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

2025

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 агрегирования по меткам

Определение

Пусть $\mathcal{Y} = \{y_1, ..., y_M\}$, $|\mathcal{Y}| = M$. Последовательность множеств: $\mathcal{S} = (s_1, ..., s_T)$, $s_t \subseteq \mathcal{Y}$. Обучаемые эмбеддинги: $\mathbf{x}_m \in \mathbb{R}^D$, $X = (\mathbf{x}_1, ..., \mathbf{x}_M)^\top \in \mathbb{R}^{M \times D}$.

Агрегация по меткам: $N_m = \sum_{t=1}^{T} \mathbf{1}(y_m \in s_t), \quad \mathbf{Z}_{\text{label}}(m, :) = N_m \cdot \mathbf{x}_m$

Агрегация по времени: $\mathbf{Z}_{\text{time}}(m, :) = \sum_{t:\, y_m \in s_t} \mathbf{t}_t$

Теорема 1 им. Боевой (об эффективности агрегирования по меткам)

Если $M \ll T$, то:

$$T_{\text{label}} = O(M^2 D) \ll T_{\text{time}} = O(T^2 D)$$

Эффективное агрегирование по меткам

Лемма: Пусть $G \in \mathbb{R}^{N \times D}$ — матрица входных представлений, где N — количество объектов (меток или временных событий), а D — размерность эмбеддинга. Тогда время выполнения одного трансформерного слоя внимания:

$$T_{\text{attn}} = O(N^2 D).$$

- для агрегирования по меткам: $T_{\text{label}} = O(M^2 D)$
- для агрегирования по времени: $T_{\text{time}} = O(T^2 D)$

Сравнение вычислительной сложности:

Рассмотрим отношение времени работы моделей:

$$\frac{T_{\text{label}}}{T_{\text{time}}} = \frac{O(M^2 D)}{O(T^2 D)} = O\left(\frac{M^2}{T^2}\right)$$

Если выполняется условие $M \ll T$, то есть $\lim_{T \to \infty} \frac{M}{T} = 0$, то:

$$\frac{T_{\text{label}}}{T_{\text{time}}} \to 0 \quad \Rightarrow \quad T_{\text{label}} \ll T_{\text{time}}$$

Агрегирование по меткам в процессах Хокса

Пусть $(\Omega, \mathcal{F}, P)$ — вероятностное пространство. Рассмотрим многомерный процесс Хокса $\{N_t^m\}_{t\geq 0}$, где $m = 1, ..., M$ обозначает тип события (метку), и интенсивность определяется как:

$$\lambda_t^m = \mu_m + \sum_{n=1}^{M} \int_0^t g_{mn}(t-s)\, dN_s^n,$$

где $\mu_m > 0$ — базовая интенсивность, $g_{mn} : \mathbb{R}_+ \to \mathbb{R}_+$ — ядро возбуждения, $\sup_{m,n} \int_0^\infty g_{mn}(u)\, du < \infty$. Обозначим через $\mathcal{H}_t = \sigma(N_s^m : 0 \leq s \leq t,\ m = 1, ..., M)$ — историю событий до времени t.

Теорема 2 им. Боевой (применение к процессам Хокса)

Если $\phi(m, N_s)$ является достаточной статистикой для λ_t^m, то существует последовательность оценок $\{\hat{\lambda}_t^m\}_{t\geq 0}$, основанных на $\{S_t^m\}_{t\geq 0}$, такая что:

$$\lim_{t\to\infty} \mathbb{E}\left[\ell(\hat{\lambda}_t^m, \lambda_t^m)\right] = 0,$$

и эта сходимость равномерна по $m = 1, ..., M$.

Доказательство (ключевые шаги):

1. Из условия $\rho(\Gamma) < 1$ следует существование стационарного режима.
2. Эмпирическая частота: $\bar{N}_t^m = \frac{1}{t} N_t^m \xrightarrow{a.s.} \nu_m$.
3. При выборе $\phi(m, N_s) = f\left(\frac{d}{ds} N_s^m\right)$ получаем:

$$S_t^m = \frac{N_t^m}{t} f(1) \to \nu_m f(1).$$

4. Если $\hat{\lambda}_t^m = h(S_t^m)$, и $h(\nu_m f(1)) = \nu_m$, то:

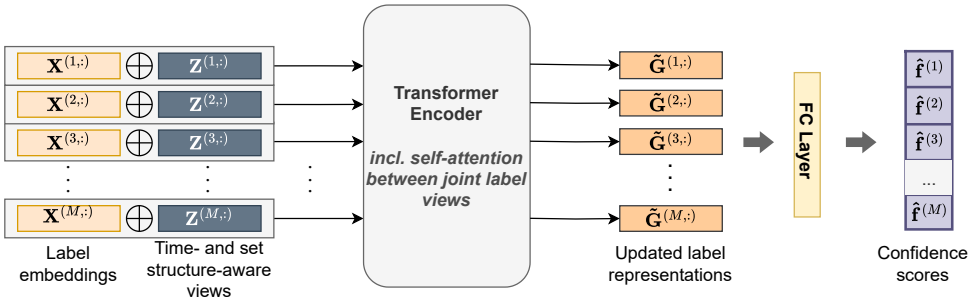
$$\hat{\lambda}_t^m \to \mathbb{E}[\lambda_t^m].$$

5. Для $\ell(x, y) = \|x - y\|^2$:

$$\mathbb{E}[\|\hat{\lambda}_t^m - \lambda_t^m\|^2] \to 0,$$

и сходимость равномерна по m.

Предложенный метод на основе внимания на метках



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

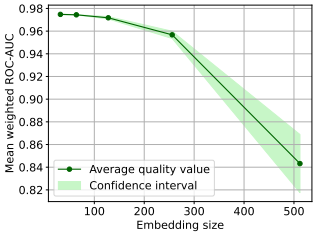
- **Mimic III** — датасет, состоящий из медицинских карт пациентов из отделения интенсивной терапии. Событие, связанное с пациентом, включает в себя время поступления в больницу и набор классификационных кодов заболеваний.
- **Instacart** — набор данных содержит записи о заказах товаров пользователями.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

Вычислительный эксперимент: Основные результаты

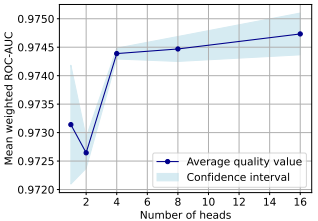
Сравнение подхода our LANET с существующими моделями для прогнозирования временных наборов на основе четырех наборов данных. Выделены наилучшие значения, а вторые по значению подчеркнуты.

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**

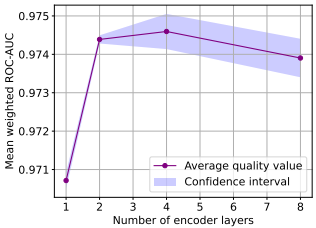
Вычислительный эксперимент: Дополнительные исследования



Зависимость качества LANET от размера векторных представлений.

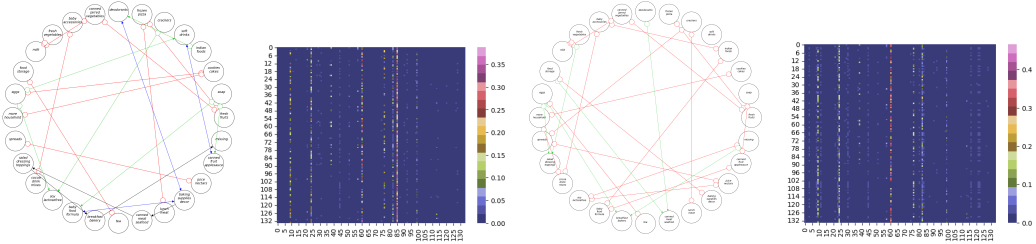


Зависимость качества LANET от количества голов во внимании.



Зависимость качества LANET от количества слоев энкодера.

Графовая интерпретация внимания на метках



Интерпретация взаимосвязи [1] 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. Vladislav Zhuzhel, Galina Boeva, Vsevolod Grabar, Artem Zabolotnyi, Alexander Stepikin, Vladimir Zholobov, Maria Ivanova, Mikhail Orlov, Ivan Kireev, Evgeny Burnaev, Rodrigo Rivera-Castro, Alexey Zaytsev. Continuous-time convolutions model of event sequences (2023). Статья подана в журнал Experts Systems With Applications.
3. Ilya Kuleshov, Galina Boeva, Vladislav Zhuzhel, Evgeni Vorsin, Evgenia Romanenkova, Alexey Zaytsev. DeNOTS: Stable Deep Neural ODEs for Time Series (2024). Статья подана на NeurIPS 2025.

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

Благодарность

Алексей Зайцев
Елизавета Ковтун
Владислав Жужель
Илья Кулешов


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.