УДК 004.056, 004.8

М. В. Онай, к.т.н., доцент, А. І. Северін
(Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Україна)

КОМПЛЕКСНИЙ ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ ЗБЕРЕЖЕННЯ ПРИВАТНОСТІ В МАШИННОМУ НАВЧАННІ

M. V. Onai, PhD, Assoc. Prof., A. I. Severin

COMPREHENSIVE COMPARATIVE ANALYSIS OF PRIVACY-PRESERVING METHODS IN MACHINE LEARNING

З точки зору захисту приватності наборів даних, ключовими загрозами є атаки на логічний висновок [1]. Наприклад, атаки логічного висновку (inference attack) дозволяють зловмиснику зробити висновок про використання конкретного профілю пацієнта для навчання класифікатора, пов'язаного із захворюванням. Іншим прикладом є атаки на інверсію моделі (model inversion attacks), які можуть використовувати доступу «чорної скриньки» до моделей передбачення для оцінки аспектів геномної інформації особи. Також, глибокий витік із градієнтів (deep leakage from gradients) може виводити приватні дані зі спільних градієнтів, що виникають при використанні машинного навчання у завданнях комп’ютерного зору та обробки природної мови.

Основними способами забезпечення захисту приватних наборів даних є [2-11]: генерація синтетичних наборів даних, обробка приватних наборів даних (анонімізація даних, диференційна приватність, гомоморфне шифрування), федеративне навчання. Генерація синтетичних наборів даних (synthetic data generation), що полягає в генерації штучних даних за певним алгоритмом (наприклад, прихована модель Маркова або генеративні конкуруючі нейронні мережі) з наміром перенести результати навчання на реальні дані. Анонімізація даних – це процес захисту приватної інформації шляхом видалення або зміни ідентифікаторів (наприклад, придушення атрибутів, перестановка даних, підміна даних, узагальнення, дисперсія чисел та дат), які з'єднують особу із збереженими даними [10, 11]. Диференційна приватність – метод захисту даних, який захищає конфіденційність користувача шляхом додавання випадкового шуму до даних. Його метою є забезпечення жорстких статистичних гарантій того, що зловмисник не зможе зробити висновок про приватні дані, на основі результатів даних, що отримані за допомогою рандомізованого алгоритму. Гомоморфне шифрування – це форма шифрування, яка дозволяє виконувати обчислення над зашифрованим текстом, розшифрований результат яких буде таким самим, як і результат операцій над відкритим текстом [5]. Федеративне навчання (federated learning) [6, 12] – децентралізований архітектурний підхід, ідея якого полягає в навчанні алгоритму штучного інтелекту на різних кінцевих пристроях або серверах, які містять локальні набори даних. Ці дані залишаються на пристрої під час навчання, тобто вони не обмінюються між пристроями. Такий підхід відрізняється від традиційних централізованих методів машинного навчання, коли всі зразки даних завантажуються на один сервер, а також від більш класичних децентралізованих підходів, які припускають, що локальні зразки даних рівномірно розподіляються між пристроями.

Порівняльний аналіз методів захисту приватних наборів даних було проведено використовуючи наступні п’ять критеріїв: складність, практичність, потреба у великій кількості даних для використання методу, надійність, точність системи штучного інтелекту (на модифікованих даних). На основі проведеного аналізу можна зробити висновок, що генерація синтетичних наборів даних є практичним і надійним методом, але досить складним і вимагає великої кількості вхідних даних для формування більш

точної системи штучного інтелекту. Анонімізація даних досить проста, практична і не потребує великих масивів даних, але цей метод недостатньо надійний. Диференціальна конфіденційність – це практичний метод, який вимагає великих наборів даних, і залежно від кількості використовуваного шуму ефективність захисту може варіюватися від дуже надійного, але неточного в оцінці результатів, до ненадійного, але дуже точного. Гомоморфне шифрування є надійним і може бути використане для побудови високоточних систем, але цей метод є обчислювально витратним, і може застосовуватися до обмеженого класу завдань. Федеративне навчання є надійним і точним методом без розповсюдження локальних даних навчання, але його передумовою є наявність принаймні декількох незалежних користувачів, що мають достатньо даних для навчання.

Розглянуто основні типи атак на системи машинного навчання, а також проаналізовано методи протидії атакам (їх переваги та недоліки), що загрожують витоку приватних даних. Зокрема, були розглянуті методи генерування синтетичних даних, анонімізацію даних, диференційну приватність, гомоморфне шифрування та федеративне навчання. Актуальними напрямками подальших досліджень є розроблення альтернативних і модифікація існуючих методів захисту приватних наборів даних, які дозволять мінімізувати розглянуті недоліки.

Література

1. Xu R. Privacy-preserving machine learning: Methods, challenges and directions / R. Xu, N. Baracaldo, J. Joshi. // arXiv preprint arXiv:2108.04417. — 2021 — DOI: 10.48550/arXiv.2108.04417.
2. Lauter K. Faculty Summit 2017: Private AI [Electronic resource] / Kristin Lauter // Microsoft Research. — 2017. — Access mode: https://www.microsoft.com/en-us/research/wp-content/uploads/2017/07/Private_AI_Kristin_Lauter.pdf.
3. Nikolenko S. I. Synthetic Data for Deep Learning [Electronic resource] / Sergey I. Nikolenko. — 2019. — Access mode: https://arxiv.org/pdf/1909.11512.pdf.
4. Dwork C. The Algorithmic Foundations of Differential Privacy [Text] / C. Dwork, A. Roth. // Foundations and Trends® in Theoretical Computer Science. — 2014. — Vol. 9, №3-4. — С. 211–407. — DOI 10.1561/0400000042.
5. Minelli M. Fully homomorphic encryption for machine learning [Text] / Michele Minelli., 2018. — 157 p.
6. Communication-efficient learning of deep networks from decentralized data. [Text] / [H. Brendan McMahan, E. Moore, D. Ramage and others]. — 2016.
7. Konečný J. Federated Optimization: Distributed Optimization Beyond the Datacenter [Electronic resource] / J. Konečný, B. McMahan, D. Ramage. — 2015. — Access mode: https://arxiv.org/pdf/1511.03575.pdf.
8. Abadi M. Learning to Protect Communications with Adversarial Neural Cryptography [Electronic resource] / M. Abadi, D. G. Andersen. — 2016. — Access mode: https://arxiv.org/abs/1610.06918.
9. Lindell Y. Secure Multiparty Computation (MPC) [Electronic resource] / Yehuda Lindell — Access mode: https://eprint.iacr.org/2020/300.pdf.
10. Data Anonymization Techniques [Electronic resource]. — 2019. — Access mode: https://www.solarwindsmsp.com/blog/data-anonymization-overview.
11. Guide to basic data anonymisation techniques [Electronic resource] // Personal Data Protection Commission Singapore (PDPC). — 2018. — Access mode: https://iapp.org/media/pdf/resource_center/Guide_to_Anonymisation.pdf.
12. Brendan McMahan H. Federated Learning: Collaborative Machine Learning without Centralized Training Data [Electronic resource] / H. Brendan McMahan, D. Ramage. — 2017. — Access mode: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html.