УДК 004.89

ОЦЕНКА ВОЗДЕЙСТВИЯ РЕПРЕЗЕНТАТИВНЫХ ДАННЫХ НА ПРИВАТНОСТЬ МОДЕЛИ

ШевякоВ В.Г., младший научный сотрудник
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

В современном развитии технологий искусственного интеллекта и машинного обучения вопрос обеспечения приватности становится одним из наиболее актуальных и сложных. Одним из ключевых аспектов, напрямую влияющих на степень защиты конфиденциальной информации, является состав и качество обучающих данных, а именно их репрезентативность. Репрезентативные данные – это те, которые адекватно и полно отражают характеристики, вариации и распределения реального мира, на котором предназначена работать модель [1].

При этом влияние репрезентативности на приватность модели оказывается двояким и требует глубокого анализа для нахождения баланса между качеством обучения и сохранением конфиденциальности.

С одной стороны, использование высоко репрезентативных данных способствует более точному обучению и улучшению обобщающих свойств модели. Такой набор данных содержит множество разнообразных примеров, позволяющих модели усвоить истинные зависимости и снизить переобучение на отдельные шумовые или случайные элементы. Однако, вместе с повышением полноты и разнообразия данных возрастает и риск того, что модель может запомнить или непреднамеренно раскрыть информацию, относящуюся к отдельным элементам обучающей выборки. В частности, если в данных присутствуют уникальные или редкие случаи, модель способна «запомнить» их особенности, что делает возможным восстановление этих данных через методы атак на приватность, такие как membership inference или extraction attacks. Таким образом, репрезентативность, с одной стороны, улучшает качество модели, а с другой – увеличивает уязвимость к утечкам чувствительной информации [2].

С другой стороны, неполная или нерепрезентативная выборка данных может казаться более безопасной в плане приватности, поскольку в ней отсутствуют уникальные паттерны и редкие случаи, снижающие риск запоминания и восстановления отдельных элементов. Однако такая ситуация приводит к ухудшению качества модели, снижению её точности и способности к

обобщению, а также может привести к появлению предвзятости и несправедливости в предсказаниях. Это особенно критично в системах, используемых в медицине, финансах и других областях, где ошибки имеют серьёзные последствия. Более того, снижение репрезентативности не всегда гарантирует адекватную защиту – даже ограниченный набор данных может содержать чувствительные закономерности, которые модель способна выявить и раскрыть.

Для анализа влияния репрезентативности на приватность в последнее время развиваются методы количественной оценки, основанные на формальных метриках и практических тестах. Например, с помощью атак типа membership inference исследуется вероятность того, что конкретный элемент данных входит в обучающую выборку, при этом учитывается его частотность и уникальность. Кроме того, используются методы моделирования риска восстановления данных и оценки степени «запоминания» модели, которые позволяют оценить, насколько информация о тренировочных примерах может быть извлечена из модели в открытом или полуоткрытом доступе [3].

Важным направлением является также анализ распределения данных и балансировка классов в выборке. Нерепрезентативные данные с перекосами по классам способствуют усилению предвзятости и могут создавать «узкие места», где риск приватностных нарушений особенно высок для недостаточно представленных групп. Репрезентативность, напротив, способствует более равномерному распределению внимания модели, что снижает вероятность дискриминационных и уязвимых зон. При этом сама репрезентативность должна оцениваться не только по количеству, но и по качеству данных, учитывая наличие прокси-признаков и сложных зависимостей.

Методы обеспечения приватности, такие как дифференциальная приватность, тесно связаны с понятием репрезентативности. Применение шумовых механизмов и ограничений влияния отдельных данных на модель требует тонкой настройки параметров, чтобы сохранить баланс между защитой данных и сохранением точности модели. В ряде случаев достижение этого баланса предполагает компромисс – частичное снижение качества обучения ради усиления приватности и наоборот [4].

Практическая реализация оценки воздействия репрезентативности включает в себя построение тестовых сценариев, проведение стресс-тестов и аудитов, а также использование синтетических данных и генеративных моделей для анализа чувствительности модели к вариациям в данных. Особое внимание уделяется выявлению областей повышенного риска, где конфиденциальная информация может быть наиболее уязвима к раскрытию, и применению дополнительных мер защиты или ограничений доступа к таким участкам.

Таким образом, влияние репрезентативных данных на приватность моделей машинного обучения представляет собой сложное многогранное явление,

требующее комплексного подхода с использованием как теоретических, так и эмпирических методов. Балансировка между точностью, обобщающей способностью и уровнем защиты данных является фундаментальной задачей при создании современных доверенных и ответственных ИИ-систем, особенно в условиях ужесточающихся требований к безопасности и этике применения технологий. Только интегрированный подход к анализу и контролю репрезентативности данных способен обеспечить эффективную защиту приватности без ущерба для функциональности и надёжности искусственного интеллекта [5].

Список литературы:

1. Li, N. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity / N. Li, T. Li, S. Venkatasubramanian // IEEE International Conference on Data Engineering. – Piscataway : IEEE, 2007. – P. 106–115. – DOI 10.1109/ICDE.2007.367856.

2. Narayanan, A. Robust De-Anonymization of Large Sparse Datasets / A. Narayanan, V. Shmatikov // IEEE Symposium on Security and Privacy. – Piscataway : IEEE, 2008. – P. 111–125. – DOI 10.1109/SP.2008.33.

3. Ganta, S. R. Composition Attacks and Auxiliary Information in Data Privacy / S. R. Ganta, S. P. Kasiviswanathan, A. Smith // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – New York : ACM, 2008. – P. 265–273. – DOI 10.1145/1401890.1401926.

4. Domingo-Ferrer, J. A Critique of k-Anonymity and Some of Its Enhancements / J. Domingo-Ferrer, V. Torra // International Conference on Availability, Reliability and Security. – Piscataway : IEEE, 2008. – P. 990–993. – DOI 10.1109/ARES.2008.97.

5. Ohm, P. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization / P. Ohm // UCLA Law Review. – 2010. – Vol. 57. – P. 1701–1777.