Addition of external information for enhancement of local embeddings for event sequences data models

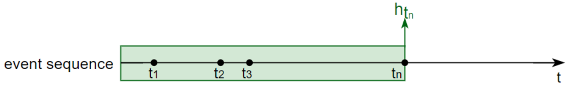
Maria Kovaleva
Research advisor: Ph.D. Alexey Zaytsev

Department of Intelligent Systems
Moscow Institute of Physics and Technology
&
Data Science
Skolkovo Institute of Science and Technology

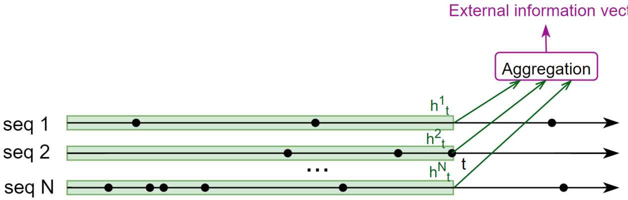
2024

Introduction. External information

General problem: building of embeddings $\mathbf{h}_t$ for event sequences



Gaps: self-supervised models for representation learning ignore external information



Idea: the external information is contained in sequences themselves and can be represented as their aggregation

Aim and Objectives

The aim of the work:

To enhance the embeddings for event sequences data models using aggregation of external information.

Objectives:

1. Development of aggregation methods for external aggregation accounting
2. Validation of developed methods on bank transactions data

Problem statement and baseline approaches

$D = \left\{S^i\right\}_{i=1}^{n}$ — set of n sequences

$S^i = \left\{\left(t_j^i, \mathbf{Z}_j^i\right)\right\}_{j=0}^{T^i}$ — event sequence

$t_j^i \in [0, T^i]$ — time of the event; $\mathbf{Z}_j^i \in \mathbb{R}^d$ — description of the event

Embeddings construction: $e\left(S^i\right) = H^i$

- e is encoder: usually dense NN for event description encoding + recurrent NN
- $H^i = \left\{\left(t_j^i, \mathbf{h}_j^i\right)\right\}$ — embeddings

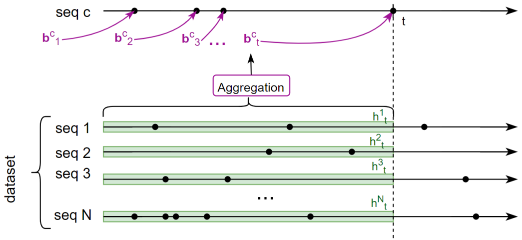
Contrastive learning for encoder:

$L_{km} = I_{k=m} d\left(\mathbf{h}^k, \mathbf{h}^l\right)^2 + \frac{1}{2}\left(1 - I_{k=m}\right) \max\left\{0, \rho - d\left(\mathbf{h}^k, \mathbf{h}^l\right)\right\}^2$,
where d is a distance between embeddings and ρ is a hyperparameter

Autoregressive learning for encoder:
A loss function consists of the cross-entropy for the categorical features and MSE for the continuous features

Receiving representations of external information and taxonomy of aggregation methods

Construction of a vector of external information



Taxonomy of aggregation methods

When constructing aggregation methods, one can use the similarity of the current sequence and sequences from the training set

Similarity of sequences:

1. by embeddings
2. by time

General formula for aggregation: $\mathbf{b}^c_\tau = f\left(H, f_e\left(H, \mathbf{h}^c_{t<\tau}\right), f_t\left(\tau, T\right)\right)$

Where $H = \left[\mathbf{h}^1_{t<\tau}, \ldots, \mathbf{h}^n_{t<\tau}\right]$ is a matrix with embeddings of all sequences at current time,
$T = \left[t^1_{t<\tau}, \ldots, t^n_{t<\tau},\right]$ is a vector of the last event times for all sequences;
f_e and f_t are functions to measure similarity by embeddings and time,
f is aggregation function, usually weighed sum of vectors from H.

Proposed aggregation methods

Classic:

1. **Mean**: $\mathbf{b}_\tau^c = \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_{t<\tau}^i$
2. **Max**: $\mathbf{b}_\tau^c = \max\left(H\right)$

Inspired by the Hawkes process*:

3. **Exp Hawkes**:
$\mathbf{b}_\tau^c = H\exp\left(-\left(\tau\mathbf{1} - T\right)\right)$
4. **Exp learnable Hawkes**: $\mathbf{b}_\tau^c =$
$\phi_{NN}\left(\text{concat}\left(H, \mathbf{h}_\tau^c\right)\right)\exp\left(-\left(\tau\mathbf{1} - T\right)\right)$
5. **Attention Hawkes**: $\mathbf{b}_\tau^c =$
$H\left(\text{softmax}\left(H^T\mathbf{h}_\tau^c\right)\odot\exp\left(-\left(\tau\mathbf{1} - T\right)\right)\right)$

Attention based:

6. **Attention**:
$\mathbf{b}_\tau^c = H\text{softmax}\left(H^T\mathbf{h}_\tau^c\right)$
7. **Learnable attention**:
$\mathbf{b}_\tau^c = H\text{softmax}\left(H^T A\mathbf{h}_\tau^c\right)$
8. **Symmetrical attention**:
$\mathbf{b}_\tau^c = H\text{softmax}\left(H^T S^T S\mathbf{h}_\tau^c\right)$
9. **Kernel attention**:
$\mathbf{b}_\tau^c = H\text{softmax}\left(\phi\left(H^T\right)\phi\left(\mathbf{h}_\tau^c\right)\right)$

Where A and S are matrices with learnable parameters,
ϕ and ϕ_{NN} are learnable transformation (two-layer NN).

*Laub P. J., Taimre T., Pollett P. K. Hawkes processes. arXiv preprint arXiv:1507.02822. – 2015.

Validation of proposed methods

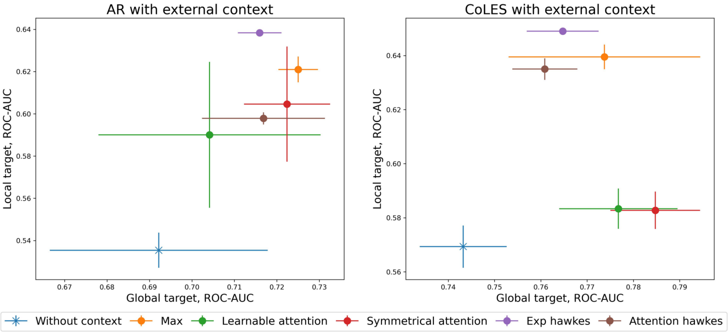
Dataset of the bank transactions:

1. Each sequence S^i: one bank client transactions
2. Each event $\mathbf{Z}^i_j$: transaction (merchant category code + amount)
3. Target: whether the client left the bank

Validation:

1. **Global** : the model inference on the whole sequence to check its global patterns
   a binary downstream task
   b one output vector for a boosting model
2. **Local**: the model inference on sliding windows to check local properties.
   a next event type prediction
   b local target prediction
   c MLP head for prediction

Results: external information addition enhances metrics



The best models are to the right and higher.

1. Accounting of the **external information** improve metrics
2. **Exp Hawkes** method is the best for the local task
3. **Classic** and **attention based** methods are the best for the global task

Defence statements

1. It was suggested to use embeddings aggregations to account for external information in event sequences.
2. A taxonomy of aggregation methods is proposed and specific methods are implemented.
3. It was shown that addition of external information improves the quality of embeddings when used in various applied problems with real data.

Publications

1. Bazarova, A.*, Kovaleva, M.†*, Kuleshov, I.*, Romanenkova, E.*, Stepikin, A.*, Yugay, A.*, Mollaev, D., Kireev, I., Savchenko, A., and Zaytsev, A. Universal representations for financial transactional data: embracing local, global, and external contexts. arXiv preprint arXiv:2404.02047 (2024)

---

*Equal contribution

†Contribution: analysis of the external information addition

Additional slides. Results.

	Global target			
	Contrastive learning		Autoregressive learning	
	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
Without contex	0.743 ± 0.009	0.792 ± 0.014	0.692 ± 0.025	0.734 ± 0.032
Mean	0.773 ± 0.004	0.828 ± 0.003	0.722 ± 0.007	0.776 ± 0.005
Max	0.774 ± 0.021	0.818 ± 0.032	**0.725 ± 0.005**	**0.777 ± 0.002**
Attention	0.760 ± 0.014	0.808 ± 0.017	0.696 ± 0.014	0.744 ± 0.017
Learn. attention	0.777 ± 0.013	0.830 ± 0.013	0.704 ± 0.026	0.751 ± 0.020
Sym. attention	**0.785 ± 0.010**	**0.835 ± 0.005**	0.722 ± 0.010	0.769 ± 0.004
Kernel attention	0.775 ± 0.003	0.824 ± 0.002	0.709 ± 0.019	0.760 ± 0.003
Exp Hawkes	0.765 ± 0.008	0.814 ± 0.009	0.716 ± 0.005	0.767 ± 0.013
Exp learn. Hawkes	0.764 ± 0.008	0.812 ± 0.008	0.714 ± 0.025	0.758 ± 0.020
Attention Hawkes	0.761 ± 0.007	0.796 ± 0.009	0.717 ± 0.014	0.751 ± 0.023
	Local target			
Without contex	0.569 ± 0.008	0.321 ± 0.003	0.535 ± 0.008	0.299 ± 0.011
Mean	0.592 ± 0.005	0.342 ± 0.005	0.543 ± 0.006	0.312 ± 0.006
Max	0.640 ± 0.005	**0.400 ± 0.006**	0.621 ± 0.006	0.256 ± 0.008
Attention	0.600 ± 0.009	0.348 ± 0.010	0.534 ± 0.016	0.301 ± 0.007
Learn. attention	0.583 ± 0.007	0.330 ± 0.008	0.590 ± 0.035	0.338 ± 0.025
Sym. attention	0.583 ± 0.007	0.329 ± 0.007	0.605 ± 0.027	0.350 ± 0.020
Kernel attention	0.582 ± 0.007	0.329 ± 0.007	0.572 ± 0.021	0.330 ± 0.023
Exp Hawkes	**0.649 ± 0.000**	0.366 ± 0.003	**0.638 ± 0.001**	**0.351 ± 0.001**
Exp learn. Hawkes	0.581 ± 0.012	0.322 ± 0.013	0.539 ± 0.034	0.293 ± 0.025
Attention Hawkes	0.635 ± 0.004	0.359 ± 0.005	0.598 ± 0.003	0.331 ± 0.001

Results of validation for different methods. The best values are **bolded**, the second values are underlined, the third values are double underlined.