УДК 004.89

КОМБИНИРОВАННЫЕ АТАКИ НА ПРИВАТНОСТЬ: ВЗАИМОДЕЙСТВИЕ MEMBERSHIP И ATTRIBUTE INFERENCE

Танака Л., студент направления MSc Cybersecurity, II курс
Королевский технологический институт, г. Стокгольм, Швеция

Комбинированные атаки на приватность в системах машинного обучения представляют собой одну из наиболее серьёзных угроз для безопасности персональных данных в эпоху повсеместного распространения предобученных и адаптивных моделей. Особенно опасными являются сценарии, в которых различные виды атак – в частности, Membership Inference Attacks (MIA) и Attribute Inference Attacks (AIA) – используются в связке, усиливая друг друга. Такое взаимодействие приводит к возникновению синергетического эффекта, при котором совокупный ущерб для конфиденциальности данных значительно превышает риск, присущий каждой из атак по отдельности [1].

Анализ этих комбинированных стратегий становится критически важным как для разработки защищённых моделей, так и для выработки нормативных и этических стандартов использования ИИ в чувствительных доменах.

Атака Membership Inference направлена на определение того, принадлежал ли конкретный образец входных данных обучающей выборке модели. Она эксплуатирует различия в поведении модели на обученных и незнакомых примерах – например, различия в уровнях уверенности, частоте ошибок или распределении логитов. Несмотря на то, что такие атаки могут быть успешны и по отдельности, они часто дают только бинарную информацию (был/не был), не раскрывая содержание или структуру самих данных. Однако даже этого достаточно, чтобы нарушить конфиденциальность, особенно в сценариях, где само участие в обучении модели является чувствительной информацией (например, в медицинских, юридических или финансовых задачах).

Атака Attribute Inference преследует другую цель: восстановление скрытых признаков (атрибутов) объекта, которые не присутствуют в открытом виде во входных данных, но оказывают влияние на выход модели. Классическим примером является ситуация, в которой модель получает в качестве входа неполный профиль пользователя и выдает результат, по которому можно восстановить, например, его пол, возраст, диагноз или политические предпочтения. Такая атака обычно требует знания некоторых статистических закономерностей в распределении признаков, а также возможности наблюдать или моделировать поведение модели на ограниченном множестве входов [2].

Когда эти два типа атак объединяются, возникает ситуация, в которой знание факта принадлежности объекта обучающей выборке (через MIA) может существенно повысить эффективность восстановления его скрытых признаков (через AIA). Например, если атакующий с высокой вероятностью установил, что конкретный пользователь входил в обучающую выборку, это позволяет ему применить к данному примеру более агрессивные методы атрибутивного восстановления, полагаясь на то, что модель "запомнила" поведение именно этого образца. В свою очередь, успешная атака на атрибут может быть использована для повышения точности membership inference – особенно если атрибут является уникальным или редко встречающимся.

Комбинированные атаки также позволяют использовать боковую информацию, собранную из внешних источников. Например, если атакующий располагает неполными метаданными о пользователе (например, время активности, город, устройство), он может сначала использовать эту информацию для запуска AIA, восстановив один-два скрытых признака, и затем на её основе выполнить MIA, более точно оценивая вероятность включения в обучающую выборку. Это делает такие атаки особенно опасными в сценариях с частичным раскрытием информации, типичным для онлайн-платформ и API-сервисов.

Технически комбинированные атаки могут строиться как в пошаговой модели (сначала MIA, затем AIA или наоборот), так и в совместной – через построение многоцелевых атакующих моделей (multi-objective inference models), которые обучаются одновременно на задачу идентификации участия и задачу восстановления признаков. Такие модели показывают высокую эффективность на реальных данных, особенно при наличии доступа к ограниченному числу запросов к модели или возможности повторного взаимодействия (например, через атакующие запросы с вариациями входов) [3].

Особенно уязвимыми к таким атакам оказываются системы, использующие тонкую настройку (fine-tuning) на пользовательских данных без должного разделения и обфускации параметров. Кроме того, в системах федеративного обучения, где модель обучается по частям на разных устройствах, но агрегирует результаты централизованно, возникает возможность объединения информации из нескольких клиентов, повышающая вероятность успешной атаки.

Методы защиты от комбинированных атак на приватность должны учитывать их многошаговый и стратегический характер. Одной из возможных мер является внедрение дифференциальной приватности на уровне градиентов или логитов, что снижает влияние отдельных образцов на поведение модели и затрудняет извлечение информации. Другой подход – рандомизация выходов и ограничение количества запросов, при которых точность обеих атак резко снижается. Также применяются методы обнаружения атакующих шаблонов(например, частого запроса редких классов или аномальных

последовательностей логитов), позволяющие заблаговременно заблокировать или замедлить подозрительную активность [4].

Особое внимание следует уделить оценке остаточной приватности – то есть возможности успешного выполнения комбинированной атаки даже после внедрения базовых защитных мер. Такие оценки требуют построения метрик, способных учитывать не только вероятность утечки отдельного признака или факта участия, но и совместную вероятность раскрытия чувствительной информации в комплексной атаке.

Таким образом, взаимодействие Membership Inference и Attribute Inference создаёт новые вызовы для обеспечения приватности в системах ИИ. Комбинированные атаки поднимают планку угроз с уровня статистического анализа до уровня персонализированного взлома, где атакующий способен не просто нарушить анонимность, но и восстановить чувствительную информацию о конкретных пользователях.

Разработка устойчивых моделей, способных противостоять таким угрозам, требует междисциплинарного подхода, сочетающего методы теории информации, криптографии, анализа данных и инженерии ИИ. Только на этой основе можно сформировать действительно доверенные системы машинного обучения, устойчивые к всё более изощрённым и комбинированным атакам на приватность [5].

Список литературы:

1. Bousquet, O. Stability and Generalization / O. Bousquet, A. Elisseeff // Journal of Machine Learning Research. – 2002. – Vol. 2. – P. 499–526.

2. Xu, A. Information-Theoretic Analysis of Generalization Capability of Learning Algorithms / A. Xu, M. Raginsky // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 2524–2533.

3. Russo, D. Controlling Bias in Adaptive Data Analysis Using Information Theory / D. Russo, J. Zou // Artificial Intelligence and Statistics. – 2016. – Vol. 51. – P. 1232–1240.

4. Bassily, R. Algorithmic Stability for Adaptive Data Analysis / R. Bassily, K. Nissim, U. Stemmer [et al.] // ACM Symposium on Theory of Computing. – New York : ACM, 2016. – P. 1046–1059. – DOI 10.1145/2897518.2897566.

5. Vadhan, S. The Complexity of Differential Privacy / S. Vadhan // Tutorials on the Foundations of Cryptography. – Cham : Springer, 2017. – P. 347–450. – DOI 10.1007/978-3-319-57048-8_7.