ポイゾニングを利用したモデル再構築によるセンシティブ情報の復元に関する一考察

披田野 清良[1]　村上 隆夫[2]　清本 晋作[1]　花岡 悟一郎[2]

概要：近年，様々な IT サービスがクラウド化され，機械学習に基づくサービスもまたその潮流を受け，様々な予測サービスがオンラインで提供され始めている．一方で，プライバシー上の問題も懸念されており，2014 年に Fredrikson らにより，システムの出力情報からユーザのセンシティブ情報を復元する Model Inversion 攻撃が提案された．しかしながら，本攻撃では，システムへの入力情報の一部を事前に取得する必要がある．そこで，本稿では，ユーザが学習データを提供できる予測システムを想定し，悪性データを注入して予測モデルを操作することにより，出力情報のみからユーザのセンシティブ情報を復元する新たな攻撃を提案する．

キーワード：センシティブ情報, Model Inversion 攻撃，ポイゾニング

Model Rebuilding Attacks by Poisoning for Exploiting Sensitive Information

Seira Hidano[1]　Takao Murakami[2]　Shinsaku Kiyomoto[1]　Goichiro Hanaoka[2]

Abstract: While online prediction services using machine learning are rapidly gaining momentum, there are concerns about privacy issues. Model inversion attacks proposed by Fredrikson at el. exploited sensitive information of users from output data from the system. However, this attack requires adversaries to obtain a portion of personal information of the target user in advance. In this paper, focusing on prediction systems to which users can provide their own data as training data, we propose a new attack that rebuilds the prediction model by poisoning malicious data and exploits the sensitive information from only output data.

Keywords: Sensitive Information, Model Inversion Attacks，Poisoning

1. はじめに

近年，様々な IT サービスがクラウド化され，機械学習に基づくサービス等もまたその潮流を受け，様々な予測サービスがオンラインで提供され始めている．しかしながら，その一方で，個人情報をクラウドに提供することへのプライバシー上の問題も懸念されており，2014 年に Fredrikson らにより，機械学習に基づく予測モデルをブラックボックとして利用し，ユーザのセンシティブ情報を復元する Model Inversion 攻撃が提案された [1]．本攻撃では，ユーザが予測モデルを利用した際の出力情報を事前に取得し，予測モデルを利用して入力情報の候補の絞り込みを行う．本攻撃は提案時は予測モデルとして線形回帰モデルを想定していたが，2015 年に Fredrikson らにより非線形の予測モデルにも適用された [2]．しかしながら，Fredrikson らの Model Inversion 攻撃では，攻撃対象のユーザの出力情報だけでなく，入力情報のうち復元したいセンシティブな属性以外のすべての属性も事前に取得する必要があった．センシティブな属性ではないもののそれらの属性がユーザに関わる情報であれば事前にすべてを取得することは難し

[1] KDDI 総合研究所
KDDI Research
[2] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

く，Fredrikson らの Model Inversion 攻撃は攻撃者にとっていくぶん有利な設定であった．

また，機械学習のセキュリティに関する研究事例としては，ポイゾニングと呼ばれるアプローチがある [3], [4]．これらの研究では，機械学習に基づく異常検知システム等の撹乱を目的とし，学習性能を著しく劣化させる悪性データを学習データとして学習システムに注入することによりこれを実現する．IoT パラダイムへの関心の高まりから，IoT データの価値は高まる一方であり，昨今ユーザがサービス提供者へ自身のデータを提供する機会が増えている．このため，ポイゾニングは今後非常に大きな脅威となりうる攻撃手段の 1 つであると考えられる．そこで，本稿では，上述した Fredrikson らの Model Inversion 攻撃の問題を解決する 1 つの手段として，ユーザが自身のデータを学習データとして提供できる予測システムを想定し，ポイゾニングにより予測モデルを操作することについて考える．

1.1 貢献

本稿では，機械学習に基づく予測モデルを利用した予測システムについて，システムへの入力情報は一切に使わずに，出力情報のみからセンシティブ情報を復元する新たな Model Inversion 攻撃を提案する．本研究の貢献は以下の通りである．

- 出力情報のみからセンシティブ情報を復元する Model Inversion 攻撃の新たな攻撃モデルを提案する．提案する攻撃モデルでは，予測モデルを利用してユーザの入力に対して出力を返す機能と，ユーザが提供する学習データを用いて予測モデルを順次更新する機能を持った予測システムを想定し，攻撃者は学習データとして悪性データを注入することにより予測モデルを再構築する．
- 予測モデルとして線形回帰モデルを想定し，具体的なポイゾニングのアルゴリズムを提案する．出力情報のみを用いて Model Inversion 攻撃を行うための予測モデル（ターゲットモデル）として，入力情報の非センシティブな属性が出力に寄与しないモデルを選択した場合，出力情報のみから Model Inversion 攻撃が可能であることを明らかにする．また，線形回帰モデルの学習アルゴリズムとして SGD (Stochastic Gradient Descent) を想定した場合，回帰係数の更新式を利用して悪性データを作成することにより，少量の悪性データで上記のターゲットモデルを構築できることを示す．
- 実データを用いた評価実験により，上記の方法で実際に少量の悪性データでターゲットモデルを構築できることを示す．

また，本稿の構成は以下の通りである．まず，2 章で本研究の関連研究について概説する．次いで，3 章でポイゾニングを利用した新たな Model Inversion 攻撃の攻撃モデルを提案する．そして，4 章で線形回帰モデルを想定した具体的なポイゾニング方法を提案するとともに，5 章で評価実験を通してその有用性を示す．最後に，6 章で本稿のまとめと今後の研究課題について述べる．

2. 関連研究

本稿で提案する攻撃の関連研究として，Model Inversion 攻撃と学習システムへのポイゾニングについて概説する．

2.1 Model Inversion 攻撃

Model Inversion 攻撃は，2014 年に Fredrikson らにより提案された機械学習に基づく予測モデルを利用してセンシティブ情報を復元する攻撃の 1 つである [1], [2]．本攻撃では，何らかの予測モデルに基づくユーザの出力情報を事前に取得し，予測モデルを利用して入力情報のセンシティブな属性を復元することを目的とする．具体的には，予測モデル f の入力情報を $(x_1, \ldots, x_T, \ldots, x_D)$，出力情報を y として以下の手順でセンシティブな属性の復元を試みる．ただし，$(x_1, \ldots, x_T)$ をセンシティブな属性，$(x_{T+1}, \ldots, x_D)$ を非センシティブな属性とする．

Step 1. 攻撃対象ユーザの非センシティブな属性 $(x_{T+1}, \ldots, x_D)$ および出力情報 y を取得する．また，入力情報の各属性の事前分布 $(p_1, \ldots, p_D)$ を取得する．

Step 2. 予測モデル f を利用して，Step 1 で取得した $(x_{T+1}, \ldots, x_D)$ と組み合わせて y を出力するセンシティブな属性 $(x_1, \ldots, x_T)$ の値を抽出する．複数の組み合わせがある場合はそれらすべての値を候補として記録する．

Step 3. Step 2 で得られた $(x_1, \ldots, x_T)$ の候補のそれぞれについて，$(p_1, \ldots, p_D)$ を利用して出力情報 y についての尤度を計算し，もっとも可能性の高い候補を $(x_1, \ldots, x_T)$ の復元情報とする．

Fredrikson らは，上記の攻撃モデルを，文献 [1] で線形回帰モデルに，文献 [2] で決定木およびニューラルネットワーク等の非線形モデルに適用している．いずれの場合においても，実データによる評価実験を通して，単に事前分布 $(p_1, \ldots, p_D)$ だけを用いてセンシティブ情報を復元する方法に比べて，攻撃性能が向上することが示されている．ただし，Fredrikson らの攻撃モデルでは，攻撃者が攻撃対象ユーザの非センシティブな属性 $(x_{T+1}, \ldots, x_D)$ を事前に取得できることを前提としており，攻撃者にとって非常に優位な状況下での攻撃を想定している．そこで，本稿では，Fredrikson らのアプローチとは異なり，非センシティブな属性を一切使わずに，出力情報のみから Model Inversion 攻撃を行う方法を考える．

2.2 ポイゾニング

学習システムに対する攻撃の 1 つであるポイゾニングは，

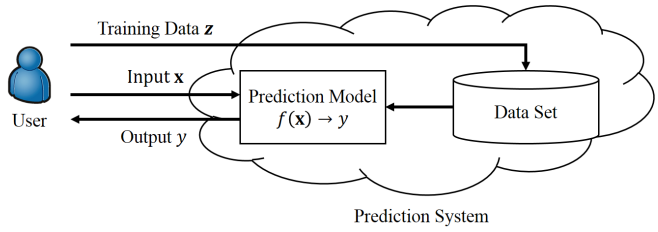

図 1 予測システム

機械学習を利用したセキュリティ検知機能等の撹乱を目的とし，学習システムの性能を著しく低下させる悪性データを学習データとしてシステムに注入する [5]．代表的な研究事例としては，Biggio らによるサポートベクタマシーンに対する攻撃があげられる [3], [4]．Biggio らは少量の悪性データでサポートベクタマシーンの予測性能を著しく劣化させるためのポイゾニングによる攻撃アルゴリズムを提案している．しかしながら，それはあくまで予測性能の劣化を目的としており，本稿で想定するポイゾニングとは目的が相反する（詳細については 4.2 節で述べるが，本稿で想定するポイゾニングはポイゾニングの前後で精度が変わらないことが望ましい）．また，Biggio らの手法はサポートベクタマシーンに特有の性質を利用したものであり，本稿で想定する線形回帰モデルへの適用は難しいと考えられる．

Cao らによるデータを削除できる学習システムもまた学習データの操作に関する研究事例である [6]．本研究では，プライバシーの観点からユーザが提供したデータを削除できることを目的とし，データが削除された後に高速で再学習を行う仕組みを提案している．しかしながら，データの削除の場合，特異なデータを攻撃に利用することが難しく，恣意的に悪性データを追加できるポイゾニングに比べて，目的の学習モデルを作成するのが難しいと考えられる．このため，本稿では，データの操作としてデータの削除については考えず，データの追加のみを想定し，予測モデルを再構築する方法について考える．

3. ポイゾニングを利用した出力情報のみからの Model Inversion 攻撃

本章では，ユーザが学習データを提供できる予測システムを想定し，悪性データを学習データとしてシステムに注入することにより，予測システムの出力情報のみからユーザのセンシティブ情報を復元する新たな攻撃のフレームワークを提案する．以下，コンテキストに応じて悪性データを学習データとしてシステムに注入することをポイゾニングと表現する．

3.1 予測システム

本稿で対象とする予測システムの概要を図 1 に示す．予測システムは，ユーザの学習データを用いて何らかの予測モデル f を構築するとともに，オンラインでユーザに f に基づくサービスを提供する．ただし，ユーザは f の学習アルゴリズム（線形回帰，決定木，ニューラルネットワーク等）のみ知ることができ，具体的なパラメータはブラックボックスとしてアクセスする．予測システムはユーザに対して少なくとも以下の 2 つの機能を提供する．

- ユーザが情報 $\mathbf{x}$ をシステムに入力した場合，予測モデル f に基づき情報 y を計算し，y を出力情報としてユーザに返す．ただし，入力情報 $\mathbf{x}$ は，ユーザのセンシティブな属性 $(x_1, \ldots, x_T)$ と，非センシティブな属性 $(x_{T+1}, \ldots, x_D)$ からなるものとする．以下，本機能を機能 1 と呼ぶ．
- ユーザは自身のデータを学習データ $\mathbf{z} = (\mathbf{x}_z, y_z)$ としてシステムに提供できる．新たな学習データが提供された場合，予測システムは予測モデル f を順次更新する．以下，本機能を機能 2 と呼ぶ．

3.2 攻撃モデル

攻撃者は，予測システムをブラックボックスとして利用し，あるユーザが予測システムを利用した際の出力情報 y から，そのユーザの入力情報のセンシティブな属性 $(x_1, \ldots, x_T)$ を復元することを目的とする．ただし，本攻撃では，Fredrikson らの攻撃モデル [1], [2] とは異なり，予測システムはセンシティブな属性値の候補を絞り込むために用いるが，非センシティブな属性は一切使わないものとする．

攻撃者は以下の手順によりユーザのセンシティブな属性 $(x_1, \ldots, x_T)$ を復元する．

Step 1. 予測システムの機能 1 の入出力情報を利用して現在の学習モデル f を推定する．

Step 2. 出力情報 y と予測システムのみからセンシティブな属性 $(x_1, \ldots, x_T)$ を復元するための予測モデル f'（以下，ターゲットモデル）を決定する．

Step 3. 現在のモデル f を f' に近づけるための悪性データを作成し，それを予測システムの機能 2 を利用して，学習データ $\mathbf{z} = (\mathbf{x}_z, y_z)$ としてシステムに提供する．

Step 4. 予測モデルがターゲットモデル f' になるまで，Step 3 を繰り返す．

Step 5. 攻撃対象のユーザが入力情報 $(x_1, \ldots, x_D)$ を用いて予測システムの機能 1 を利用した際の出力情報 y を取得する．

Step 6. 予測システムの機能 1 を利用して出力情報が y となるセンシティブな属性値の候補を抽出し，それらの候補から何らかの方法により復元情報を一意に決定する．ただし，候補が 1 つしかない場合は，その値を復元情報とする．

以下，Step 1 から Step 4 の処理を「モデル再構築」と呼ぶ．Step 5 および Step 6 の処理は，Model Inversion 攻撃だが，[1], [2] とは異なり，（非センシティブな属性を使わ

ずに）出力情報のみからセンシティブな属性を復元する．また，攻撃者は f の学習に用いたデータセットに含まれるすべてのデータもしくはその一部か，それに類するものを入手できるものとする．

4. 線形回帰モデルに対するポイゾニング

本章では，予測モデル f として，線形回帰モデルに着目し，3.2 節で述べた出力情報のみからセンシティブ情報を復元をするための具体的なターゲットモデルを明らかにするとともに，そのターゲットモデルを構築するためのポイゾニング方法を提案する．

4.1 線形回帰モデル

本章では，予測システムへの入力情報 $\mathbf{x}$ を次のように定義する．

$$\mathbf{x} = (1, x_1, \ldots, x_T, \ldots, x_D)^T \tag{1}$$

ただし，3.2 節と同様に，$(x_1, \ldots, x_T)$ をセンシティブな属性，$(x_{T+1}, \ldots, x_D)$ を非センシティブな属性とする．

このとき，線形回帰モデル f は次式で表せる．

$$f = w_0 + w_1 x_1 + \ldots + w_T x_T + \cdots + w_D x_D \tag{2}$$

$$= \mathbf{w}^T \mathbf{x} \tag{3}$$

ただし，$\mathbf{w}$ は回帰係数とする．

SGD (Stochastic Gradient Descent)

3.1 節で述べたように，予測システムは，ユーザより新たな学習データが提供された場合，予測モデル f を順次更新する．このため，本稿では，予測モデルの逐次的な更新を実現するオンライン学習アルゴリズムの 1 つである確率的勾配降下法（SGD：Stochastic Gradient Descent）[7] を f の学習アルゴリズムとして想定する．すなわち，学習データ $(\mathbf{x}, y)$ が与えられたとき，回帰係数 $\mathbf{w}$ は次式を用いて更新する．

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \eta(y - \mathbf{w}^{(\tau)T}\mathbf{x})\mathbf{x} \tag{4}$$

ただし，τ は繰り返しの回数であり，η は学習率とする．

4.2 ターゲットモデルの決定

線形回帰モデルにおいて，非センシティブな属性を使わずに，出力情報のみからセンシティブな属性を復元するためのターゲットモデルについて考える．予測モデル f において非センシティブな属性が出力にまったく寄与しない場合，出力値はセンシティブな属性から一意に決定すると考えられる．(3) 式より，$w_{T+1} = \cdots = w_D = 0$ のとき，予測モデル f は非センシティブな属性が出力に寄与しない．したがって，本稿では，次式のモデルをターゲットモデル f' とする．

$$f' = w_0 + w_1 x_1 + \cdots + w_T x_T \tag{5}$$

また，ポイゾニングにより学習モデルの予測精度が著しく劣化する場合，攻撃が行われていることを検知される可能性がある．このため，ポイゾニングに使う悪性データは，以下の 2 つの条件を満たすように作成する．

- $w_{T+1} = \cdots = w_D = 0$
- 予測モデルの性能を著しく劣化させない．

また，センシティブな属性が 1 つだけの場合，その属性を x_1 とすると，ターゲットモデル f' は次式で表せる．

$$f' = w_0 + w_1 x_1 \tag{6}$$

(5) 式では，同一の y に対して $(x_1, \ldots, x_T)$ の候補が複数存在する場合があるのに対し，(6) 式では，x_1 は一意に決まる．センシティブな属性数が少なければ，Model Inversion 攻撃の成功確率は上昇するが，この場合，ターゲットモデル f' に関与する属性数も減るため，（特にモデルへの寄与率が高い属性を削除する場合）予測精度が著しく劣化する可能性がある．このため，5 章において，センシティブな属性数とポイゾニング後の予測モデルの予測精度の関係についての調査結果を示す．

4.3 予測モデルの再構築

ポイゾニングにより現在のモデル f から (5) 式のターゲットモデル f' を構築する方法について述べる．モデルの再構築は，3.2 節で示したように，予測システムの機能 1 を利用して予測モデル f のパラメータを推定した後に，予測システムの機能 2 を利用して予測システムに悪性データを繰り返し注入することにより行う．

パラメータの推定

予測システムを利用して回帰係数 $\mathbf{w}$ と学習率 η を推定する方法について述べる．回帰係数 $\mathbf{w}$ は予測システムに入力情報 $\mathbf{x}$ を複数回に入力し，それらの出力結果から推定する．具体的な手順を以下に示す．

Step 1. $\mathbf{x} = (1, 0, 0, ..., 0)^T$ を入力情報として予測システムに入力し，出力結果として $y = w_0$ を得る．

Step 2. $\mathbf{x} = (1, 1, 0, ..., 0)^T$ を入力情報として予測システムに入力し，出力結果として $y = w_0 + w_1$ を得る．

Step 3. Step 1 および Step 2 と同様に，$\mathbf{x}$ を $(1, 0, 1, ..., 0)^T, ..., (1, 0, 0, ..., 1)^T$ と変化させながら入力情報として予測システムに入力し，出力結果 y として $w_0 + w_2, ..., w_0 + w_D$ を得る．

Step 4. Step 1 から Step 3 で得られた $w_0, w_0 + w_1, ..., w_0 + w_D$ から $\mathbf{w}$ を算出する．

学習率 η は任意のデータを学習データとして予測システムに提供し，学習モデルが更新される前後の回帰係数 $\mathbf{w}^{(\tau)}, \mathbf{w}^{(\tau+1)}$ と，(4) 式を用いて算出する．ただし，$\mathbf{w}^{(\tau)}$ および $\mathbf{w}^{(\tau+1)}$ の推定は，上述した方法で行う．

ポイゾニング

ターゲットモデルを得るための悪性データ $\mathbf{z}' = (\mathbf{x}'_z, y'_z)$

(ただし，$\mathbf{x}'_z = (x'_{z_1}, \ldots, x'_{z_T}, \ldots, x'_{z_D})$ の作成方法について述べる．まず，$(x'_{z_1}, \ldots, x'_{z_T})$ および y'_z については，4.2 節の 2 つ目の条件を考慮して，ポイゾニング前のモデル f の学習に用いられたデータセット，もしくはそれに類するデータセットのデータを利用する．そして，この条件下で，4.2 節の 1 つ目の条件 $w_{T+1} = \cdots = w_D = 0$ を考慮して，残りの $(x'_{z_{T+1}}, \ldots, x'_{z_D})$ の値は以下の $D - T$ 個の連立方程式を解くことにより算出する．

$$0 = w_j + \eta y' - \mathbf{w} x'_j \ (T+1 \leq j \leq D) \qquad (7)$$

以上より，線形回帰モデルに対するポイゾニングは以下の手順で行う．

Step 1. 既存のデータセットからランダムにデータを選択し，悪性データ $\mathbf{z}' = ((x'_{z_1}, \ldots, x'_{z_T}, \ldots, x'_{z_D}), z'_y)$ を生成する．

Step 2. (7) 式の連立方程式を解き，得られた解と，Step 1 で作成した $\mathbf{z}'$ の該当の属性値 $(x'_{z_{T+1}}, ..., x'_{z_D})$ を置き換える．ただし，得られた解のうち，各属性のドメイン $[x_j^{min}, x_j^{max}]$ を超えるものは，属性値 x'_{z_j} を最大値 x_j^{max}，もしくは最小値 x_j^{min} で置き換える．

Step 3. Step 2 で作成した悪性データ $\mathbf{z}'$ を学習データとして予測システムに提供する．

Step 4. $w_{T+1} = \cdots = w_D = 0$ となるまで，Step 2 および Step 3 の処理を繰り返す．

5. 評価実験

4 章で示した線形回帰モデルに対するポイゾニングについて，実データを用いたシミュレーション実験を通して攻撃性能を評価した．以下，実験諸元と本実験の評価結果について述べる．

5.1 実験諸元

本実験では，データセットとして，FiveThirtyEight の“ How Americans Like Their Steak ”[8] を使用した．本データセットは，複数のアメリカ人に対して実施されたステーキの焼き加減についての調査結果に関するものであり，焼き加減の好み，被験者の情報，ならびにいくつかのアンケート結果が紐付けられて収録されている．それらのデータのうち，表 1 に本実験に使用した属性と各属性の取りうる値を示す．ただし，実験には表 1 の属性について欠損値を含まない 335 人のデータを使用した．また，属性値は実験時には離散値で置き換え，各属性の平均が 0 となるように正規化した．(1) 式の入力情報 $\mathbf{x}$ の各属性および出力情報 y との対応関係についても表 1 に示す．ただし，$D = 10$ とし，T は変化させながら実験を行った．(3) 式の線形回帰モデル f の学習には 200 人のデータを使用しており，それ以外のデータは評価データとした．また，攻撃用のデータは学習に使用したデータからランダムに選択し，4.3 節の方法で作成した．(4) 式の学習率は $\eta = 0.01$ の固定値とした．上記の条件の下，予備実験を行った結果，f の予測精度は 95.3%であった．

表 1 データセット

属性名	属性値	$(\mathbf{x}, y)$ との対応
性別	女性，男性	x_{10}
年齢	18–29，30–44，45–60, 60 以上	x_9
学歴	高卒未満，高卒，短大卒，大卒，院卒	x_8
宝くじの好み	低確率高額当選，高確率少額当選	x_7
喫煙の経験	No，Yes	x_6
飲酒の経験	No，Yes	x_5
ギャンブルの経験	No，Yes	x_4
スピード超過の経験	No，Yes	x_3
詐欺の経験	No，Yes	x_2
所得	\$0–\$24,999，\$25,000–\$49,999，\$50,000–\$99,999，\$100,000–\$149,999，\$150,000 以上	x_1
ステーキの焼き加減の好み	レア，ミディアムレア，ミディアム，ミディアムウェルダン，ウェルダン	y

表 2 実験結果

センシティブな属性数	悪性データ数	ポイゾニング後の予測精度
1	9	95.3%
2	10	95.3%
3	10	95.3%
4	10	95.3%
5	10	95.3%
6	10	95.3%
7	10	95.3%
8	10	95.3%
9	1	95.3%

5.2 評価結果

4.3 節の方法で実際に予測システムへのポイゾニングを試み，モデルを再構築するために必要な悪性データ数について調査した．表 2 に実験結果を示す．本実験は，T の値を 1 から 9 まで変化させ，センシティブな属性数を変えて複数回実施した．また，ポイゾニング後の予測精度も表 2 に示す．直感的ではあるが，9，10 回のポイゾニングは十分に現実的な回数といえる．たとえ一人の攻撃者が提供できる学習データの数に制限があるような状況を想定したとしても，9，10 人の攻撃者が結託してポイゾニングを行う

ことは難しくない．

また，Fredrikson らの Model Inversion 攻撃では，非センシティブな属性は事前に取得する必要があるため，非センシティブな属性数が多いほど攻撃に関わるコストも増加すると考えられる．一方，本攻撃の場合，モデルの再構築に必要な悪性データ数についてはセンシティブな属性数が 9 のとき（非センシティブな属性数が 1 のとき）を除き，ほとんど変化が見られなかった．したがって，本攻撃は特に，非センシティブな属性数が多いときに，従来手法と比べて有用であるといえる．

また，予測精度についてもポイゾニングの前後で一切変化しなかった．4.2 節では，ポイゾニング後に予測精度が著しく劣化する場合，その変化に基づき攻撃を検知できる可能性があることを示唆したが，上記の結果より，そういった対策が必ずしも有効でないことが分かる．ただし，同様に 4.2 節で述べたように，モデルに関与する属性数が減れば，多くの場合それに合わせて精度も劣化するため，この結果は本実験に使用したデータセットに特有な事象である可能性も捨てきれない．このため，今後は上記の結果について理論的に考察するとともに，異なるデータセットを用いて比較実験を実施する．

6. おわりに

本稿では，機械学習に基づく予測システムへの出力情報から，予測システムを利用して入力情報のセンシティブな属性を復元する Fredrikson らの Model Inversion 攻撃に着目し，学習データとして悪性データを注入するポイゾニングと呼ばれるアプローチを導入することにより，補助的な情報を用いない新たな Model Inversion 攻撃を提案した．Fredrikson らの Model Inversion 攻撃では，入力情報の非センシティブな属性をすべて事前に取得する必要があったが，提案手法では出力情報のみからセンシティブな属性を復元できる．また，予測モデルとして線形回帰モデルを想定し，出力情報のみからセンシティブな属性を復元するための予測モデル（ターゲットモデル）を明らかにするとともに，ターゲットモデルを構築するための具体的なポイゾニング方法を提案した．さらに，実データを用いた評価実験により，わずか 9 回程度のポイゾニングでターゲットモデルを構築できることを確認した．また，Fredrikson らの手法では非センシティブな属性数が多いほど事前に取得しなければ属性数が増えるのに対して，本手法では非センシティブな属性数が増えてもポイゾニングの回数は一定であった．この結果から，本稿で提案する攻撃モデルは，非センシティブな属性数が多い入力情報に対して特に有用であるといえる．

今後は，まず異なるデータセットを用いて提案手法の有用性のさらなる検証を行う．次いで，ポイゾニングを利用した Model Inversion 攻撃を線形回帰モデル以外の予測モデルへ適用するとともに，対策についてもあわせて検討する．

参考文献

[1] Fredrikson, M., Lantz, E. and Jha, S.: Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, *the Proceedings of the 23rd USENIX Security Symposium (USENIX 2014)*, pp. 17–32 (2014).

[2] Fredrikson, M., Jha, S. and Ristenpart, T.: Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS 2015)*, pp. 1322–1333 (2015).

[3] Biggio, B., Nelson, B. and Laskov, P.: Poisoning Attacks against Support Vector Machines, *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)* (2012).

[4] Biggio, B., Fumera, G. and Roli, F.: Security Evaluation of Pattern Classifiers under Attack, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 4, pp. 984–996 (2014).

[5] Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I. P. and Tygar, J. D.: Adversarial Machine Learning, *In Proceedings of 4th ACM Workshop on Artificial Intelligence and Security (AISec 2011)* (2011).

[6] Cao, Y. and Yang, J.: Towards Making Systems Forget with Machine Unlearning, *Proceedings of the 36th IEEE Symposium on Security and Privacy (S&P 2015)*, pp. 463–480 (2015).

[7] Bishop, C.: *Pattern Recognition and Machine Learning*, Springer (2010).

[8] Hickey, W.: FiveThirtyEight: How Americans Like Their Steak, http://fivethirtyeight.com/datalab/how-americans-like-their-steak/ (2014).