Эффективное агрегирование по меткам для задачи последовательностей событий

Галина Леонидовна Боева
Научный руководитель: к.ф.-м.н. А. А. Зайцев

Кафедра интеллектуальных систем ФПМИ МФТИ
Специализация: Интеллектуальный анализ данных
Направление: 03.04.01 Прикладные математика и физика

2024

Агрегирование по меткам для задачи последовательностей событий

Проблема

Современные подходы фокусируются на архитектуре преобразования последовательных данных, агрегируя данные по временным меткам, но теряя информацию о взаимозависимостях меток.

Цель работы

Создание подхода, основанного на механизме собственного внимания над метками, предшествующими прогнозируемому шагу.

Задачи работы

1) разработка метода на основе внимания для предсказания множества меток
2) валидация разработанных методов
3) обоснование причинно-следственных связей с помощью построения графа на основе внимания

Постановка задачи предсказания временных наборов

Пусть $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ - это набор из N элементов.
Каждый элемент $u_i, 1 \leq i \leq N$, связан с последовательностью временных множеств $\mathcal{S}_i = \{s_i^1, s_i^2, \ldots, s_i^T\}$, где T - число наблюдаемых временные метки.
Набор $s_i^j, 1 \leq i \leq N, 1 \leq j \leq T$, представляет собой набор произвольного количества меток, выбранных из словаря $\mathcal{Y} = \{y_1, y_2, \ldots, y_L\}$ размера L.
Цель задачи предсказания временных множеств состоит в том, чтобы предсказать последующий набор меток $\hat{s}_i^{T+1}$, то есть,

$$\hat{s}_i^{T+1} = g(s_i^1, s_i^2, , s_i^T, \mathbf{W}), \tag{1}$$

где $\mathbf{W}$ относится к обучаемым параметрам функции g.

Предложенный метод на основе внимания на метках

Пусть $\mathbf{X} \in \mathbb{R}^{L \times D}$ — матрица представлений всех меток из словаря $\mathcal{Y} = \{y_1, y_2, \ldots, y_L\}$. Для каждой временной метки $j, 1 \leq j \leq T$ создается временное представление $\mathbf{t}_j \in \mathbb{R}^D$, как это сделано в [1]. Для каждого момента времени $t_j, 1 \leq j \leq T$ образуется матрица представлений $\mathbf{Z} \in \mathbb{R}^{L \times D}$. l-я строка, $1 \leq l \leq L$, матрицы $\mathbf{Z}$, обозначаемая как $\mathbf{Z}^{(l,:)}$, равна сумме представлений временных меток, в которых метка $y_l \in \mathcal{Y}$ отображается как элемент набора:

$$\mathbf{Z}^{(l,:)} = \sum_{j | y_l \in s_i^j} \mathbf{t}_j. \tag{2}$$

Тогда:

$$\mathbf{G} = \mathbf{X} \oplus \mathbf{Z}. \tag{3}$$

Для выявления зависимостей меток $\tilde{\mathbf{G}}$:

$$\tilde{\mathbf{G}} = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\mathbf{T}}{\sqrt{2D}})\mathbf{V}. \tag{4}$$

Агрегирование по меткам

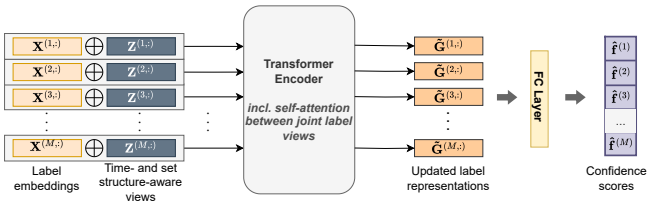
Теорема: Сложность алгоритма, основанного на агрегации на основе меток имеет меньшую вычислительную сложность, чем алгоритм на основе агрегирования по временным меткам для задачи последовательностей событий.

Этапы	Агрегирование по меткам	Агрегирование по времени
Create Emb	$O(T \times M)$	$O(T \times M)$
Transformer	$O(M^2)$	$O(T^2)$
Updated Emb	$O(M^2)$	$O(T^2)$
Affine transform	$O(M)$	$O(T)$
Общая сложность	$O(T \times M + M^2)$	$O(T \times M + T^2)$

Так как $M << T$, то сложность по меткам меньше, чем сложность по времени, получается агрегация по меткам эффективнее.

Предложенный метод на основе внимания на метках

Гипотеза: Рассматривая два типа агрегирования по меткам и по времени для процесса Хокса, можно утверждать, что функция интенисивности, агрегированная по меткам ближе к истинной функции интенсивности заданного процесса.



Общий пайплайн получения глобальных представлений

Вычислительный эксперимент: Данные

Статистика наборов данных для прогнозирования временных наборов.

Dataset	#Sets	MdnSS	MaxSS	Vocab	MnLen	#Seqs
Mimic III	17 849	5	23	169	2.7	6636
Instacart	115 604	6	43	134	16.5	7000

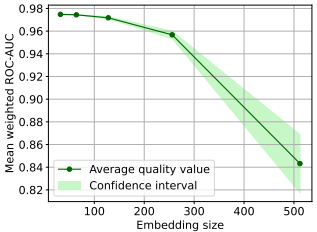
- **Mimic III** — датасет, состоящий из медицинских карт пациентов из отделения интенсивной терапии. Событие, связанное с пациентом, включает в себя время поступления в больницу и набор классификационных кодов заболеваний.
- **Instacart** — набор данных содержит записи о заказах товаров пользователями. Товары из маркетплейсов и магазинов.

Вычислительный эксперимент: Основные результаты

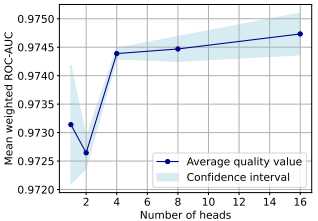
Сравнение подхода our LANET с существующими моделями для прогнозирования временных наборов на основе четырех наборов данных. Выделены наилучшие значения, а вторые по значению подчеркнуты.

Data	Model	Weighted F1↑	Weighted ROC-AUC↑	Hamming Loss↓
Mim	SFCNTSP	0.3791 ± 0.0081	0.7034 ± 0.0024	0.0377 ± 0.0004
	DNNTSP	0.3928 ± 0.0030	0.6926 ± 0.0003	0.0365 ± 0.0003
	GPTopFreq	0.4291 ± 0.0073	0.6912 ± 0.0028	0.0398 ± 0.0005
	TCMBN	0.4979 ± 0.0180	0.8670 ± 0.0095	0.0305 ± 0.0008
	LANET(ours)	**0.8214 ± 0.0224**	**0.9852 ± 0.0023**	**0.0220 ± 0.0001**
Ins	SFCNTSP	0.1672 ± 0.0112	0.6852 ± 0.0448	0.0581 ± 0.0004
	DNNTSP	0.4160 ± 0.0009	0.7913 ± 0.0004	0.0541 ± 0.0002
	GPTopFreq	0.4087 ± 0.0079	0.7736 ± 0.0039	0.0529 ± 0.0008
	TCMBN	0.3687 ± 0.0065	0.8187 ± 0.0030	0.0530 ± 0.0005
	LANET(ours)	**0.6159 ± 0.0029**	**0.9445 ± 0.0008**	**0.0474 ± 0.0003**

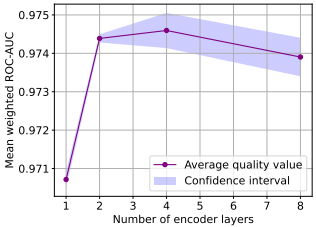
Вычислительный эксперимент: Дополнительные исследования



Зависимость качества LANET от размера векторных представлений.

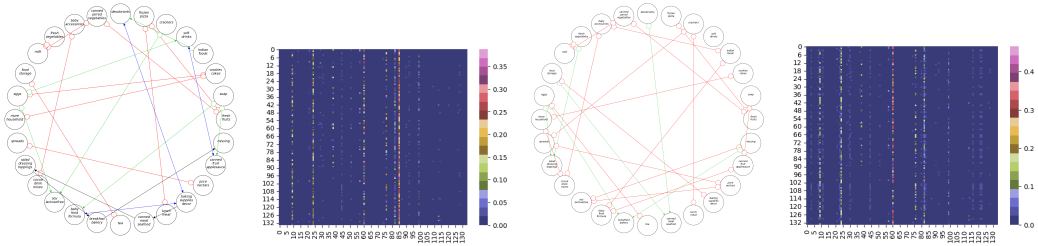


Зависимость качества LANET от количества голов во внимании.



Зависимость качества LANET от количества слоев энкодера.

Графовая интерпретация внимания на метках



Интерпретация взаимосвязи [2] надписей с помощью слоя attention и последующим удалением метки с наибольшим весом внимания из всех возможных значений.

Результаты, которые выносятся на защиту

- Проведены исследования по анализу различных наборов данных, используемых при сравнении реализованной модели LANET.
- Проведены ряд экспериментов для задачи классификации с несколькими метками на двух различных выборках и сравнение с базовыми подходами в данной области.
- Проведен анализ причинно-следственных связей в self-attention, где используется графовый подход на основе построения PAG для взаимосвязи меток.
- Проведена оценка метрики в зависимости от гиперпараметра, отвечающего за размер входных представлений, количество голов во внимании и также количества слоев энкодера.

Список работ автора по теме диплома

Статья **опубликована** в октябре 2024 года **на конференцию ранга A ECAI**.

1. Elizaveta Kovtun, Galina Boeva, Andrey Shulga, and Alexey Zaytsev. Label Attention Network for Temporal Sets Prediction: You Were Looking at a Wrong Self-Attention, IOS Press, October 2024.
2. *Zhuzhel, V., Grabar, V., Boeva, G., Zabolotnyi, A., Stepikin, A., Zholobov, V., Ivanova, M., Orlov, M., Kireev, I., Burnaev, E., Rivera-Castro, R., Zaytsev, A.:Continuous-time convolutions model of event sequences (2023)*

Вклад: разработка идеи статьи, базовые подходы, исследование устойчивости модели и графовая интерпретация внимания.

Благодарность

Алексей Зайцев
Елизавета Ковтун
Андрей Шульга
Владислав Жужель
Александр Степикин
Всеволод Грабарь
Артем Заболотный
Владимир Жолобов

Xiao Shou, Tian Gao, Shankar Subramaniam, Debarun Bhattacharjya, and Kristin Bennett.
Concurrent multi-label prediction in event streams.
In *AAAI Conference on Artificial Intelligence*, 2023.

Raanan Y Rohekar, Yaniv Gurwicz, and Shami Nisimov.
Causal interpretation of self-attention in pre-trained transformers.
Advances in Neural Information Processing Systems, 36, 2024.