Learning Enhanced Representations for Tabular Data via Neighborhood Propagation

Kounianhua Du, Weinan Zhang, Ruiwen Zhou, Yangkun Wang, Xilong Zhao, Jiarui Jin,

Quan Gan, Zheng Zhang, David Wipf

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Main contributions and motivation

We model tabular data as a hypergraph, where each distinct feature value forms a node and a collection of them, i.e., a data instance, forms a hyperedge. Then we design a novel architecture that **P**ropagates and **E**nhances the **T**abular data representations based on the hypergraph for target label prediction.

- We propose a retrieval-based hypergraph to capture the feature and label correlations among tabular data instances.
- We design an end-to-end graph neural network prediction model that unifies the product feature interaction, locality mining, and label enhancement.
- We utilize the observed labels in the resulting set to guide the feature learning process and use the propagated labels to enhance predictions.

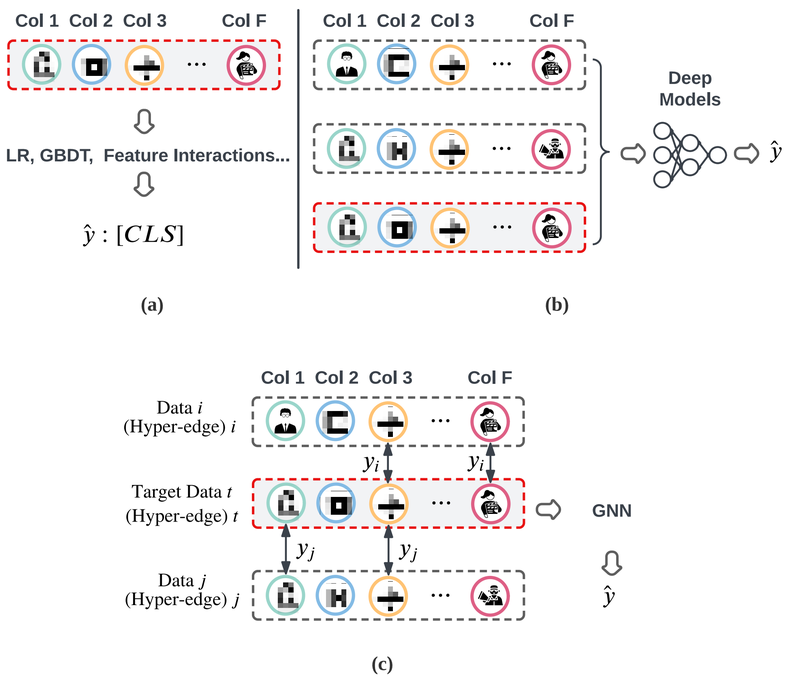


Figure. (a) The popular tree models and interaction-based models utilize a single data instance for prediction. (b) The retrieval-based methods take multiple data instances as input without sufficiently mining the interaction patterns among them. (c) The proposed PET models the multiple data instances set as a hypergraph and capture their correlations with the assistance of labels.

Methodology

The workflow of PET

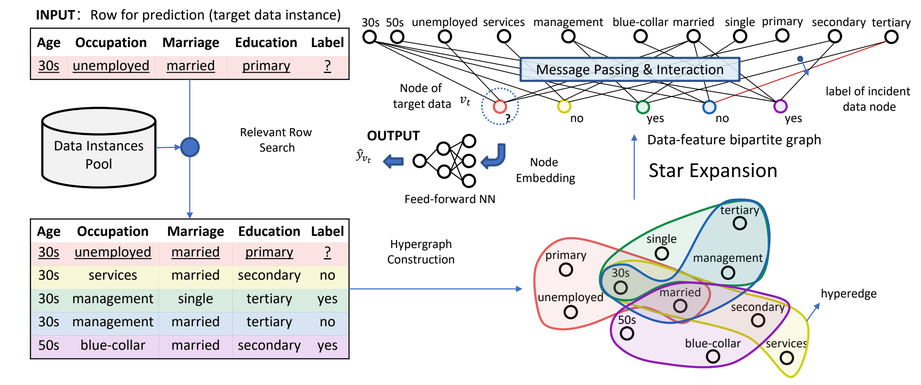


Figure. For each row to predict (top-left), PET retrieves a fixed number of relevant data instances (bottom-left) and constructs a hypergraph (bottom-right) from the resulting data instances set. After a star expansion (top-right), we get a data-feature bipartite graph with data instance nodes at the bottom and feature value nodes at the top. We then perform the proposed hypergraph neural network on the resulting graph and use the target data instance node representation for prediction.

Message Passing

- Initialization.

- Message Generation.

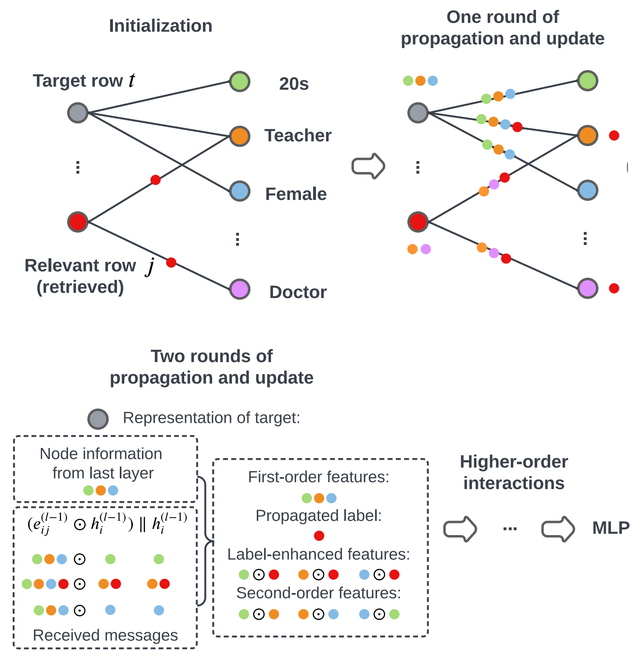
$$m_{ij}^{(l)} = (e_{ij}^{(l-1)} \odot h_i^{(l-1)}) \| h_i^{(l-1)}.$$

- Attention-based aggregation

$$Q_j^{(l)} = W_Q^{(l)} h_j^{(l-1)},$$
$$K_{ij}^{(l)} = W_K^{(l)} m_{ij}^{(l)},$$
$$V_{ij}^{(l)} = W_V^{(l)} m_{ij}^{(l)},$$
$$a_{ij}^{(l)} = softmax_{i \in N(j)}\left(Q_j^{(l)} K_{ij}^{(l)}\right),$$
$$n_j^{(l)} = \sum_{i \in N(j)} a_{ij}^{(l)} V_{ij}^{(l)}.$$

- Update

$$h_j^{(l)} = \sigma\left(W_N^{(l)}\left(h_j^{(l-1)} \| n_j^{(l)}\right)\right),$$
$$e_{ij}^{(l)} = \sigma\left(W_E^{(l)}\left(h_i^{(l)} \| h_j^{(l)} \| e_{ij}^{(l-1)}\right)\right).$$



Experiments

CTR prediction & Top-N Recommendation

Table 2: Result comparisons with baselines on CTR prediction task. ($K = 10$)

Models	Tmall			Taobao			Alipay		
	AUC	LogLoss	Rel.Impr.	AUC	LogLoss	Rel.Impr.	AUC	LogLoss	Rel.Impr.
GBDT	0.8319	0.5103	12.08%	0.6134	0.6797	44.08%	0.6747	0.9062	32.36%
DeepFM	0.8581	0.4695	8.66%	0.6710	0.6497	31.71%	0.6971	0.6271	28.1%
FATE	0.8553	0.4737	9.01%	0.6762	0.6497	30.70%	0.7356	0.6199	21.40%
TabGNN	0.8945	0.4158	4.24%	0.7294	0.6173	21.17%	0.8086	0.5849	10.44%
DIN	0.8796	0.4292	6.00%	0.7433	0.6086	18.90%	0.7647	0.6044	16.78%
DIEN	0.8838	0.4445	5.50%	0.7506	0.6084	17.74%	0.7502	0.6151	19.03%
SIM	0.8857	0.4520	5.27%	0.7825	0.5795	12.95%	0.7600	0.6089	17.50%
UBR	0.8975	0.4368	3.89%	0.8169	0.5432	8.19%	0.7952	0.5747	12.30%
RIM	0.9138	0.3804	2.04%	0.8563	0.4644	3.21%	0.8006	0.5615	11.54%
PET	**0.9324**	**0.3321**	–	**0.8838**	**0.4162**	–	**0.8930**	**0.4132**	–

Table 3: Result comparisons with baselines on top-n recommendation task. ($K = 10$)

Datasets	Metric	FPMC	TransRec	NARM	GRU4Rec	SASRec	RIM	PET
ML-1M	HR@1	0.0261	0.0275	0.0337	0.0369	0.0392	0.0645	**0.0904**
	HR@5	0.1334	0.1375	0.1418	0.1395	0.1588	0.2515	**0.2889**
	HR@10	0.2577	0.2659	0.2631	0.2624	0.2709	0.4014	**0.4404**
	NDCG@5	0.0788	0.0808	0.0866	0.0872	0.0981	0.1577	**0.1903**
	NDCG@10	0.1184	0.1217	0.1254	0.1265	0.1341	0.2059	**0.2390**
	MRR	0.1041	0.1078	0.1113	0.1135	0.1193	0.1704	**0.2006**
LastFM	HR@1	0.0148	0.0563	0.0423	0.0658	0.0584	0.0915	**0.1149**
	HR@5	0.0733	0.1725	0.1394	0.1785	0.1729	0.3468	**0.3621**
	HR@10	0.1531	0.2628	0.2227	0.2581	0.2499	0.5780	**0.6033**
	NDCG@5	0.0432	0.1148	0.0916	0.1229	0.1163	0.2165	**0.2381**
	NDCG@10	0.0685	0.1441	0.1185	0.1486	0.1409	0.2911	**0.3156**
	MRR	0.0694	0.1303	0.1083	0.1362	0.1289	0.2210	**0.2492**

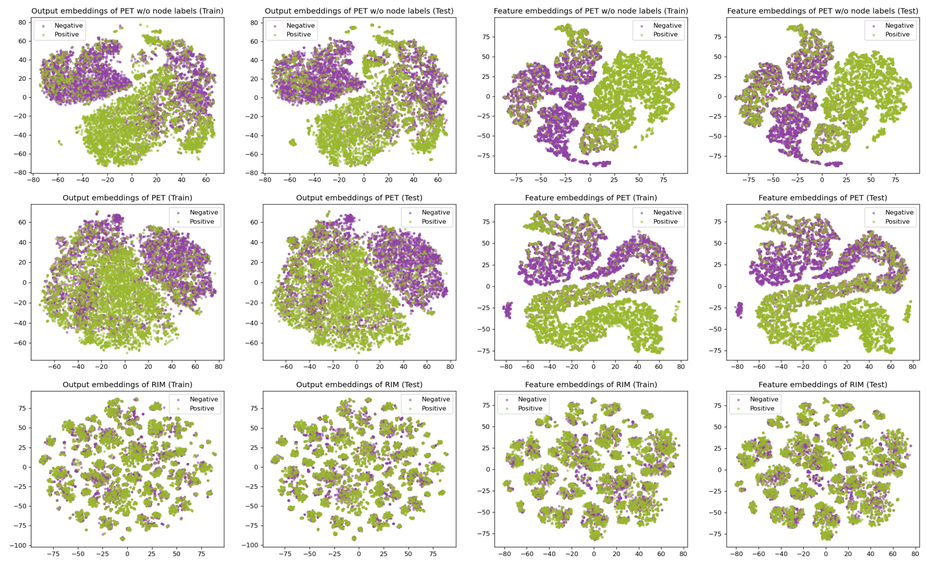
Label Usage

Table 4: Impact of label embeddings. (On randomly sampled data, $K = 10$)

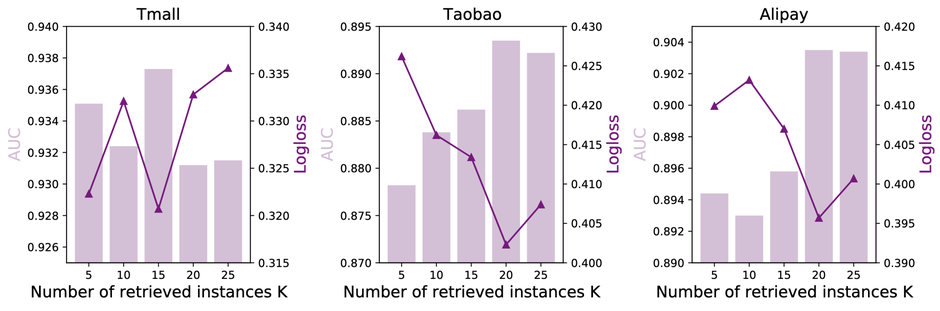
Models	Tmall		Taobao		Alipay	
	AUC	LogLoss	AUC	LogLoss	AUC	LogLoss
RIM	0.9120	0.3769	0.8587	0.4586	0.7845	0.5742
PET	0.9279	0.3387	0.8762	0.4279	0.8720	0.4201
PET (w/o edge labels)	0.9291	0.3367	0.8665	0.4465	0.8558	0.4776
PET (w/o node labels)	0.9233	0.3494	0.8431	0.4847	0.8518	0.4799
PET (w/o all labels)	0.9208	0.3568	0.8416	0.4815	0.8096	0.5719

Ablations

T-SNE visualizations for representations



Retrieval size study



References

Balasubramaniam Srinivasan, Da Zheng, and George Karypis. 2021. Learning over Families of Sets-Hypergraph Representation Learning for Higher Order Tasks. In Proceedings of the 2021 SIAM International Conference on Data Mining (SDM). SIAM, 756–764.

Jiarui Qin, Weinan Zhang, Rong Su, Zhirong Liu, Weiwen Liu, Ruiming Tang, Xiuqiang He, and Yong Yu. 2021. Retrieval & Interaction Machine for Tabular Data Prediction. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 1379–1389.