Cross-Regional Fraud Detection via Continual Learning (Student Abstract)

Yujie Li[1], **Yuxuan Yang**[1], **Qiang Gao**[1,2,*], **Xin Yang**[1,2]

[1]Southwestern University of Finance and Economics
[2]Kash Institute of Electronics and Information Industry
yujie_li@smail.swufe.edu.cn, 41810023@smail.swufe.edu.cn, qianggao@swufe.edu.cn, yangxin@swufe.edu.cn

Abstract

Detecting fraud is an urgent task to avoid transaction risks. Especially when expanding a business to new cities or countries, developing a totally new model will bring the cost issue and result in forgetting previous knowledge. This study proposes a novel solution based on heterogeneous trade graphs, namely **HTG-CFD**, to prevent knowledge forgetting of cross-regional fraud detection. Specifically, a novel heterogeneous trade graph is meticulously constructed from original transactions to explore the complex semantics among different types of entities and relationships. Motivated by continual learning, we present a practical and task-oriented forgetting prevention method to alleviate knowledge forgetting in the context of cross-regional detection. Extensive experiments demonstrate that HTG-CFD promotes performance in both cross-regional and single-regional scenarios.

Introduction

For fraud detection, traditional rule-based methods heavily rely on human prior knowledge, resulting in detection bias and the collapse of tackling more complex patterns. First, real-world transaction data contains various types of attached entities except for the customer and merchant entities. Second, existing solutions are only suitable for a narrow scope owing to the limitation of data collection. However, we usually either use the previously trained model or develop a totally new model, resulting in the uncertainty of capturing financial patterns. Developing a totally new model leads to expensive costs of resources while re-using the learned model can be a more appropriate solution. To remedy the above challenges, this paper investigates a new problem, cross-regional fraud detection (CFD), which aims to continuously detect fraudulent activities in the process of business expansion to new regions. We regard CFD as a continual learning task and propose a novel solution with **H**eterogeneous **T**rade **G**raph Learning for **C**ross-regional **F**raud **D**etection (HTG-CFD). Firstly, HTG-CFD constructs a comprehensive heterogeneous trade graph (HTG) to expose the complex but semantic interactions among different entities. Next, HTG-CFD builds multiple transaction-like meta-paths and then performs with graph attention networks to learn the complex structure and interactions among various entities. In the end, we devise a novel task-oriented forgetting prevention module including knowledge replay and parameter smoothing to address the forgetting issue.

[*]Corresponding Author (qianggao@swufe.edu.cn).

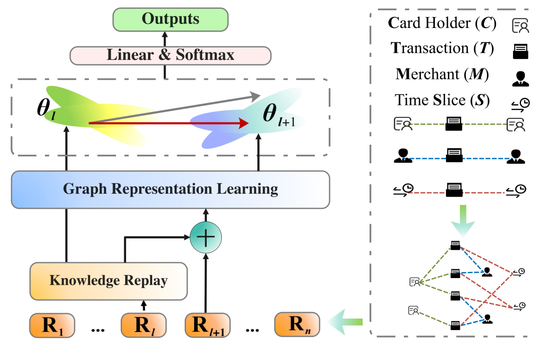

Figure 1: The framework of HTG-CFD.

Methodology

Problem definition. *Given a set of HTGs from n regions $(\mathcal{G}_1, \cdots, \mathcal{G}_l, \cdots, \mathcal{G}_n)$ with $\mathcal{G}_l = (\mathcal{V}_l, \mathcal{E}_l, \mathbf{X}_l)$, our task is to learn a series of neural function $(\Phi_1, \cdots, \Phi_l, \cdots, \Phi_n)$, and each function $\Phi_l : \mathcal{V}_l \rightarrow \mathbb{R}^d$ maps nodes to a low-dimensional space d, i.e., $d \ll |\mathcal{V}_l|$ and use a dense layer $f_l : \mathbb{R}^d \rightarrow \{0, 1\}$ to detect whether a node is fraudulent.*

Heterogeneous Trade Graph Contriving. For each HTG $\mathcal{G}_l$, we first build it from l-th region's structured trade data. Specifically, there are four types of nodes in $\mathcal{G}_l$, i.e., credit card holders C_l, merchants M_l, time slices S_l, and transaction IDs T_l. That is to say, $\mathcal{V}_l = \{C_l, M_l, S_l, T_l\}$. Next, we generate massive meta-paths to explore the diverse semantic interactions among heterogeneous entities, i.e., TCT (transaction-card holder-transaction), TMT (transaction-merchant-transaction), and TST (transaction-time slice-transaction). Notably, timestamp is a special node type in HTG because the frequency of fraudulent transactions changes abruptly across different times while legitimate transactions are much more stable. To alleviate the scale of timestamps, we split the time slices into 24 fixed intervals. Then, we extend a hierarchical graph attention network to generate each transaction node representation $\mathbf{x}_i^l$.

Prototype-based Knowledge Replay. As shown in Fig. 1, assume we have finished the learning task of l-th region. In order to preserve the existing knowledge from region l, we sample a small experience buffer $\mathbb{B}_l$ from $\mathcal{D}_l$ for knowledge replay. Then, we propose a prototype-based replay method to alleviate the training instability problem in the replay. As each transaction in $\mathbb{B}_l$ is associated with an attribute vector $\mathbf{x}_i^l$, our prototypes are based on the average attribute vector regarding $\{x_i^l\}_1^{|\mathbb{B}_l|}$, which can be defined as:

$$\mathbf{c}^l = \frac{1}{\mathbb{B}_l} \sum_{\mathbf{x}_i^l \in \mathbb{B}_l} \mathbf{x}_i^l. \quad (1)$$

We then use $\mathbf{c}^l$ and a Gaussian prior to generating a prototype $\mathbf{x'}_i^l$ for each $\mathbf{x}_i^l$. Finally, we collect a similar buffer $\mathbb{B}'_l$ and use it with $\mathbb{B}_l$ for experience replay together. Notably, we only sample transactions from the last task instead of all trained tasks to avoid memory costs.

Regularization-based Parameter Smoothing To further alleviate the knowledge forgetting, Inspired by recent smoothing research (Zhou and Cao 2021), we use Fisher Information $\mathbb{F}$ to evaluate the importance of the trained parameters Θ_l. For instance, when the new region $l+1$ arrives, we add a smoothing term to constrain the parameter optimization based on $\mathbb{F}_{l+1}$, which can be formulated as:

$$\mathcal{L}_{l+1}^{s} = \frac{\lambda}{2} \sum_i \mathbb{F}_{i+1} \left(\Theta_{l+1}(i) - \Theta_l(i)\right)^2, \quad (2)$$

where λ is a hand-craft weight. Notably, fraud patterns usually have common characteristics, so it is necessary to consider task similarity in parameter smoothing as well. Inspired by recent multi-task learning (Li et al. 2017), a regularization term can be used to measure the parameter distance. Thus, our final objective is to minimize:

$$\mathcal{L}_{l+1} = \mathcal{L}_{l+1}^{c}(\hat{y}, y) + \mathcal{L}_{l+1}^{s} + \gamma ||(||\Theta_{l+1}||_2, ||\Theta_l||_2)||_1. \quad (3)$$

where the first term is cross-entropy loss of task $l+1$, the second term is the $l_{2,1}$-norm, and γ is a hand-craft weight.

Experiments and Results

Datasets and Baselines We conduct experiments on five regions from a USA financial database (https://kaggle.com/kartik2112/fraud-detection). The proposed HTG-CFD is first compared with three fraud detection baselines: GEM (Liu et al. 2018), GATNE (Cen et al. 2019), HAN (Wang et al. 2019). Then, to verify the effectiveness of forgetting prevention, we also implement popular continual learning methods—EWC (Kirkpatrick, Pascanu et al. 2017), GEM-CL (Tang and Matteson 2020), HAT (Serra et al. 2018)—to tackle the proposed CFD problem.

Results We conduct experiments under two different settings: single-regional and cross-regional detection. From Table 1, HTG-CFD significantly outperforms the best baseline by 2%-8%. Also, HTG-CFD achieves the best performance on all metrics, indicating that the HTG-CFD is effective in handling knowledge forgetting. Notably, HTG-CFD outperforms HAT which has no forgetting, demonstrating that

	Method	Recall	AUC	F1
Single-regional	GEM	71.65	80.61	77.44
	GATNE	66.32	88.84	80.59
	GAN	74.28	87.51	80.97
	HTG-CFD	**89.74**	**95.73**	**87.76**
Cross-regional	EWC	89.62	94.54	87.00
	GEM-CL	78.15	85.51	79.41
	HAT	74.33	95.89	87.54
	HTG-CFD	**93.38**	**98.21**	**92.22**

Table 1: Overall Performance.

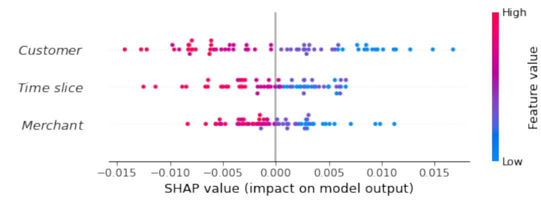

Figure 2: SHAP results.

HTG-CFD has a promising knowledge transfer ability that can promote the old tasks' performances as well. To present the interpretability investigation of the node contribution in HGT, we randomly mask 0%, 30%, 60%, and 90% of nodes for each type. We use Grid Search and SHAP to visualize the importance of each type. As Fig. 2 shows, the more important node type is closer to the top area. Clearly, $Customer$ is most important, followed by $TimeSlice$.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62102326), the Natural Science Foundation of Sichuan Province (No. 22NSFSC3049), and the Special Fund of Chengdu SWUFE & Jiaozi Fintech Innovation Institute (No. CGZH20210107).

References

Cen, Y.; Zou, X.; Zhang, J.; Yang, H.; Zhou, J.; and Tang, J. 2019. Representation learning for attributed multiplex heterogeneous network. In *KDD*, 1358–1368.

Kirkpatrick, J.; Pascanu, R.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *PNAS*, 3521–3526.

Li, C.; Yan, J.; Wei, F.; Dong, W.; Liu, Q.; and Zha, H. 2017. Self-paced multi-task learning. In *AAAI*.

Liu, Z.; Chen, C.; Yang, X.; Zhou, J.; Li, X.; and Song, L. 2018. Heterogeneous graph neural networks for malicious account detection. In *CIKM*, 2077–2085.

Serra, J.; Suris, D.; Miron, M.; and Karatzoglou, A. 2018. Overcoming catastrophic forgetting with hard attention to the task. In *ICML*, 4548–4557.

Tang, B.; and Matteson, D. S. 2020. Graph-Based Continual Learning. In *ICLR*.

Wang, X.; Ji, H.; Shi, C.; Wang, B.; Ye, Y.; Cui, P.; and Yu, P. S. 2019. Heterogeneous graph attention network. In *WWW*, 2022–2032.

Zhou, F.; and Cao, C. 2021. Overcoming catastrophic forgetting in graph neural networks with experience replay. In *AAAI*, 5, 4714–4722.