УДК 004.942

ГЕНЕРАЦИЯ СИНТЕТИЧЕСКИХ ТРАНЗАКЦИЙ В УСЛОВИЯХ ОГРАНИЧЕНИЙ КОНФИДЕНЦИАЛЬНОСТИ

Захаров К.А. (Университет ИТМО), **Ставинова Е.А.** (Университет ИТМО)

Научный руководитель – к.т.н., доцент Гулева В.Ю.
(Университет ИТМО)

Введение. Транзакционные данные являются одним из главных источников информации в финансовой сфере. Транзакции показывают активность клиентов банка или финансовой организации в виде объёма трат по некоторым категориям в определенные моменты времени. Таким образом, транзакции представляют собой табличные данные, но с временной зависимостью. Помимо перечисленных атрибутов данные могут также содержать информацию о возрасте клиента, месте совершения транзакции и т. д. Сегодня исследователи в области искусственного интеллекта активно создают новые методы для генерации синтетических данных. Однако мало исследований посвящено работе именно с транзакционными данными. Помимо генерации данных, транзакции также должны обладать должным уровнем конфиденциальности или дифференциальной приватности для сохранности шаблонов поведения клиента банка.

Требования конфиденциальности исключают возможность выявления нарушителем закономерностей персональных данных клиента по синтетическим данным, например, объёма трат в определенной категории товаров. Для обеспечения механизмов приватности в модель добавлен алгоритм зашумления функции потерь так, чтобы модель была способна порождать синтетику, схожую с данными реального мира, и при этом бы удовлетворяла метрикам приватности.

Основная часть. На первом этапе исследования был проведен аналитический обзор научных источников по теме генерации синтетических транзакций, табличных данных и временных рядов. Были выделены основные недостатки существующих подходов и, в особенности, проблемы существующих методов генерации транзакций [2]. Рассматривая методы генерации табличных данных, акцент был сделан в сторону возможности порождать многомерные данные, но без фактора времени. При обзоре методов временных рядов внимание уделялось возможности генерации упорядоченных во времени последовательностей, в том числе многомерных временных рядов [1, 4].

Вторая часть исследования направлена на разработку метода генерации синтетических транзакций. Предложенный метод включает новый механизм предобработки данных, новую схему генерации транзакций, а также механизмы конфиденциальности. Все атрибуты в транзакционных данных были разделены на 4 типа: категориальные признаки с большим числом уникальных значений, категориальные признаки с маленьким числом уникальных значений, числовые признаки и признак, отвечающий за фактор времени. После разделения атрибутов метод пропускает их через дифференциально приватные автокодировщики для получения сжатого представления и увеличения степени конфиденциальности полученных представлений. Далее также выделяется условный вектор, который включает в себя категории трат и время. После получения компактных представлений и условного вектора данные подаются на вход генеративной модели, которая была представлена в статье [5], но с модификацией в виде добавления условного вектора и подачи на вход случайного процесса. Также метод генерации модифицирован для создания дифференциально приватных данных путем использования дифференциальной приватности Реньи [3]. Для этого были доказаны теоремы для оценки чувствительности функций потерь, а также теоремы об общем уровне получаемой моделью приватности.

Финальная часть исследования включает разработку генератора синтетического времени и нового механизма дифференциальной приватности, который способен учитывать фактор времени.

Выводы. Работа предложенного метода продемонстрирована с помощью экспериментального сравнения с методами CTGAN, CopulaGAN и Banksformer на транзакционных данных двух банков, находящихся в открытом доступе. Результаты показали, что разработанный метод демонстрирует более высокое качество генерации транзакций по ряду критериев. Основными критериями выступали дивергенция Йенсена-Шеннона, расстояние Вассерштейна и статистические тесты. Также были представлены гистограммы распределений объемов трат, категорий транзакций, приращений времени и потока денежных средств в сравнении с данными реального мира. Также метод позволяет генерировать синтетическое время, что выделяет его из существующих подходов. Отдельно было проведено экспериментальное исследование по сохранению конфиденциальности данных путем проверки различий в распределениях, метриках задач машинного обучения и метриках приватности, таких как t-closeness.

Разработанный метод можно использовать в случае недостатка данных для решения задач финансовой сферы, например рекомендаций категорий кэшбека или прогнозирования трат клиентов в определенной категории или момент времени.

Список использованных источников:

1. Alaa, A., Chan, A.J., van der Schaar, M.: Generative time-series modeling with fourier flows. In: International Conference on Learning Representations (2020).
2. Chakraborti, A., Toke, I.M., Patriarca, M., Abergel, F.: Econophysics review: I. empirical facts. Quantitative Finance 11(7), 991–1012 (2011).
3. Mironov, Ilya. "Rényi differential privacy." In 2017 IEEE 30th computer security foundations symposium (CSF), pp. 263-275. IEEE, 2017.
4. Wiese, M., Knobloch, R., Korn, R., Kretschmer, P.: Quant gans: deep generation of financial time series. Quantitative Finance 20(9), 1419–1440 (2020).
5. Zakharov K., Stavinova E., and Boukhanovsky A., "Synthetic Financial Time Series Generation with Regime Clustering," Journal of Advances in Information Technology, Vol. 14, No. 6, pp. 1372-1381, 2023.

Захаров К.А. (автор) Подпись

Гулева В.Ю. (научный руководитель) Подпись