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

メンバーシップ推論攻撃に対する交差蒸留を利用した防御手法
Defense method using knowledge distillation against membership inference attacks

Rishav Chourasia * Batnyam Enkhtaivan † 伊東 邦大 † 森 隼基 †
Rishav Chourasia Batnyam Enkhtaivan Kunihiro Ito Junki Mori

寺西 勇 † 土田 光 †
Isamu Teranishi Hikaru Tsuchida

キーワード メンバーシップ推論攻撃、蒸留、機械学習、プライバシー

あらまし

機械学習は現在，社会の様々な場面で幅広く利用されており，特に画像認識や自然言語処理などの分野で大きな進歩を遂げている．しかし，近年機械学習モデルを解析することにより，学習データのプライバシーが損なわれることが報告されている．機械学習の実社会への応用場面の多くでは，学習データとしてセンシティブなデータが使用されるため，学習データのプライバシーを保護することはデータ提供者や社会からの承認を得るためにも必要不可欠である．

機械学習モデルのプライバシーに対する最も基本的な攻撃の 1 つとして，メンバーシップ推論攻撃が知られている [1]．メンバーシップ推論攻撃は，あるデータが機械学習モデルの学習データに含まれているかどうかを推測する攻撃である．メンバーシップ推論攻撃が脅威となる例は次のような場合である．癌患者のデータからある薬に対する反応を推測する機械学習モデルを考えよう．その機械学習モデルへのアクセス権をもつメンバーシップ推論攻撃者は，癌か否かの直接的な情報を持たないある患者データに対して攻撃を行うことで，その患者がモデルの学習に使用されたかどうか，すなわち癌かどうかを知ることができ，プライバシーの問題が生じる．

メンバーシップ推論攻撃に対しては様々な防御手法が提案されているが，現在精度とプライバシー保護の面で最も有効である方式は，蒸留を利用した *Distillation for Membership Privacy (DMP)* である [2]．蒸留は元々モデル圧縮の技術であり，ラベル付きのデータで学習された（大きな）モデルの出力によりラベル付けされていないデータにソフトラベルを与え，ソフトラベルが付けられたデータで別の（小さな）モデルを訓練する．この技術を利用するため，DMP は学習データに加えてラベル付けされていない大量の公開データを必要とする．したがって，公開データの入手が保証されないセンシティブな情報を扱う医療や金融などの領域では適用が難しい．

本稿では，学習データの分割とデータの使い回しをすることで，公開データを必要としない “交差蒸留” を定義し，それを利用した新たな防御手法を提案する．また，本手法は精度とプライバシー保護のトレードオフに関して，公開データを必要としない既存手法と比較して非常に良く，公開データを必要とする DMP に対しては同程度の性能を示すことを実験的に明らかにする．

参考文献

[1] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov, “Membership inference attacks against machine learning models,” Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP)，pages 3-18, 2017

[2] Virat Shejwalkar and Amir Houmansadr, “Membership privacy for machine learning models through knowledge transfer,” Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)，2021

* シンガポール国立大学, National University of Singapore
† NEC セキュアシステム研究所, NEC Secure System Research Laboratories