各攻撃クラスでは攻撃者が訓練データセットに関してアクセスできる情報や保有している情報が異なる．訓練データセットに関する知識量が少なければ少ないほどモデル複製が難しくなると考えられる．攻撃者にとって最も有利な仮定は，一部データサンプルを用いた攻撃で，訓練データセットのうち 2 割程度のデータや各クラスごとに数十から数百のデータを使う攻撃 [4], [5] が提案されている．次に有利な仮定として，代替データセットを用いた攻撃で，訓練データセットのデータを一切使わず，訓練データセットに含まれるデータと類似したデータを含むような代替データセット使う攻撃 [6] が提案されている．そして攻撃者にとって，最も不利な仮定はデータフリーな攻撃で，訓練データに関する事前知識を一切使わない（代替データセットすら使わない）攻撃 [7], [8] が提案されている．本研究で新たに提案する「補助情報を用いた攻撃」を加え，現在 4 つの攻撃者クラスに分類できる．補助情報を用いた攻撃は，各属性の「平均・分散」や「最小値・最大値」といった統計量を使う攻撃で，本研究で新たな手法を提案する．

2.3 関連研究

本章では，モデル複製攻撃のベースラインとして扱われることの多い PRADA[5]，Knockoff Nets[6]，データフリー手法 [7] について詳しく述べる．まずはじめに表 1 で，3 つのモデル複製攻撃の精度と被害者モデルへの問い合わせ数 (#query) を示す．被害者モデルに問い合わせながら，データ生成をする PRADA [5] と Data-Free ME [7] の問い合わせ数はデータセットの中から効果的なサンプルを選択する Knockoff Nets [6] よりも大きくなる傾向がある．

2.3.1 一部サンプルを用いた攻撃：PRADA

訓練データセットの一部サンプルを用いる攻撃として，Tramèr らの手法 [4] と PRADA[5] などが提案されている．2 つの手法の違いは，攻撃者が使える訓練データセットのサンプルの割合である．Tramèr らの手法は，訓練データセットのうち 2 割程度のサンプルの知識量が必要だが，PRADA では各クラスごとに 10 個程度のサンプルの知識量でモデル複製が可能である．PRADA では，攻撃者が保有するサンプルから，敵対的サンプルを繰り返し生成してデータ増強をすることで，被害者モデルと同等の精度を持つモデルを複製する攻撃手法である．表 1 に示すように，白黒画像である MNIST のほうが RGB カラー画像である GTSRB よりも少ない問い合わせ数で精度の高いモデルが得られている．RGB カラー画像に適した敵対的サンプルの生成手法を利用することで，複製モデルの精度が向上することも報告されている．

2.3.2 代替データセットを用いた攻撃：Knockoff Nets

Knockoff Nets[6] は，攻撃者が代替データセット D_{sub} を準備し，被害者モデルに問い合わせながら D_{sub} からモデル複製に効果的なサンプル $x \in D_{sub}$ を選択・訓練して被害者モデルと同等の精度を持つモデルを複製する攻撃手法である．モデルを複製するためには，訓練データセットのサンプルに似たサンプルを訓練すれば良い．代替データセットから効果的なサンプルを選択するために，訓練データセットと代替データセットが重なっている部分（積集合）に含まれるサンプルを選択・訓練する．Knockoff Nets では，被害者モデルの入出力関係から積集合部分に含まれるサンプルを強化学習を用いて選択していくことで，モデル複製を実現している．

実験では，4 つのデータセット (Caltech256, CUBS200, Indoor67, Diabetic5) を訓練データセットとして，2 つの大規模データセット (ILSVRC[*3], OpenImages[*4]) を代替データセットとして設定している．一様分布に従い独立に大規模データセットからサンプリングして得られたサンプルを問い合わせるランダム戦略と強化学習を用いて問い合わせるサンプルを決める適応的戦略を比較した結果，適応

[*3] https://www.image-net.org/challenges/LSVRC/
[*4] https://storage.googleapis.com/openimages/web/index.html

表 1: 既存のモデル複製攻撃の概要

攻撃手法	データセット	被害者モデルの精度に対する複製モデルの精度	#query
PRADA	MNIST	0.99	102,400
10 natural seed samples per class	GTSRB	0.50	110,880
Knockoff Nets D_{sub}: OpenImages	Caltech256	0.94	60,000
	CUBS200	0.92	60,000
	Indoor67	0.94	60,000
	Diabetic5	0.82	60,000
Data-Free ME	SVHN	0.99	1,000,000
	CIFAR-10	0.92	10,000,000

的戦略がランダム戦略よりも少ない問い合わせ数で精度の高いモデルを複製できていることが示された．また複製モデルのアーキテクチャは，大きくなれば大きくなるほど精度がよくなることも実験で示しており，被害者モデルのアーキテクチャの情報はモデルの精度に重要でないことを主張している．Knockoff Nets で複製されたモデルの精度は，訓練データセットと代替データセットの積集合部分に含まれるサンプルに大きく依存する．そのため，訓練データセットと相性の良い代替データセットの準備にかかるコストの高さが指摘されている．

2.3.3 データフリーな攻撃：Data-Free ME

Data-Free ME[7], [8] は，攻撃者が Generator G と被害者モデル $M_{\mathcal{V}}$・複製モデル $M_{\mathcal{A}}$ から構成される敵対的ネットワークを用いて Generator と複製モデルの訓練を行うことで，被害者モデルと同等の精度を持つモデルを複製する手法である．Data-Free ME の最適化は下記の通りである．

$$\min_{\theta_{\mathcal{A}}} \max_{\theta_g} \mathbf{E}_{z \leftarrow \mathbb{R}^n}[L(M_{\mathcal{V}}(G(z;\theta_g)), M_{\mathcal{A}}(G(z;\theta_g);\theta_{\mathcal{A}}))],$$

但し，$L(a,b)$ は a,b の類似度を表す損失関数 (Loss) である．損失が小さいほど a,b が類似していることを示す．この式は，Generator G が生成したサンプル $G(z)$ を被害者モデル $M_{\mathcal{V}}$ と複製モデルに $M_{\mathcal{A}}$ に入力し，得られた推論結果が大きく異なるように G を訓練し，その上で被害者モデルと複製モデルの推論結果が大きく異なるようなサンプル $G(z)$ に対して，$M_{\mathcal{V}}$ と $M_{\mathcal{A}}$ が一致するように複製モデル $M_{\mathcal{A}}$ を訓練することを意味する．上記の最適化を解くことで，Generator は複製モデルの訓練に効果的なサンプルを生成し続け，攻撃者が保有していないテストデータセット D_{TEST} に対する Accuracy や Fidelity が向上する．モデル複製攻撃の設定では，攻撃者は被害者モデルへのアクセス方法が Black-box となるため，被害者モデルに関する勾配を Black-box 近似しなければいけない．Black-box モデルの勾配近似には，Black-box モデルに対する敵対的サンプルの生成 [13] で使われている有限差分を用いる．

実験では，2 つのデータセット「SVHN・CIFAR-10」を訓練した被害者モデルに対してモデル複製を行った．結果，10^6 回の問い合わせで SVHN の被害者モデルの 0.99 倍の精度を持つ複製モデル，10^7 回の問い合わせで CIFAR-10 の被害者モデルの 0.92 倍の精度を持つ複製モデルを得ることを示した．

3. データフリーな手法の難しさ

本章では，表形式データのデータフリーな攻撃の難しさについて述べる．

本研究がメインに扱う表形式に関するデータフリーな手法の難しさは，各属性ごとにスケールが異なりデータ不定であること，数値変数と名義変数の取り扱いの違いが要因である．

画像は非構造化データの一種で，データベースなどで扱う構造化データと異なり特定の構造を持たないデータを指す．特定の構造を持たないとは，データの各属性や値に対して特定のルールがないことを指す．例えば，サイズが $(w \times h)$ の RGB カラー画像は，各要素の値 $x \in [0, \cdots, 255] \subset \mathbb{Z}$ から構成されるサイズが $3 \times w \times h$ の 3 次元配列で表現される．配列の各要素間には，特定のルールがなく任意の値をとることができる．

表形式は構造化データの一種で，特定の構造を持つデータを指す．特定の構造やルールを持つため，データベースなどで扱いやすい．例えば，金融や医療分野では，顧客情報を扱う名簿や患者情報を扱うカルテは各列ごとに属性と型（名義変数や数値変数（実数，整数））が決められている．また，各属性間でもある特定のルールが存在することもある．

上記で述べたように画像と表形式はデータの性質が大きく異なるので，これらのデータを訓練・推論するために，異なる前処理が行われる．例えば，画像は各要素の値 $x \in [0, \cdots, 255] \subset \mathbb{Z}$ なので，各属性間でのスケールが統一されている．しかしながら，表形式は各属性ごとにスケールが異なるケースが多い．例をあげると，身長と体重のスケールは異なる．さらに，表形式は数値変数以外に名義変数を含むことがあるため，数値しか扱えないロジスティック回帰やニューラルネットワークでは，one-hot encoding や埋め込みといった方法で数値に変換する必要がある．

従来のデータフリーな攻撃手法 [7], [8], は，データの正

則化処理や名義変数の数値変換を考慮していないことが問題である．具体的に，画像を訓練したモデルに対する攻撃を暗黙的に仮定しているため，従来手法では，被害者モデルに対する入力を $[-1,1]^d$(d は次元数) [7], [8] としている．しかしながら，先で述べたように，表形式データは国名や性別のような名義変数と身長や体重のような数値変数（実数・整数）から構築される．つまり，入力は各属性ごとに変数の種類が異なる上にスケールも異なるため，各属性のデータが不定となり正則化処理ができない．さらに，名義変数は one-hot encoding などを用いて数値に変換する必要がある．よく知られている Generator は，名義変数を直接生成することができないため，Generator が生成した数値をなんらかの変換を用いて名義変数に変換する必要があり，最適化に勾配を用いる従来手法が適用できない．これらデータ構造上の違いを踏まえ，表形式データを訓練したモデルに対して既存手法を適用することは非現実的であり，データフリーな仮定の下で現実のシステムを攻撃することは難しいといえる．

上述した問題に対し，本研究では「公開された統計情報を基にすることで，上述したデータフリー攻撃の問題が改善できる」という仮説を検討する．被害モデルが利用しているであろうデータの平均や最大値・最小値など統計情報を得ることができたとすると，攻撃者はデータそのものは一切持つことがなくモデル抽出攻撃が可能となる．一方，例えば我が国における各年代・性別ごとに，金融データに対しては平均年収，健康データに対しては身長・体重など，様々なデータが利用できることが考えられる．すなわち，統計情報の利用は実用的な側面もあるとみなせる．

4. 提案手法

本章では新しいモデル複製攻撃として，「ランダムクエリを用いたモデル複製攻撃」と敵対的ネットワークをベースとした「補助情報と遺伝的アルゴリズムを用いたモデル複製手法」の 2 手法を提案する．

ランダムクエリを用いた複製攻撃は，攻撃者が訓練データの統計情報にアクセスできるために実現可能な攻撃である．そして素朴なランダムクエリを用いたモデル複製攻撃の改善手法の１つとして，遺伝的アルゴリズムを用いた手法を提案する．遺伝的アルゴリズムは，Black-box 最適化に使われる手法のひとつで，Black-box 仮定における敵対的サンプルの生成 [14], [15] にも使われている．本研究では，3 章で述べた問題に対する解決手段として遺伝的アルゴリズムを適用し，その有効性について議論する．

4.1 遺伝的アルゴリズム

遺伝的アルゴリズムは，解の候補を遺伝子として表現し，その都度適応度が高い遺伝子を選択・交叉・突然変異を繰り返しながら解を探索するヒューリスティックアルゴリズムである．遺伝的アルゴリズムは，微分不可な目的関数に対しても解の探索できることが特徴である．

アルゴリズム 1 GA(Optimization)

入力: 適応度評価式 OF, 入力 x, threshold t, mutation probability p
出力: 最適解 θ.

Generate k random genes $\theta_{i\in[1,\cdots,k]}$
while 1 **do**
  for $i = 1$ to k do **do**
    Compute $fitness\_score_i = OF(x, \theta_i)$ for all genes
  end for
  if $fitness\_score_i \geq t$ **then**
    Return θ_i
  end if
  Select m good genes from all genes
  Apply crossover for selected genes
  Apply mutation with probability p for genes after crossover
end while

アルゴリズム 2 Random Query Model Extraction

入力: 被害者モデル $M_{\mathcal{V}}$，複製モデル $M_{\mathcal{A}}$，公開情報データ: i 番目の属性訓練データの最小値 min_i・最大値 max_i, Query budget Q
出力: $M_{\mathcal{A}}$.

Generate $Q \times n$ matrix $X = X_{j,\in[Q],i\in[n]}$
for $i = 1$ to n(属性数) **do**
  $X_{j\in[Q],i} \leftarrow Uniform(min_i, max_i)_{j\in[Q]}$
end for
$M_{\mathcal{A}} \leftarrow Training(X, M_{\mathcal{V}}(X))$
Return $M_{\mathcal{A}}$

遺伝的アルゴリズムをアルゴリズム 1 に示す．はじめに k 個の初期遺伝子をランダムに生成し，適応度を評価する．適応度の評価には，解きたい最適化問題の目的関数が用いられる．次の選択は，良い適応度を持つ遺伝子を選び，良い遺伝子を交叉させることで，更に良い遺伝子を生成していく．交叉の仕方は，一点交叉・二点交叉・一様交叉などいくつか知られており，本研究では一様交叉を採択している．交叉によって生成された遺伝子に対し，突然変異がある確率 p で発生する．突然変異は，生成した遺伝子が局所解に陥った際に，その解から抜け出すための重要な役割を担う．こうして進化した遺伝子がある一定の終了条件を満たしていれば最適解として出力され，条件を満たしていなければ条件を満たすまで繰り返し遺伝子を進化させる．

遺伝的アルゴリズムは，微分不可な目的関数でも解を探索できるが，ランダムに遺伝子を生成して交叉・突然変異を最適解が得られるまで繰り返すため，解を導き出すコストが高い．

4.2 ランダムクエリを用いたモデル複製攻撃

ランダムクエリを用いたモデル複製攻撃は，アルゴリズム 2 に示すように各属性の最小値と最大から独立に一様

分布に従いサンプリングしてデータを生成する．生成したデータを被害者モデル $M_{\mathcal{V}}$ に問い合わせ，得られた出力を用いて複製モデル $M_{\mathcal{A}}$ を訓練する．

4.3 遺伝的アルゴリズムを用いたモデル複製攻撃

本研究が提案する敵対的ネットワークを図 1 に示す．図 1 で使われている $inv\_onehot(\cdot)$ は Generator が生成した数値から名義変数に変換（逆正規化），$inv\_normalization(\cdot)$ は Generator が生成した数値を元ドメインに変換，$Concat(a,b)$ は a,b の結合を表す．本研究で最適化する目的関数 OF を式 (1) に示す．

$$
\begin{aligned}
OF = min_{\theta_{\mathcal{A}}} max_{\theta_g} \mathbf{E}_{z \leftarrow \mathbb{R}^n}[\\
L(M_{\mathcal{V}}(Concat(inv\_onehot(G(z;\theta_g)), \\
inv\_normalization(G(z,\theta_g))), \\
M_{\mathcal{A}}(Concat(inv\_onehot(G(z;\theta_g)), \\
inv\_normalization(G(z,\theta_g))))] \cdots (1)
\end{aligned}
$$

既存のデータフリーな手法 [7], [8] の目的関数との違いは，Generator が生成した数値に対して逆 one-hot encoding や逆正規化を適用し，$M_{\mathcal{V}}, M_{\mathcal{A}}$ の入力を表形式データに変換していることである．

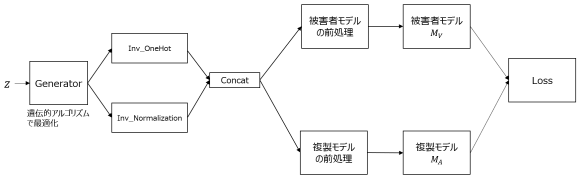

図 1: 提案手法のネットワーク構成図

提案手法のネットワーク構成は，Generator と被害者モデル・複製モデルから構成される敵対的ネットワークである．既存のデータフローな手法では，画像を訓練したモデルに対する攻撃を暗黙的に仮定しているため，被害者モデルの前処理や複製モデルの前処理を意識する必要がない．加えて，画像には数値変数のみで構成されるため，前処理が Standard Scaling や MinMax Scaling といった線形変換のみとなる．

一方で，表形式を訓練したモデルを複製する際には，各属性ごとにデータドメインが異なるため前処理を意識する必要がある．しかしながら，攻撃者は被害者モデルのデータ前処理がどのように行われているか観測できない．また，生成したデータを被害者モデルに入力するために，攻撃者は Generator が生成したデータに逆変換を適用して表形式データに戻す必要がある．そのため，攻撃者の構成する敵対的ネットワークでは，Generator を訓練するために必要な勾配近似をすることが困難であるので，微分不可な最適化手法で Generator を訓練する必要がある．

アルゴリズム 3 GA Model Extraction

入力: 被害者モデル $M_{\mathcal{V}}$, 複製モデル $M_{\mathcal{A}}$, 公開情報データ: 訓練データの最小値 min_i・最大値 max_i. 目的関数 $OF = Chebyshev\ dist()$
出力: $M_{\mathcal{A}}$.

```
while Q > 0 do
    for i = 1 ··· n_ga do
        z ~ N(0, 1)
        x = G(z; θ_g)
        Execute genetic algorithm GA(OF = OF, x = (M_V, M_A, G), t = query_budget, p = 0.1)
    end for
    for i = 1 ··· n_A do
        z ~ N(0, 1)
        x = G(z; θ_g)
        compute M_V(x), M_A(x), L(M_V(x), M_A(x)), ∇_{θ_A} L
        θ_A = θ_A − η · ∇_{θ_A} L
    end for
end while
```

本研究が提案する手法をアルゴリズム 3 に示す．Generator の最適化に遺伝的アルゴリズムを用いる．つまり，Generator のパラメータを遺伝子として扱い，被害者モデルと複製モデルの出力が大きく異なるデータを生成するように遺伝的アルゴリズムで訓練していく．そして複製モデルは，Generator が被害者モデルと複製モデルの出力が異なるように生成したデータに対し，2 つのモデルの出力が一致するように勾配降下法で訓練することで，被害者モデルのモデルを複製する．

5. 実験

本章では，提案手法の有効性を検討するための実験について説明する．本研究が提案する素朴な手法と遺伝的アルゴリズムを用いた手法を比較することで，表形式データ学習モデルに対する攻撃手法としてそれらが有効か議論する．

5.1 実験設定

5.1.1 モデルのアーキテクチャと訓練

本実験で用いるモデルのアーキテクチャと訓練について説明する．モデルアーキテクチャは，全結合三層のニューラルネットワークを用いており，中間層のノード数は 90 個である．本実験では，被害者モデルのアーキテクチャと複製モデルのアーキテクチャは同一のものを扱う．ハイパーパラメータは，被害者モデルと複製モデルの学習率を 0.01，エポック数を 30 とする．表形式データは数値変数と名義変数からなるので，それぞれの変数に適した前処理を行う．被害者モデルでは，数値変数に対して平均と分散を用いた正規化処理（Standard Scaling）を，名義変数に対して one-hot encoding を行った．

本研究では，2.2.2 節で述べた「補助情報を用いた攻撃」を仮定している．そこで，本実験では攻撃者が被害者モ

表 2: 本実験で用いるデータセット

データセット名	属性数	インスタンス数	クラス数
Adult	14	32600	2
Breast Cancer	32	569	2
Pima Indians diabetes	9	768	2
Arrhythmia	279	452	16

表 3: 被害者モデルの精度

データセット名	被害者モデルの精度
Adult	84.8%
Breast Cancer	99.1%
Pima Indians diabetes	77.9%
Arrhythmia	72.5%

デルの訓練データに関する各属性の「最小値・最大値」を利用できるとした．「最小値・最大値」を利用することで，攻撃者は訓練データドメインを決めることができる．また，複製モデルを訓練する際に行われる前処理は，数値変数に対して最小値と最大値を用いた正規化処理（MinMax Scaling）を，名義変数に対して one-hot encoding を行った．遺伝的アルゴリズムで Generator の最適化を行う際に利用した目的関数には，Chebyshev 距離を用いた．停止条件は，クエリ数が攻撃者のクエリ予算を超えるタイミングで，突然変異が起きる確率は 10%とした．

5.1.2 データセット

本実験で用いるデータセットを表 2 に示す．各データセットの訓練データとテストデータは 8:2 で分割した．各データセットの被害者モデルの精度を表 3 に示す．

5.2 実験結果

本実験の実験結果について述べる．

ランダムクエリと遺伝的アルゴリズムベースの手法の複製モデルの精度とクエリ数の関係をそれぞれ図 2, 3 に示す．ランダムクエリでは 20000 個程度のクエリを被害者モデルに問い合わせることで，被害者モデルに対して 70–80% 程度のモデルの複製ができていることがわかる．遺伝的アルゴリズムベースの手法では，Adult と Arrhythmia がクエリ数を増やすと複製モデルの精度も向上している．しかしながら，その他 2 つのデータセットは，クエリ数を増やしても複製モデルの精度が向上していない．加えて，ランダムクエリと提案手法のクエリ数と比較すると，提案手法はランダムクエリよりも 10-100 倍程度のクエリ数を要していることがわかる．

データセットのクラス数・属性数とそれぞれのクエリ戦略における複製モデルの比較を図 4 に示す．Adult(C=2) は，Adult が 2 クラス分類問題であることを意味する．クラス数が少ない Adult, Breast Cancer, Diabetes ではランダムクエリの複製モデルのほうが提案手法の複製モデルよりも精度が高い．

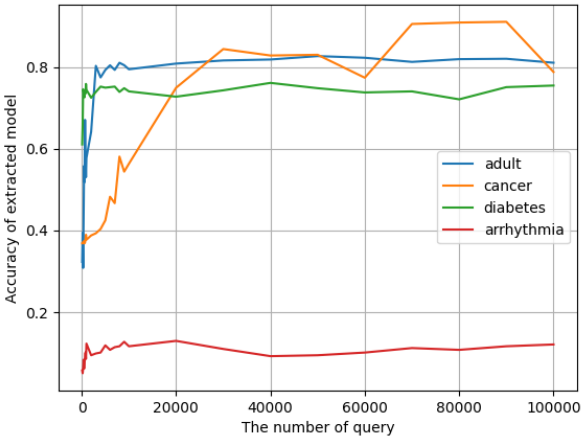

図 2: ランダムクエリの複製モデルの精度とクエリ数

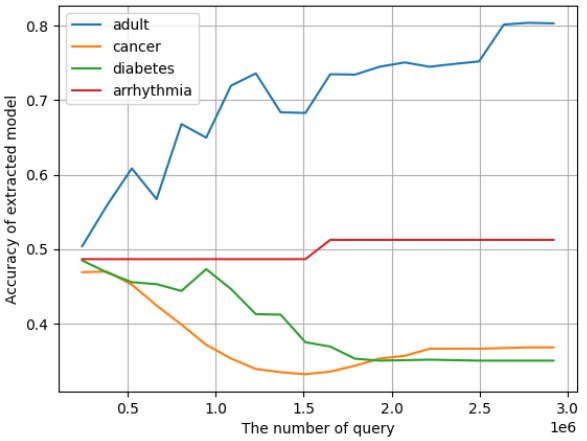

図 3: 提案手法の複製モデルの精度とクエリ数

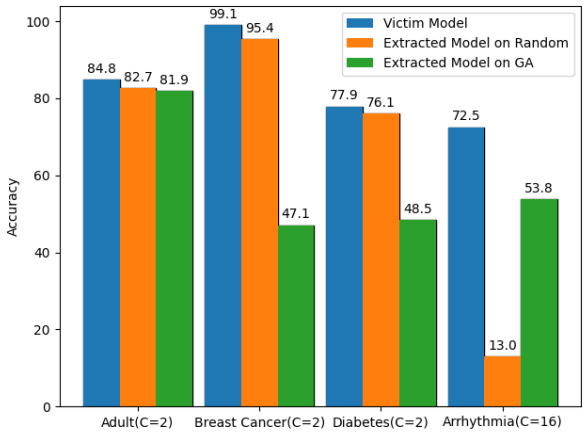

図 4: クラス数と精度の関係

6. 実験結果の考察

5.2 節の結果を踏まえ，表形式データ学習モデルに対する

モデル複製におけるクエリ戦略の性質について考察する．

6.1 ランダムクエリを用いた手法の有効性

素朴な攻撃であるランダムクエリによるモデル複製について述べる．図 2, 4 をみるとわかるように，Adult, Breast Cancer, Diabetes に対して提案手法よりも優れた性質を持っている．20000 個程度のクエリで，被害者モデルに対して，Accuracy が 0.7–0.8 倍程度のモデルを複製することに成功している．表 1 に示した既存手法が要するクエリ数と比較しても，如何にランダムクエリが優れているかがわかる．しかしながら，Arrhythmia に関して言えば複製モデルの精度が 13%に留まっており，データセットの属性数やクラス数が増えると精度が上がらないことが予想される．

6.2 遺伝的アルゴリズムベースの手法の有効性

提案手法である「遺伝的アルゴリズムを用いたモデル複製攻撃」の性質について述べる．実験結果を踏まえると，表形式データ学習モデルに対するモデル複製の観点で，提案手法はランダムクエリよりも劣る場合が多かった．特に，モデル複製攻撃で最も重要な観点である「クエリ数」は，ランダムクエリで複製したモデルと同等のモデルを得るために，10–100 倍程度のクエリ数を要した．その一方で，Arrhythmia に関していえば，ランダムクエリよりも良い精度のモデルを複製している．これらの結果と表 2，3 に示している各データセットの情報を踏まえると，Adult, Breast Cancer, Diabetes のような属性数が比較的少なく，中間層のノード数が 90 個である全結合 3 層ニューラルネットワークで分類精度が 80–90%となるような比較的簡単な問題に対して，提案手法は不向きな可能性が高い．その一方で，属性数が数百個と多く，Adult, Breast Cancer, Diabetes に比べて分類が難しい問題に対して，提案手法は有効な可能性があり，手法改善の余地がある．

7. 結論

本研究では，表形式データ学習モデルに対するモデル複製攻撃として，遺伝的アルゴリズムを用いた新たな攻撃手法を示した．本研究の実験を通し，問題によっては最も素朴なランダムクエリでさえ，被害者モデルに対して，Accuracy が 0.8 倍程度のモデルを複製できることを示した．我々が提案した遺伝的アルゴリズムベースの手法は，データセットの属性数が比較的多く，分類が難しい問題に向いていることが期待される．

参考文献

[1] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R. and Le, Q. V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding, *Proc. of NeurIPS 2019*, Curran Associates, Inc., pp. 5753–5763 (2019).

[2] Herbert-Voss, A.: Practical defenses against adversarial machine learning, *briefing at Black Hat USA* (2020).

[3] Goodfellow, I. J., Shlens, J. and Szegedy, C.: Explaining and Harnessing Adversarial Examples, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Bengio, Y. and LeCun, Y., eds.), (online), available from ⟨http://arxiv.org/abs/1412.6572⟩ (2015).

[4] Tramèr, F., Zhang, F., Juels, A., Reiter, M. K. and Ristenpart, T.: Stealing Machine Learning Models via Prediction APIs, *Proc. of USENIX Security 2016*, USENIX Association, pp. 601–618 (2016).

[5] Juuti, M., Szyller, S., Marchal, S. and Asokan, N.: PRADA: Protecting against DNN Model Stealing Attacks, *Proc. of EuroS&P 2019*, IEEE, pp. 512–527 (2019).

[6] Orekondy, T., Schiele, B. and Fritz, M.: Knockoff Nets: Stealing Functionality of Black-Box Models, *Proc. of CVPR 2019*, IEEE, pp. 4954–4963 (2019).

[7] Truong, J.-B., Maini, P., Walls, R. J. and Papernot, N.: Data-Free Model Extraction, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4771–4780 (2021).

[8] Kariyappa, S., Prakash, A. and Qureshi, M. K.: MAZE: Data-Free Model Stealing Attack Using Zeroth-Order Gradient Estimation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13814–13823 (2021).

[9] Wang, B. and Zhenqiang Gong, N.: Stealing Hyperparameters in Machine Learning, *Proc. of IEEE S&P*, IEEE, pp. 36–52 (2018).

[10] Szyller, S., Duddu, V., Gröndahl, T. and Asokan, N.: Good Artists Copy, Great Artists Steal: Model Extraction Attacks Against Image Translation Generative Adversarial Networks, *CoRR*, Vol. abs/2104.12623 (online), available from ⟨https://arxiv.org/abs/2104.12623⟩ (2021).

[11] Krishna, K., Tomar, G. S., Parikh, A. P., Papernot, N. and Iyyer, M.: Thieves on Sesame Street! Model Extraction of BERT-based APIs, *CoRR*, Vol. abs/1910.12366, pp. 1–18 (online), available from ⟨`http://arxiv.org/abs/1910.12366`⟩ (2019).

[12] Keskar, N. S., McCann, B., Xiong, C. and Socher, R.: The Thieves on Sesame Street are Polyglots - Extracting Multilingual Models from Monolingual APIs, *Proc. of EMNLP 2020*, ACL, pp. 6203–6207 (2020).

[13] Bhagoji, A. N., He, W., Li, B. and Song, D.: Practical Black-box Attacks on Deep Neural Networks using Efficient Query Mechanisms, *Proceedings of the European Conference on Computer Vision (ECCV)* (2018).

[14] Alzantot, M., Sharma, Y., Chakraborty, S., Zhang, H., Hsieh, C.-J. and Srivastava, M. B.: GenAttack: Practical Black-Box Attacks with Gradient-Free Optimization, *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '19, New York, NY, USA, Association for Computing Machinery, p. 1111–1119 (online), DOI: 10.1145/3321707.3321749 (2019).

[15] Chen, J., Su, M., Shen, S., Xiong, H. and Zheng, H.: POBA-GA: Perturbation optimized black-box adversarial attacks via genetic algorithm, *Comput. Secur.*, Vol. 85, pp. 89–106 (online), DOI: 10.1016/j.cose.2019.04.014 (2019).