公募シンポジウム
# 公募シンポジウム8
# 医療情報の Public Use File提供にむけた検討
2018年11月25日(日) 09:00 〜 10:30 B会場 (4F 409+410)

[4-B-1-3] 医療分野における人工知能セキュリティ・プライバシー
○佐久間 淳 （筑波大学大学院システム情報工学研究科）

近年の人工知能技術の急速な発展により、近い将来における様々な領域での人工知能技術の応用可能性が期待されている。特に大量のデータを入力として構築される深層学習は機械による画像や音声の認識能力の獲得という意味において、人間による認識能力を上回るほどの能力を獲得しつつある。一方で、機微な情報を用いて人工知能を学習させた場合に、その人工知能をなんらかの意味でプライバシーを侵害しないかどうかは事前に検討しておく筆御湯がある。また、悪意を持つ者が外部から敵対的な入力を人工知能の学習データやテストデータとして与えた場合、人間の認識とはかけはなれた挙動を示すことがあり、そのような人工知能の「バグ」を利用して、人工知能による認識や意思決定の結果を恣意的に操るようなセキュリティ上の懸念がある。講演では人工知能における標準的なセキュリティ・プライバシー上の問題点を紹介するとともに、人工知能の医療応用における課題を示す。

医療分野における人工知能セキュリティ・プライバシー

佐久間 淳[*1*2]
*1 筑波大学 システム情報工学研究科 *2 理化学研究所 革新知能統合研究センター

AI Security and Privacy in Medical Domain

Jun Sakuma [*1*2]
*1 University of Tsukuba, Graduate School of SIE, *2 RIKEN AIP.

Abstract

Machine learning and deep learning (Deep Neural Network, DNN) have achieved state-of-the-art in various types of recognition tasks, such as speech recognition and natural language processing, as well as image classification. When DNN is trained with data samples containing personal or sensitive information, we need to pay careful attention to the possibility of privacy breach through utilization or distribution of the DNN. In this paper, we discuss the risks of privacy breach caused by distributing predictions or models of AI by taking biomedical applications of AI as examples.

Keywords: security, privacy, medical application, AI

1. はじめに

機械学習や深層学習(Deep Neural Network, DNN)は画像分類にとどまらず, 音声認識, 自然言語処理などの分野で既存の手法を上回る成果を挙げている. 一方で, 機微な情報を用いて人工知能を学習させた場合に, その人工知能がなんらかの意味でプライバシーを侵害しないかどうかは事前に検討しておく必要がある. また, 悪意を持つ者が外部から敵対的な入力を人工知能の学習データやテストデータとして与えた場合, 人間の認識とはかけはなれた挙動を示すことがあり, そのような人工知能の「バグ」を利用して, 人工知能による認識や意思決定の結果を恣意的に操るようなセキュリティ上の懸念がある. 本稿では, そのような機械学習を医療応用した場合における問題点について考察する.

敵対的設定では, 機械学習システムに対する悪意を持った攻撃者を仮定し, その攻撃者からのシステムに対する攻撃を想定する. 想定される攻撃の種類は, それを用いてDNN を学習させた場合に DNN になんらかの影響(ある特定のデータを与えた時に, 本来とは異なる特定のラベルとして予測する等)を与えることを意図して設計された poisoning 攻撃[1] や, DNN そのものやその予測値から学習に用いた訓練データや予測に用いたテストデータを推定する逆推定攻撃[2], データに何らかの細工を加え, そのデータを DNN に与えた時に本来とは異なる認識結果が与えられるようなデータを作成する敵対的サンプル[3]などがある. 本稿では特にモデルの逆推定攻撃が, 機械学習を用いた医療サービスに与える影響について考察する.

2. 予測値からの逆推定

個人に関する情報の利用が進み, 個人に関する情報を入力値として用い個人に特化した予測を行なうサービスも広がりつつある. 例として, 遺伝情報から薬剤の投与量を予測するアプリケーションなどが存在する.

個人に関する情報を入力値として統計モデルに基づき予測を行なうサービスについて, ユーザーが予測値を公開する・第三者が予測値を盗み見るなどし, 第三者が予測値を得た場合, 予測値から入力値を推定されるリスクが存在する. 予測値から入力値が容易に推定可能であるならば, その予測値は入力値と同等のセンシティブな情報として扱う必要がある. 例えば, 遺伝情報から疾患罹患リスクを予測するサービスを考える. ある個人において, サービスの予測した疾患罹患リスクから入力値である遺伝情報が容易に推定できたとすると, この疾患罹患リスクは遺伝情報と同等のセンシティブな情報になりうる. 一方, 容易に推定できないのであれば入力値と同等にセンシティブな情報として扱う必要は必ずしもない. このように予測値をどの程度のセンシティブな情報として扱うかを判断するために, 予測値から入力値がどの程度推定されるか評価する必要がある.

[4]では複数の臨床情報(年齢, 性別, BMI, 喫煙歴の有無, 血中クレアチニン濃度, 糖尿病の有無・高トリグリセリド血症の有無・低 HDL コレステロール血症の有無・高 LDL コレステロール血症の有無・高血圧症の有無)および 13 個の SNP を用いて, 脳出血の発症・罹患リスク(以降, 罹患リスク)を評価する線形モデルを用いて推定した脳出血の罹患リスクを求めた時に, その数値的リスクを A-E の五段階に丸めたとしても, それを公開した場合, 2 つの SNP が一意に特定されるケースが存在することを報告した. また同じ五段階表示であっても, その階級幅を適切にコントロールすることでそのリスクは大幅に軽減できることを報告した. このように, 機械学習によって機微な情報を用いて何らかの予測を行った場合, その予測値自体が機微な情報になり得ることに注意する必要がある.

3. 学習モデルからの逆推定

深層学習の予測モデルは, 訓練データを入力とした学習アルゴリズムによって構築する. 例として, 顔画像から個人を識別するようなモデルでは, 入力顔画像が対象の個人と識別される確率を出力する分類モデルを利用する.

深層学習技術を利用したアプリケーションにおいて, サービス上の制約により予測モデルを利用者に公開する必要 g あある場合が場合が存在する. 例えば個人の機微な情報を用いて画像診断を行う深層学習モデルでは, その利用者はローカル環境で予測モデルにアクセスできる必要があるため, 予測モデルを提供する必要がある

予測モデルから, その学習に利用したデータが復元されることをモデル逆推定攻撃と呼ぶ. 一般的に深層学習に対するモデル逆推定は困難であると考えられていた. 画像や音

声のような高次元なデータの生成分布を正確に推定することがそもそも困難な問題であるからである．しかし[5]では攻撃者が対象ドメインについて（訓練データとは無関係であるが，同様度ドメインから収集された）ある程度の量のデータを保持している場合，深層学習モデルからその訓練データに非常に近いデータを抽出できる場合があることを示した．例えば，医療画像診断における予測モデルを患者の画像で学習した場合には，その患者の画像に近い画像が予測モデルから復元可能なケースがあることを示した．このような予測モデルが一般公開されることは考えにくいため，直ちにリスクがあるとは考えられないが，機微な情報で学習させたモデルは，機微な情報を抽出させる可能性があることは，モデルを第三者に譲渡する場合に留意する必要がある．

4. おわりに

本稿では機械学習の医療応用にけるセキュリティ・プライバシーの話題として，機微な情報から機械学習によって得られた予測や，機微な情報を訓練データとして得た機械学習モデルから，元の機微な情報あるいはそれに近い情報が逆推定されるケースについて例示した．今後，機械学習が医療応用される場合にはそのようなリスクについて留意する必要がある．

参考文献

1) Battista Biggio, Blaine Nelson, and Pavel Laskov. ICML2012, pp. 1467-1474, 2012.
2) Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures, ACMCCS2015, pp. 1322-1333, 2015.
3) Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. ECML2013, pp. 387-402, 2013.
4) Kosuke Kusano, Ichiro Takeuchi, Jun Sakuma, Privacy-preserving and Optimal Interval Release for Disease Susceptibility, AsiaCCS 2017, pp. 532-545, 2017
5) 草野 光亮, 佐久間 淳, Generative Adversarial Networks を用いた深層学習モデルに対する Concept Extraction 攻撃, コンピュータセキュリティシンポジウム 2017.