УДК 004.048

СРАВНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПОСТРОЕНИЯ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ ПОЛЬЗОВАТЕЛЕЙ НА ОСНОВЕ ГЕОГРАФИЧЕСКОЙ АКТИВНОСТИ ИЗ ТРАНЗАКЦИОННЫХ ДАННЫХ

Глухов Г.И. (Университет ИТМО)
Научный руководитель – к.т.н. Шиков Е.Н
(Университет ИТМО)

Введение. Построение векторных представлений пользователей в настоящее время используется для решения различных задач, например утверждения кредита [1], сегментации клиентов банковских услуг [2], прогнозирования жизненного цикла клиента [3], и других бизнес-задач. Одной из проблем в сфере финансовых технологий является интерпретируемость векторных представлений пользователей [4]. Эта задача имеет большое значение из-за необходимости для пользователя модели (т. е. заинтересованного лица в бизнесе) объяснять результаты, предоставляемые моделью, и использовать их для поддержки принятия решений, направленных на достижение бизнес-целей. С другой стороны, векторные представления пользователей должны повышать качество моделей прогнозирования за счет предоставления явной информации о поведении клиента. В данной работе мы предлагаем подход на основе кластеризации для построения интерпретируемых векторных представлений пользователей на основе географических характеристик, извлеченных из их истории транзакций.

Основная часть. Предложенный метод можно разбить на несколько этапов. На первом этапе рассчитываются пользовательские вектора признаков, а именно вектора географической активности клиентов. Для этого данные группируются по пользователям и географическим характеристикам с использованием различных функций агрегирования. В данной работе в качестве географических характеристик рассматриваются районы города и муниципальные округа, а в качестве функций агрегации сумма, количество, а также минимальный и максимальный размер транзакции. После чего каждый пользователь имеет набор векторов признаков для каждой пары характеристики и агрегирующей функции. Эти вектора нормализуются при помощи L2-нормы. На втором этапе полученные вектора географической активности клиентов кластеризуются (метод кластеризации может быть любым). Для каждого полученного кластера вычисляется его центр. Центр кластера определяется как среднее значение всех векторов географической активности пользователей, которые входят в этот кластер. Количество компонент (кластеров) определяется при помощи метода локтя. На последнем этапе итоговые низкоразмерные векторные представления пользователей рассчитываются как вектор (размер которого равен количеству кластеров, полученном на предыдущем этапе), где каждое значение показывает расстояние между вектором географической активности пользователя и центром кластера. В качестве меры расстояния в данной статьей рассматривается евклидово и косинусное расстояние между векторами.

Выводы. Мы сравниваем результаты, полученные из данных, сгенерированных с помощью предложенного метода, с данными, сгенерированными с использованием метода word2vec [5] и методов, основанных на моделях автокодировщиков [6]. Результаты показывают, что наш метод обеспечивает сравнимую (в некоторых случаях превосходящую) эффективность при сохранении интерпретируемости признаков.

Список использованных источников:

1. X. Song, M. T. Liu, Q. Liu, and B. Niu, ‘Hydrological cycling optimization-based multiobjective feature-selection method for customer segmentation’, Int. J. Intell. Syst., vol. 36, no. 5, pp. 2347–2366, May 2021, doi: 10.1002/int.22381.

2. C. Calvo-Porral and J. P. Lévy-Mangin, ‘An emotion-based segmentation of bank service customers’, Int. J. Bank Mark., vol. 38, no. 7, pp. 1441–1463, 2020, doi: 10.1108/IJBM-05-2020-0285.

3. J. Bauer and D. Jannach, ‘Improved Customer Lifetime Value Prediction With Sequence-To-Sequence Learning and Feature-Based Models’, ACM Trans. Knowl. Discov. from Data, vol. 15, no. 5, pp. 1–37, 2021.

4. D. Smilkov, N. Thorat, C. Nicholson, E. Reif, F. B. Viégas, and M. Wattenberg, ‘Embedding projector: Interactive visualization and interpretation of embeddings’, arXiv Prepr. arXiv1611.05469, 2016.

5. L. Baldassini and J. A. R. Serrano, ‘client2vec: towards systematic baselines for banking applications’, arXiv Prepr. arXiv1802.04198, 2018.

6. Y. Bengio, A. Courville, and P. Vincent, ‘Representation learning: A review and new perspectives’, IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, 2013.

Глухов Г.И. (автор) Подпись

Шиков Е.Н. (научный руководитель) Подпись