文章编号:1672-3961(2025)06-0001-12 DOI:10.6040/j.issn.1672-3961.0.2024.269

融合主客观评价的图数据 Top-k 频繁模式挖掘

黄芳,王欣*,高国海,沈玲珍,付勋,方宇
(西南石油大学计算机与软件学院, 四川 成都 610500)

摘要:为解决传统 Top-k 模式挖掘结果难以满足用户实际需求的问题,提出一种融合主客观评价的图数据 Top-k 频繁模式挖掘。通过基于最小 DFS 编码的模式表征技术,实现对模式的编码;搭建基于孪生神经网络的图模式评价模型(graph patterns evaluation model, GPEM),学习模式对之间的偏好关系,实现对模式的主观偏好预测;设计融合主客观的模式兴趣度评价函数,指导 Top-k 模式挖掘。在 6 个真实图数据集上的试验结果表明,GPEM 在多项指标上优于其他模型,准确率最高可达 93%。

关键词:孪生神经网络;频繁模式挖掘;兴趣度评价函数
中图分类号:TP311 **文献标志码**:A
引用格式:黄芳,王欣,高国海,等. 融合主客观评价的图数据 Top-k 频繁模式挖掘[J]. 山东大学学报(工学版),2025,55(6):1-12.
HUANG Fang, WANG Xin, GAO Guohai, et al. Mining Top-k frequent patterns for graphs based on subjective and objective metrics[J]. Journal of Shandong University (Engineering Science), 2025, 55(6):1-12.

Mining Top-k frequent patterns for graphs based on subjective and objective metrics

HUANG Fang, WANG Xin*, GAO Guohai, SHEN Lingzhen, FU Xun, FANG Yu
(School of Computer Science and Software Engineering, Southwest Petroleum University, Chengdu 610500, Sichuan, China)

Abstract: In order to solve the problem that traditional Top-k pattern mining results failed to meet the users′ practical needs, a graph data Top-k frequent pattern mining approach that integrates subjective and objective evaluations was proposed. A pattern representation technique based on minimum DFS coding was introduced to encode patterns. The graph patterns evaluation model (GPEM) was built based on a siamese neural network, which learned the biased order relationships between pattern pairs and predicted subjective preference of patterns. A pattern interestingness evaluation function that combined subjective and objective factors was designed to guide Top-k pattern mining. Experiments on six real graph datasets demonstrated that GPEM outperformed other models on various metrics, with up to 93% accuracy.

Keywords: siamese neural network; frequent pattern mining; interestingness evaluation function

0 引言

频繁模式挖掘(frequent pattern mining, FPM)是图数据挖掘领域重要问题之一,旨在从图数据中发现支持度不低于设定阈值的频繁模式。FPM 主要有两种设定:基于图集的 FPM[1] 和基于单一大图的 FPM[2]。近年来,基于单一大图的 FPM 广泛应用于社交网络分析、生物信息学、化学信息学等领域,但现有方法运算开销随输入规模指数增长,在大图上难以实施[3],且大图上结果集往往较大,不利于用户理解与使用。为此,研究人员设计 Top-k FPM 算法[4-6],用客观兴趣度指标如支持度、模式大小,对模式进行度量排序,以发现排名前 k 名的频繁模式。但以支持度为评估指标的 Top-k 模式较多为结构简单的单边模式,以模式大小为评估指标的 Top-k 模式支持度往

收稿日期:2024-10-15
基金项目:国家自然科学基金资助项目(62172102);四川省科技创新人才基金资助项目(2022JDRC0009)
第一作者简介:黄芳(2000—),女,四川眉山人,硕士研究生,主要研究方向为数据挖掘、机器学习。E-mail:huangfang1632021@163.com
*通信作者简介:王欣(1981—),男,江苏扬州人,教授,博士生导师,博士,主要研究方向为机器学习、数据挖掘及油气人工智能。
E-mail:xinwang@swpu.edu.cn

往不高且不一定符合用户实际兴趣。因此,设计一种能够兼顾用户主观偏好和模式客观信息的模式评价方法尤为重要。挖掘此类 Top-k 频繁模式需解决三个关键问题:

(1)如何将模式转为向量表示,从而输入评价模型?

(2)如何有效捕获用户偏好信息并设计模式评价模型,实现对模式主观评价?

(3)如何将主观评价与客观评价结合,设计评价指标,实现 Top-k 频繁模式挖掘?

针对上述问题,本研究提出一种结合最小深度优先搜索(depth first search, DFS)编码和 One-hot 编码的模式表示方法,解决模式初始化问题;提出一种图模式主观评价预测模型(graph patterns evaluation model, GPEM),用孪生神经网络学习模式对偏好关系,并利用单塔结构作为主干网络,实现对模式主观评价;结合用户主观偏好和模式客观信息挖掘满意度更高的 Top-k 频繁模式。

1 相关工作

以往工作中,大量 FPM 算法致力于挖掘精准结果集,但用户通常只需前 k 个最感兴趣模式,因此 Top-k 模式挖掘的重要性日益凸显。在 Top-k 模式挖掘算法方面,文献[4]提出 FastPat 框架,通过设定最小图像支持度上界,实现算法提前终止以发现频繁模式;文献[5]提出保持提前终止属性的高效 Top-k 模式挖掘并行算法 DisMiner,通过"前瞻回溯"和"部分求值"策略发现频繁模式。为满足用户对不同兴趣度指标的需求,文献[7]提出 Resling 框架,采用基于随机游走的算法进行排名,挖掘前 k 个代表性频繁模式,但处理大规模图数据时内存需求可能较高;文献[2]设计一种近似算法 AprTopK,采用"逐层"策略,保证算法提前终止性,从而发现频繁且有趣的子模式。但这些算法可能产生冗余且缺乏多样性的结果,需额外机制确保结果实用性。为降低枚举成本,文献[8]引入基于等价顶点的图压缩技术,通过使用紧凑的图快照表示、合适的时间索引及估计权重上界方法,减少冗余验证并提高整体性能;文献[9]采用整体最佳优先探索策略和"最佳"压缩数据结构,识别单个图中频繁共存的子模式对;文献[6]提出一种将挖掘和排名阶段结合方法;文献[10]提出 ItrMiner 挖掘算法,设计兴趣度指标对模式排名,该指标能同时兼顾模式支持度和大小,有效减少低兴趣度候选模式生成。然而,上述基于客观评价指标的挖掘技术难以准确反映用户实际兴趣。为此,文献[11]利用层次分析法学习用户偏好,确定不同兴趣度指标权重,优化模式学习和排序;文献[12]提出不确定数据上的交互式 Top-k 频繁模式挖掘技术 ITUFP,考虑挖掘过程中不确定性,通过交互式技术提升挖掘效率;文献[13]提出 WaveLSea 方法,将用户主观兴趣融入模式搜索过程中,引导用户浏览感兴趣的挖掘结果。

2 基本概念

定义 1 图和子图[14]:一个数据结构图被定义为三元组 $G=(V,E,L)$,其中 V 是顶点集合;E 是边集合;L 表示节点 V 的标签。图 $G_s=(V_s,E_s,L_s)$ 是 $G=(V,E,L)$ 的子图,其中,$V_s\subseteq V$, $E_s\subseteq E$,针对每个节点 $v\in V_s$,都有 $L_s(v)=L(v)$。

定义 2 模式和子模式[2]:模式 Q 定义为 $Q=(V_p,E_p,f_v)$,其中,V_p 和 E_p 分别是节点和边的集合。对于节点 $u\in V_p$, $f_v(u)$ 被定义为 $A=a$ 形式的原子公式连结。A 表示节点 u 的一个属性,a 是属性 A 对应值。子模式 $Q'=(V'_p,E'_p,f'_v)$ 和模式 $Q=(V_p,E_p,f_v)$,若满足 (V'_p,E'_p) 是 (V_p,E_p) 一个子图,并且 f'_v 是 f_v 一个限制,则模式 Q' 是 Q 模式的子模式,用 $Q'\subseteq Q$ 表示。

定义 3 支持度[2]:给定模式 Q 和图 G,支持度为模式 Q 在图 G 中对应匹配出现的频率,记为 Sup(Q,G)。基于图像的最小支持度是一个广泛使用的度量标准,它保证模式扩展的反单调性。

$$\mathrm{Sup}(Q,G)=\min\{|P(u)|,\ u\in V_p\},\tag{1}$$

式中,$P(u)$ 为模式中节点 u 在图 G 上的匹配去重后的节点集合。

定义 4 频繁模式挖掘[15]:给定图 G 和支持度阈值 θ,频繁模式挖掘旨在从图 G 中发现一个频繁模式集合 S, S 中的任意模式 Q 的支持度满足 $\mathrm{Sup}(Q,G)\geqslant\theta$。

定义 5 最小 DFS 编码[16]:通过深度优先方式遍历图 G,可以生成多棵 DFS 编码树,每棵树对应一个唯一 DFS 编码序列。该序列是一组有序五元组,每个五元组定义为 $(u,v,L(u),L(e),L(v))$,其中 u 和 v 是边 e 节点,$L(u)$ 和 $L(v)$ 是节点标签,$L(e)$ 是边标签。在所有 DFS 编码序列中,按字典序排列最小编码称为图 G 的最小 DFS 编码。对于模式 Q,其最小 DFS 编码常被用于表征它。

3 主客观融合的 Top-k FPM

一个社交网络图如图 1(a)所示,其中每个节点

表示一个具有特定职称的用户(如数据库管理员(database administrator, DBA)、程序员(programmer, PRG)、业务分析师(business analyst, BA)、软件测试人员(software tester, ST)和项目经理(project manager, PM));边表示用户之间联系,虚线框表示一组等价节点,它们不仅标签相同,且邻居节点也相同。基于支持度和基于模式大小的Top-k频繁模式分别如图 1(b)、(c)所示。以支持度作为评估指标的 Top-k 模式均为结构简单的单边模式,而以模式大小作为评估指标的 Top-k 模式虽然结构复杂、信息量大,但支持度往往不高且与用户的实际兴趣未必相符。

针对上述问题,本研究结合用户主观偏好和模式客观信息挖掘满意度更高的 Top-k 频繁模式,所提方法整体流程如图 2 所示。首先,对单一大图使用共同邻居感知随机游走采样算法[17]采样,在采样图上进行模式挖掘[10],得到一组具有不同结构和大小的频繁模式;其次,通过融合模式最小 DFS 编码和标签 One-hot 编码表征模式;然后,搭建参数共享的孪生神经网络 GPEM 学习模式对偏序关系,实现模式主观评价;最后,设计融合主客观的模式兴趣度评价函数,指导 Top-k模式挖掘。通过融入用户兴趣度偏好所发现的 Top-k 模式如图 1(d)所示,由图可见以用户实际兴趣为度量指标能更精准地发现满足用户需求的模式。

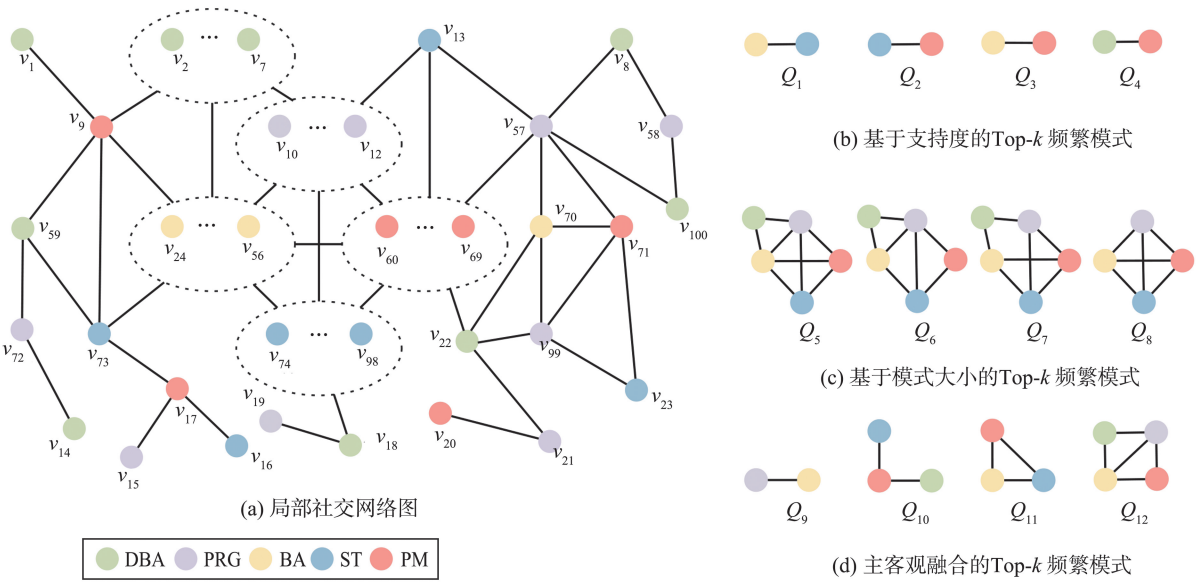


图 1　社交网络图和不同评价指标下的 Top-k 频繁模式

Fig.1　A social network graph and the Top-k frequent patterns under different evaluation metrics

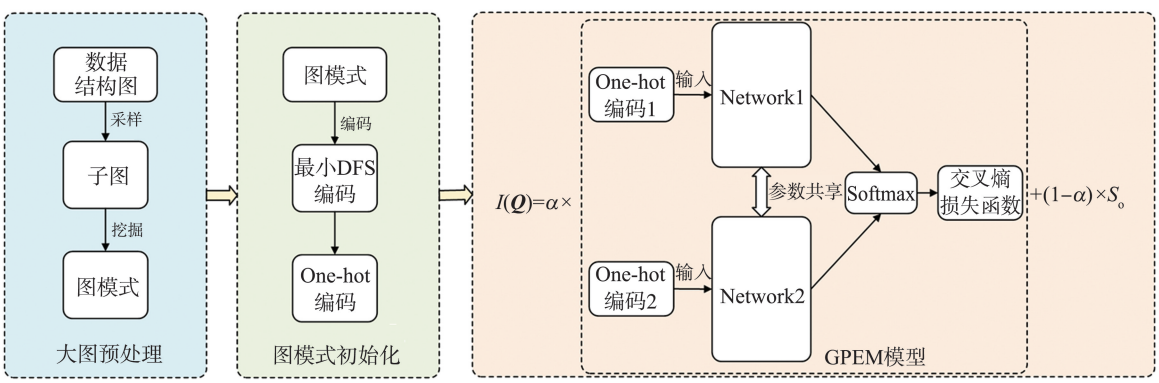


图 2　整体流程图

Fig.2　Overall flow chart

3.1　模式评价模型

3.1.1　图模式初始化

DFS 编码通过深度优先搜索完成。在 DFS 遍历中,对于当前访问的边 $e=(u,v)$,若 v 为首次访问,则 $u<v$,称 e 为前向边,否则为后向边。对于节点 A、B,假设 A 和 B 的前向边集合和后向边集合分别为 $E_{A,f}$、$E_{A,b}$、$E_{B,f}$、$E_{B,b}$。有 $e_a=(u_a,v_a,L(u_a),L(e_a),L(v_a))<e_b=(u_b,v_b,L(u_b),L(e_b),L(v_b))$[16]。若下列条件之一成立:

(1) $e_a\in E_{A,b}$且 $e_b\in E_{B,f}$;

(2) $e_a \in E_{A,b}$, $e_b \in E_{B,b}$且 $v_a < v_b$;

(3) $e_a \in E_{A,b}$, $e_b \in E_{B,b}$, $v_a = v_b$ 且 $L(e_a) < L(e_b)$;

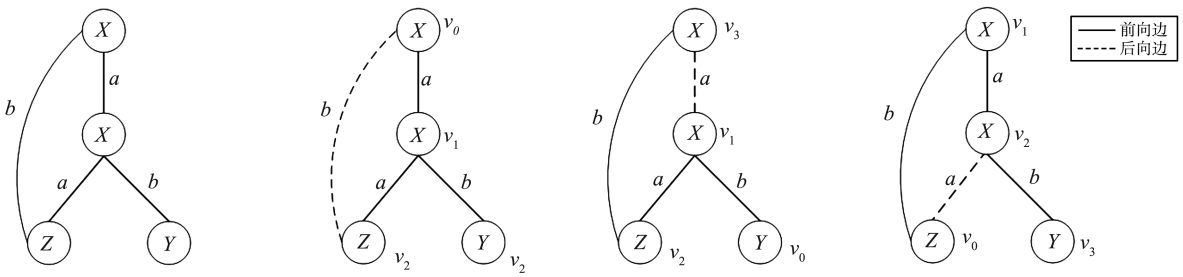
(4) $e_a \in E_{A,f}$, $e_b \in E_{B,f}$且 $u_b < u_a$;

(5) $e_a \in E_{A,f}$, $e_b \in E_{B,f}$, $u_b = u_a$ 且 $L(u_a) < L(u_b)$;

(6) $e_a \in E_{A,f}$, $e_b \in E_{B,f}$, $u_b = u_a$, $L(u_a) = L(u_b)$且 $L(e_a) < L(e_b)$;

(7) $e_a \in E_{A,f}$, $e_b \in E_{B,f}$, $u_b = u_a$, $L(u_a) = L(u_b)$, $L(e_a) = L(e_b)$ 且 $L(v_a) < L(v_b)$。

如图 3 所示,执行深度优先遍历,构造其 DFS 树。一个图可以有多个不同 DFS 树。图 3(b)~(d)中实线边形成的树是图 3(a)三个不同的 DFS 树。遍历过程不同,导致节点序不同(节点下标表示其访问顺序),进而形成不同 DFS 编码。图 3(b)~(d)DFS 树对应 DFS 编码如表 1 所示。对比 3 个编码,边 $e_0=(0,1)$为前向边,先按规则(4)比较源节点,源节点均为 0,再按规则(5)比较源节点标签,由字典序得 $X<Y<Z$,可知表 1 中 b 对应编码是图 3(a)的最小 DFS 编码。



(a) 原始图 (b) DFS树1 (c) DFS树2 (d) DFS树3

图 3 深度优先搜索树

Fig.3 Depth-first searchtree

表 1 图 3(b)~3(d)的 DFS 编码

Table 1 DFS codes for Fig.3(b) to 3(d)

边	DFS 编码		
	图 3(b)	图 3(c)	图 3(d)
e_0	(0,1,X,a,X)	(0,1,Y,b,X)	(0,1,Z,b,X)
e_1	(1,2,X,a,Z)	(1,2,X,a,Z)	(1,2,X,a,X)
e_2	(2,0,Z,b,X)	(2,3,Z,b,X)	(2,0,X,a,Z)
e_3	(1,3,X,b,Y)	(3,1,X,a,X)	(2,3,X,b,Y)

先根据频繁模式最小 DFS 编码涉及的节点和边,构造节点和边的标签词典,确定编码长度,生成标签的 One-hot 编码。再把模式的最小 DFS 编码与其中涉及标签的 One-hot 编码融合,形成更新编码,作为模式表征。

给定模式 Q_1 和 Q_2,令节点标签 DBA、PRG、BA、ST、PM 分别为 1、2、3、4、5,边标签为 0,得两个模式最小 DFS 编码:$Q_1=[(0,1,2,0,3)]$, $Q_2=[(0,1,1,0,5),(1,2,5,0,4)]$。将它们与标签 One-hot 编码融合,产生最终表示。具体而言,先分别统计所有节点、节点标签和边标签的维度,节点维度为 3(编号 0、1、2),节点标签维度为 5(标签 1、2、3、4、5),边标签维度为 1(边标签固定为 0)。以 5 为编码长度,对节点、节点标签、边标签进行独立编码,对于维度小于 5 的编码,在其末尾以 0 填充;若模式边数少于最大边数,剩余部分也用 0 填充,保证所有模式编码长度一致。模式初始化过程如图 4 所示。

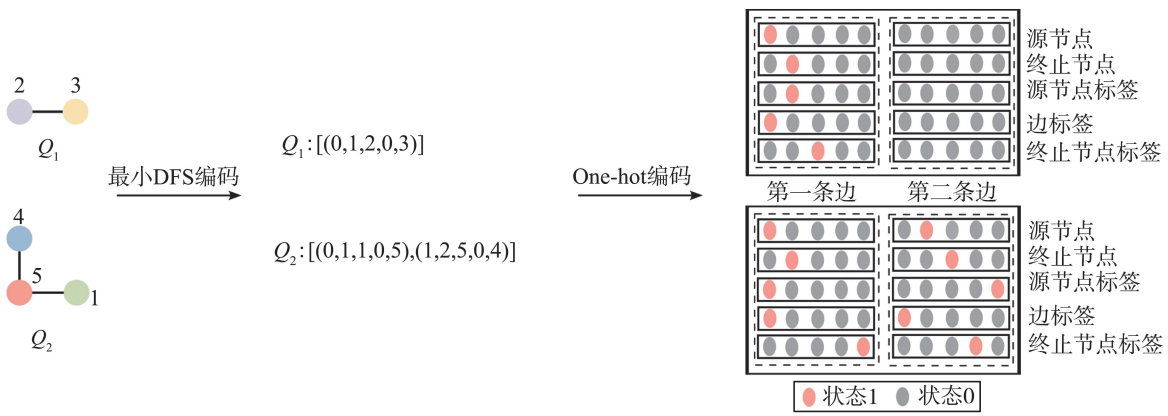


图 4 模式初始化过程

Fig.4 Patterns initialization process

3.1.2　GPEM

本研究构建基于孪生神经网络[18]的预测模型 GPEM,其结构图如图 5 所示。首先,输入的 One-hot 编码通过两个参数共享的卷积神经网络(convolutional neural network, CNN)模块处理,每个模块包含三个卷积层,每个卷积层之后接 ReLU 激活函数及池化层,可自动检测并提取有效特征,提升模型性能及其泛化能力,最终输出经 Flatten 层转为一维向量,作为后续多层感知机(multilayer perceptron, MLP)模块输入。随后,一维特征向量被送入后续的两个 MLP 模块(每个模块由一系列全连接层组成,每层都融合了激活函数、批归一化和 Dropout 正则化机制)训练以预测用户偏好。具体 MLP 模块如图 6 所示。整个网络架构通过其层次化和模块化设计,确保从输入数据中有效提取关键信息,并构建准确的偏好预测模型。在此基础上,每个偏序对生成的二维向量得分反映用户偏好强弱,其中分数较高的向量代表更受偏好的选择。通过 Softmax 函数将得分向量转换为(0, 1)区间内的概率分布,确保两个概率值之和为 1。模型通过交叉熵损失函数评估预测值与真实标签之间的差异,并通过该损失指导梯度下降过程,从而优化模型性能。最后,从训练完成的孪生模型中抽取单塔结构,用于预测单个输入向量分数。通过比较所预测得分的排名与原始偏序对的拓扑排名,评估排名的相似性。

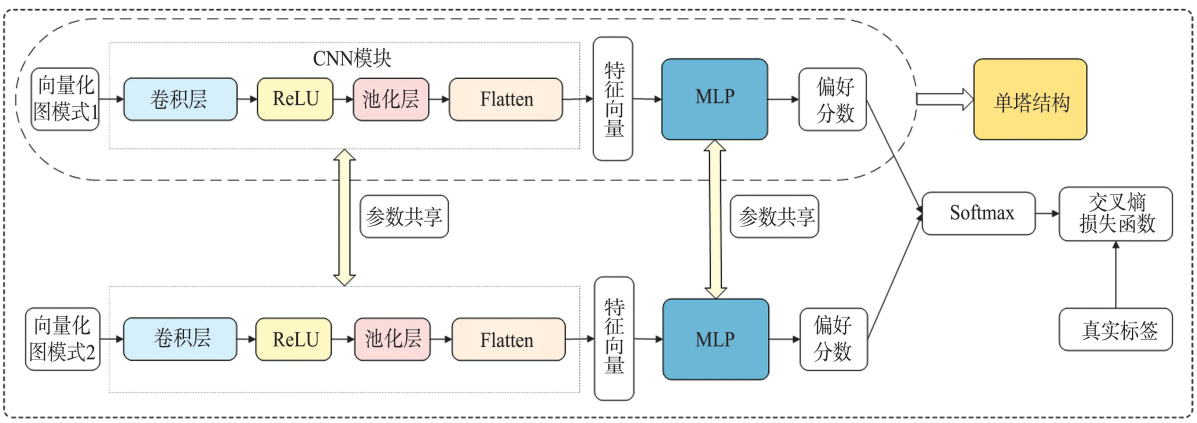


图 5　基于孪生神经网络的主观评价模型流程图

Fig.5　Flow chart of subjective evaluation model based onsiamese neural network

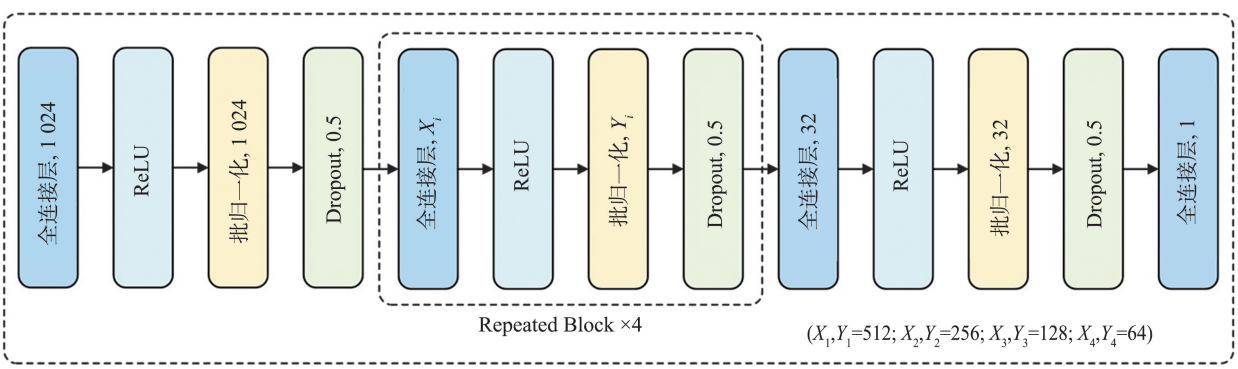


图 6　孪生神经网络中 MLP 模块

Fig.6　MLP module insiamese neural network

为衡量模型预测结果和实际标签间差异,引入交叉熵损失函数

$$L=(y)(-\ln(\hat{y}))+(1-y)(-\ln(1-\hat{y})), \quad (2)$$

式中,y 为真实标签值,$\hat{y}$ 为模型相应类别的预测概率。通过最小化交叉熵损失指导梯度下降,可以优化模型参数,使预测的概率分布尽可能接近真实标签分布。构建参数共享的孪生神经网络后,进一步采用单塔结构作为主观评价的模型。

3.1.3　基于偏好关系的全局排序策略

由于输入模式以偏序对形式呈现,无全局排序结果,无法直接对比预测结果与真实排序评估模型性能。为此,本研究提出一种基于偏好关系的全局排序策略,根据模式间偏序关系构建拓扑排序实现模式全局排序。对模式集合中的每对模式,按偏好关系在图 G 中构建有向边,如模式 Q_1 相较 Q_2 更受偏好,则在有向图中建立一条从 Q_1 指向 Q_2 的有向边。在图 G 上计算节点拓扑排序时,先通过 DFS 进行环检测,识别强连通分量并将其收缩为超点形成有向无环图(directed acyclic graph, DAG),再执行拓扑排序得到全体模式偏好

顺序。以图 1 中 Q_1—Q_8 为例,根据用户偏好信息得 Q_1—Q_8 各模式偏好关系,如图 7(a)所示,如[Q_1,Q_2,[1,0]]表示对于模式 Q_1 和 Q_2,用户更偏向 Q_1,据此建立的偏好图如图 7(b)所示。初始化时所有节点设为未访问状态,经 DFS 检测到两个环,即 Q_1→Q_2→Q_3→Q_1 和 Q_8→Q_5→Q_6→Q_7→Q_8,将这两个环分别合并为超节点 Q_{123} 和 Q_{5678},原图转换为新 DAG,如图 7(c)所示。对每个节点执行 DFS 操作,完成后将其放入栈中。所有节点搜索完毕,按栈顶到栈底顺序得全局拓扑排序 Q_{123}、Q_4、Q_{5678},如图 7(d)所示。对于两个超节点中模式,内部顺序随机处理。

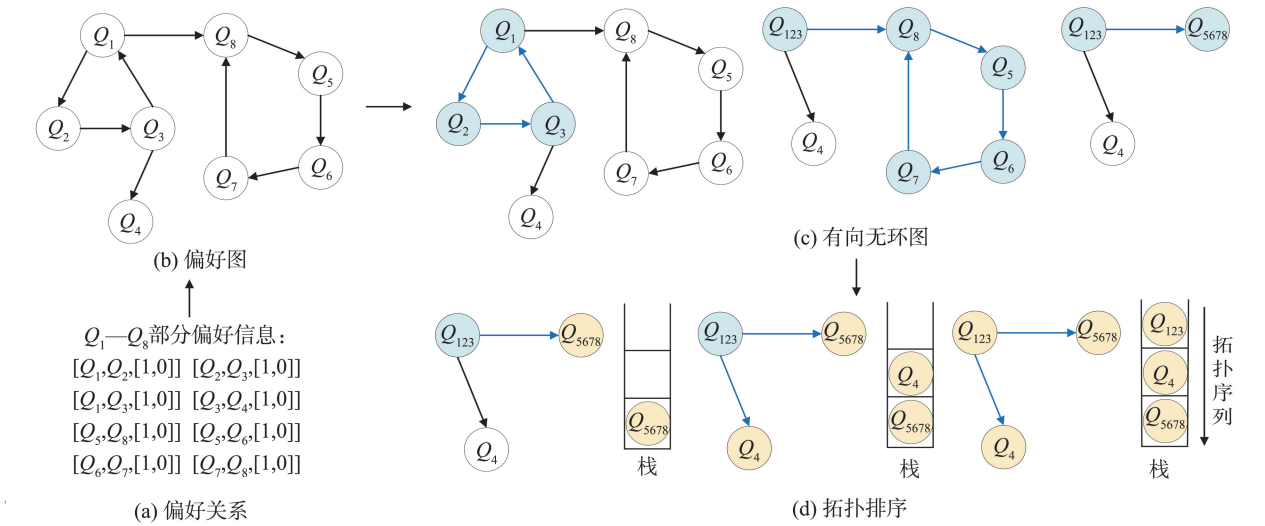


图 7 偏好关系的拓扑排序

Fig.7 Topological sorting of preference relations

3.2 主客观融合的 Top-k FPM

本研究提出一种新的兴趣度指标 I 对模式进行评估,旨在挖掘主客观融合 Top-k 频繁模式。模式 Q 的兴趣度定义为

$$I(Q)=\alpha\times S_s+(1-\alpha)\times S_o, \tag{3}$$

式中:α 为主观分数权重,$\alpha\in[0,1]$;S_s 为主观分数,由 GPEM 得出;S_o 为客观分数,由公式$\frac{1}{1+2^{-(|Q|)}}\times \mathrm{Sup}(Q,G)$得出[10],$\mathrm{Sup}(Q,G)$为模式 Q 在图 G 中的支持度,$|Q|$为模式大小。

图 1(d)中模式 Q_9、Q_{11} 支持度分别为 5、11,模式大小分别为 1、3,Q_{11} 的客观分数为 9.78,Q_9 的客观分数为 3.34,客观上 Q_{11} 被认为是更有价值和更有趣的模式。这在平衡模式大小和支持度方面具有一定优势,可以弱化小模式支持度过大以及大模式支持度偏低带来的影响。

为便于数据比较分析,用最小最大归一化法[18]将数据线性变换到[0,1]范围。将归一化后主观分数和客观分数代入公式(3)得模式兴趣度。

4 试验

为评估本研究所设计方法性能,在真实数据集上将本研究所提模型与其他模型对比,考察模型评估准确性和稳定性。试验环境为一台配备 2.4 GHz CPU、16 GB RAM 的 Windows11 主机。代码均用 Python3.8 编写。模型优化器采用 AdamW,初始学习率设为 0.000 1,权重衰减设为 0.3。模型损失函数采用交叉熵,引入 ReduceLROnPlateau 作为学习率调度器。测试集上损失无改进时,调度器通过设置衰减为 0.3 和容忍轮数为 5 自动调整学习率。

4.1 试验数据

试验采用如下 6 组真实图数据集,其统计信息如表 2 所示。

(1) Amazon[19],产品联合采购网络,当两个商品 a 和 b 被客户同时购买频次达到一定数量时,就会形成边(a,b)。

(2) DBLP[20],论文出版网络图,每个节点代表一篇论文,每条边代表两位作者间合作关系。

(3) Mico[21],微软合作作者构建网络,每个节点代表作者,每条边代表两位作者间合作关系。

(4) Patent[22],专利引文网络,每个节点代表一个专利,每条边被表示为一个引用关系。

(5) Twitch[19],社交网络图 Twitch,每个节点代表一个 Twitch 用户,每条边代表两个用户的关注。

(6) Twitter[23],社交网络图 Twitter,每个节点

代表一个 Twitter 用户,每条边代表两个用户的关注。

表 2　试验数据集
Table 2　Experimental dataset

数据集	点集数量	边集数量
Amazon	410 236	3 356 824
DBLP	317 080	1 049 866
Mico	100 000	1 080 298
Patent	2 745 761	13 965 409
Twitch	168 114	6 797 557
Twitter	81 306	1 768 149

4.2　评价指标

选用以下 4 个评价指标,前 3 个指标用于评估拓扑排序和输出模式排名间相似程度,第 4 个指标用于衡量预测结果和真实结果高排名模式准确性。

(1) Spearman 等级相关系数 ρ[24]:用于衡量两个变量排名之间单调关系。假设有两组数据 X_i、Y_i,其中 $i=1,2,\cdots,n$,对于每个值 X_i、Y_i,可以匹配等级 $R(X_i)$、$R(Y_i)$,那么 Spearman 等级相关系数表达式为

$$\rho=1-\frac{6\sum d_i^2}{n(n^2-1)},\tag{4}$$

式中:$d_i=R(X_i)-R(Y_i)$,为 X_i 和 Y_i 之间等级差;ρ 位于[-1,1],大于 0 为正相关,小于 0 为负相关。ρ 越接近于 1,代表两个序列之间正相关性越大。

(2) Kendall 等级相关系数 τ[24]:用于衡量两个变量排名之间单调关联程度。

$$\tau=\frac{O_s-O_d}{n(n-1)/2},\tag{5}$$

式中,O_s 为顺序关系相同元素对数,O_d 为顺序关系不同元素对数。τ 位于[-1,1],大于 0 为正相关,小于 0 为负相关。τ 越接近于 1,代表两个序列之间正相关性越大。

(3)余弦相似度 $S_{\cos}$[25]:用于衡量两个变量在方向上相似程度。通过计算两个变量夹角的余弦值来确定它们之间相似性。

$$S_{\cos}=\frac{A\cdot B}{|A|\times|B|}=\frac{\sum_{i=1}^{n}(A_i\times B_i)}{\sqrt{\sum_{i=1}^{n}A_i^2}\times\sqrt{\sum_{i=1}^{n}B_i^2}},\tag{6}$$

式中,$A\cdot B$ 为 A 和 B 点积,$|A|$ 和 $|B|$ 为变量长度。$S_{\cos}$ 越接近 1,表明两个变量在方向上相似度越高。

(4)准确率 A_c:用于衡量预测结果和真实结果高排名模式的准确性。

$$A_c=\frac{|\text{top}^C(k)\cap\text{top}^D(k)|}{k},\tag{7}$$

式中,$\text{top}^C(k)$ 为 GPEM 预测得到的排名前 k 个模式集合,$\text{top}^D(k)$ 为用户实际偏好的排名前 k 个模式集合,$|\text{top}^C(k)\cap\text{top}^D(k)|$ 为 $\text{top}^C(k)$ 和 $\text{top}^D(k)$ 交集的个数。

4.3　对比模型与对比算法

4.3.1　对比模型

(1) MLP 模型:为单塔结构,作为 GPEM 对比模型,超参数设置、相应网络层数与 GPEM 相同,输出偏好分数,用均方误差损失函数衡量预测值与实际值间差异。

(2) CNN+MLP 模型:为 CNN 和 MLP 结合的单塔结构,同为 GPEM 对比模型,超参数设置、相应网络层数与 GPEM 相同,输出偏好分数,用均方误差损失函数衡量预测值与实际值间差异。

(3) 孪生 MLP 模型(siamese MLP, SMLP):为消融实验模型,通过去除 CNN 模块并采用孪生 MLP 模块构建主观评价模型,分析 CNN 模块对模型性能贡献。

4.3.2　对比算法

(1) ItrMiner 算法[10]:综合考虑模式大小和支持度挖掘出 k 个频繁模式。

(2) AprTop$k$ 算法[2]:以模式大小为指标衡量每个模式趣味性,从图 G 中识别前 k 个有趣模式。

(3) BMiner 算法[14]:用支持度作为指标评估模式,从图 G 中识别出前 B 个最频繁模式。

(4) PT4AL-RF 算法[26]:PT4AL 是基于自监督的主动学习方法,用于高效选择最具信息量模式进行标注。PT4AL-RF 在其基础上加入随机森林模型,实现模式主观偏好预测。

4.4　试验结果

4.4.1　主观模型效果评估

数据输入时,按 7:3比例将数据集划分为训练集和测试集。依据用户偏好分别对训练集和测试集构建偏序关系,形成偏序对,将偏序对作为孪生网络结构输入,通过评估测试集偏序对准确率衡量孪生模型性能。评估 GPEM 时,将测试集输入模型,输出偏好分数,并基于这些偏好分数生成模型预测排序;对测试集偏序对进行拓扑排序以获得实际排序;计算预测排序与实际排序间相关性,以及高排名模式准确性,评估 GPEM 模型效果。对比模型与 GPEM 划分训练集、测试集方式一致。因对比模型均采用单塔结构,需将训练集和测试集偏序对进行拓扑排序,以生成向量排序序列。训练集排名经归一化处理,映射到[0,1]范围内分

数。训练时,拓扑排序后的训练集模式和相应归一化分数作为两个对比模型输入,测试集输入为测试集模式。

MLP、CNN+MLP、GPEM 模型在 6 个真实数据集上的试验结果如表 3 所示,加粗数据表示各模型在不同指标下的最优性能。由表 3 可以发现,GPEM 在 6 个数据集上均优于其他模型。相较 MLP 模型,GPEM 在 Kendall 等级相关系数上平均提高了 14%,GPEM 在所有数据集上余弦相似度均达到 99%以上。

各模型在 6 个真实数据集上的准确率试验结果如图 8 所示,预测模式数量比例 p 从 0.05 开始,每次递增 0.05,直至 0.20 结束。由图 8 可以发现,GPEM 除在 DBLP 数据集中 $p=0.20$ 时准确率略低于 CNN+MLP 模型外,其余情况表现均优于对比模型。在 Mico 和 Twitch 数据集上,GPEM 准确率呈现先下降后上升趋势,原因在于当数据集选取比例较小时,样本数量不足以充分训练模型,导致模型在小样本阶段的泛化能力较弱,准确率短暂下降;随着样本量增加,模型能更好地学习数据特征,准确率随之提升。这表明,GPEM 在小样本条件下可能受数据量不足限制,但在样本量充足情况下,其性能显著提高,表现出更强的泛化能力。

表 3 主观模型评价结果评估

Table 3 Assessment of subjective model evaluation results

数据集	模型	ρ	τ	$S_{\cos}$
Amazon	MLP	0.927 7	0.771 9	0.942 9
	CNN+MLP	0.950 5	0.805 5	0.967 6
	GPEM	**0.989 4**	**0.910 1**	**0.997 4**
DBLP	MLP	0.948 4	0.794 2	0.967 1
	CNN+MLP	0.963 0	0.859 2	0.973 2
	GPEM	**0.987 2**	**0.900 1**	**0.996 8**
Mico	MLP	0.915 9	0.745 0	0.939 0
	CNN+MLP	0.937 6	0.792 1	0.964 4
	GPEM	**0.987 1**	**0.903 2**	**0.995 8**
Patent	MLP	0.948 2	0.804 3	0.957 1
	CNN+MLP	0.960 3	0.818 6	0.970 1
	GPEM	**0.987 3**	**0.901 0**	**0.996 9**
Twitch	MLP	0.903 9	0.722 4	0.936 1
	CNN+MLP	0.932 2	0.768 6	0.962 2
	GPEM	**0.986 7**	**0.901 5**	**0.996 3**
Twitter	MLP	0.898 6	0.728 6	0.944 7
	CNN+MLP	0.968 7	0.841 8	0.972 2
	GPEM	**0.984 3**	**0.890 9**	**0.996 1**

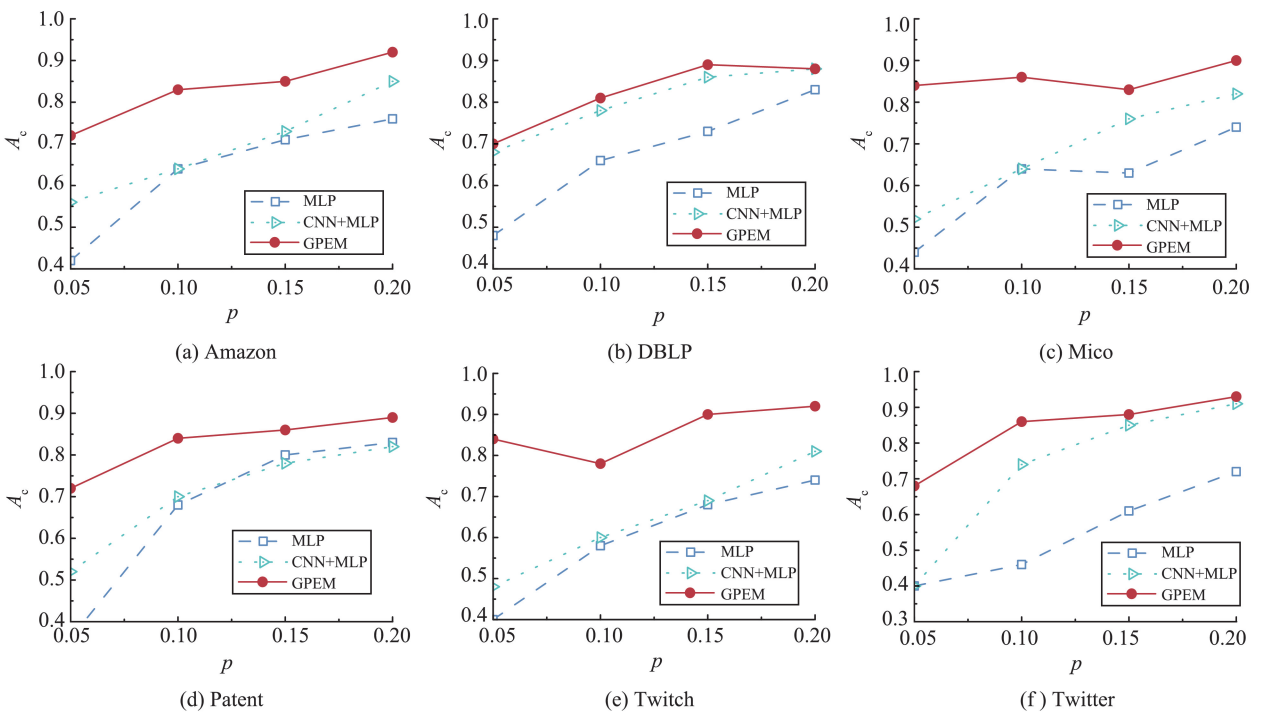


图 8 主观模型评价准确率

Fig.8 Subjective model evaluation accuracy

4.4.2 消融试验

进行消融试验评估 GPEM 中各模块有效性,试验在相同数据集上进行,样本大小、训练轮次参数保持一致。SMLP、GPEM 在 6 个数据集上试验结果如表 4 所示。由表 4 可知,GPEM 在 6 个数据集上的三个评价指标均略高于 SMLP。这主要是因为

CNN 模块增强了模型特征提取能力,提高了模型预测准确率。

表 4　消融试验结果评估

Table 4　Evaluation of ablation experiment results

数据集	模型	ρ	τ	$S_{\cos}$
Amazon	SMLP	0.978 2	0.864 9	0.989 9
	GPEM	0.989 4	0.910 1	0.997 4
DBLP	SMLP	0.964 9	0.855 3	0.984 0
	GPEM	0.987 2	0.900 1	0.996 8
Mico	SMLP	0.979 8	0.878 8	0.993 7
	GPEM	0.987 1	0.903 2	0.994 9
Patent	SMLP	0.967 2	0.865 5	0.991 8
	GPEM	0.987 3	0.901 0	0.996 9
Twitch	SMLP	0.965 3	0.850 5	0.983 8
	GPEM	0.986 7	0.901 5	0.996 3
Twitter	SMLP	0.977 2	0.875 0	0.994 3
	GPEM	0.984 3	0.890 9	0.996 1

2 种模型在 6 个真实数据集上准确率试验结果如图 9 所示,预测模式数量比例 p 从 0.05 开始,每次递增 0.05,直至 0.20 结束。由图 9 可以发现,GPEM 总体表现较好,在 Amazon、Mico、Patent、Twitter 数据集上均优于 SMLP 模型。在 DBLP 和 Twitch 数据集上,部分情况下 GPEM 准确率低于 SMLP 模型。这是由于在数据集特定采样比例下,模型受样本数量不足影响,无法充分捕捉偏好模式结构和标签信息,模型特征提取优势未能充分发挥。

前述试验展现仅考虑主观评价指标下的消融试验结果。进一步地,在主客观评价指标下,针对 Top-k 挖掘任务展开消融试验。先用 ItrMiner 算法获取模式集合,再根据公式[10]得模式客观评价分,用 GPEM 和 SMLP 分别得模式主观分,最后挖掘 Top-k 频繁模式。设定 α 为 0.5,p 从 0.05 起,每次增加 0.05,直至 0.20。6 个真实图数据上对比结果如图 10 所示,GPEM 准确率均高于 SMLP,得益于 CNN 模块特征提取能力。

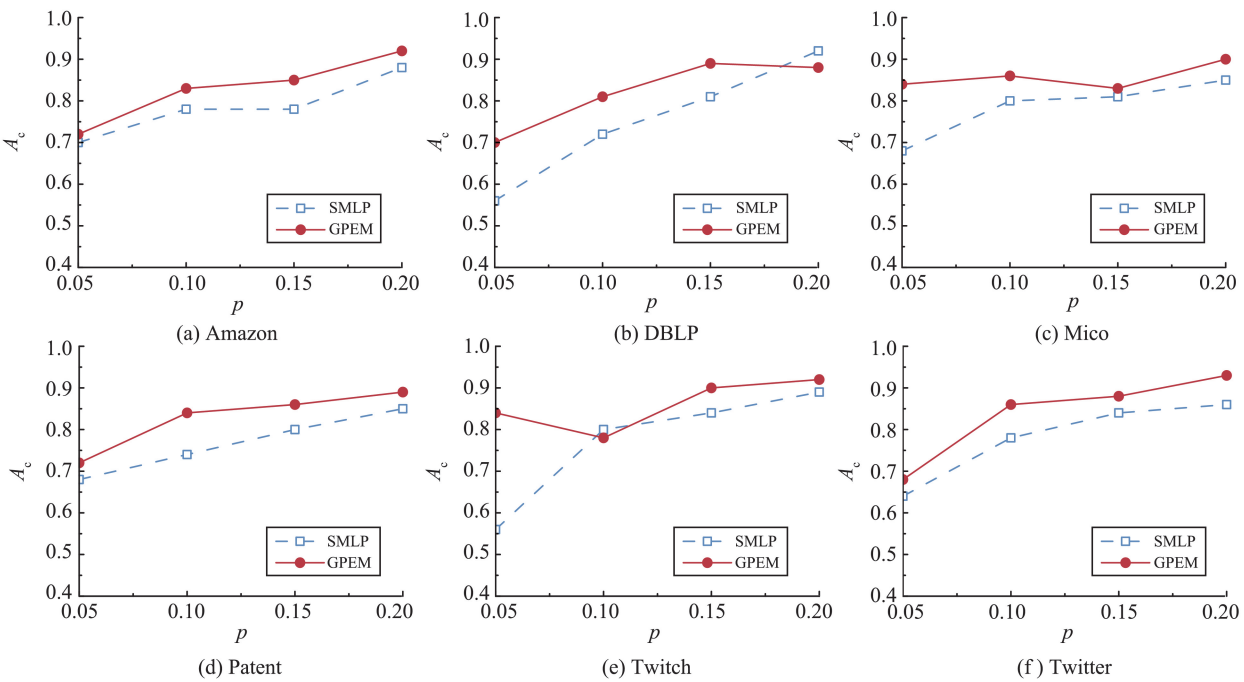
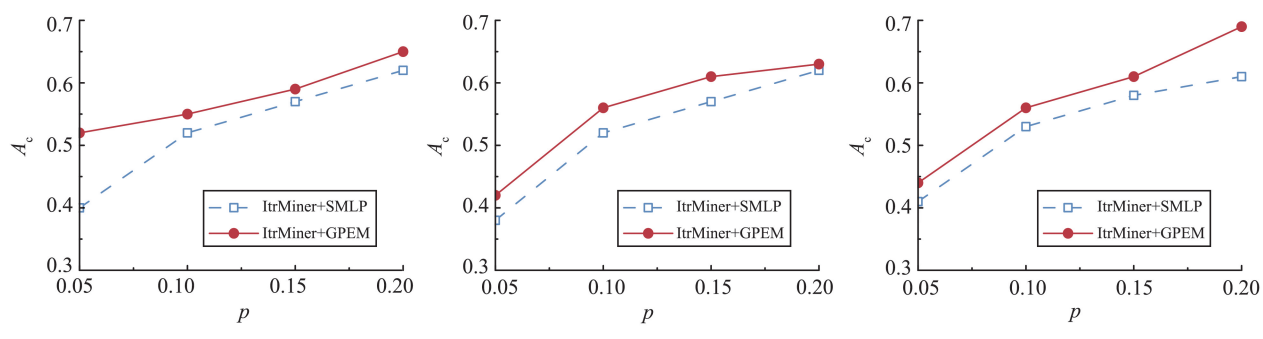
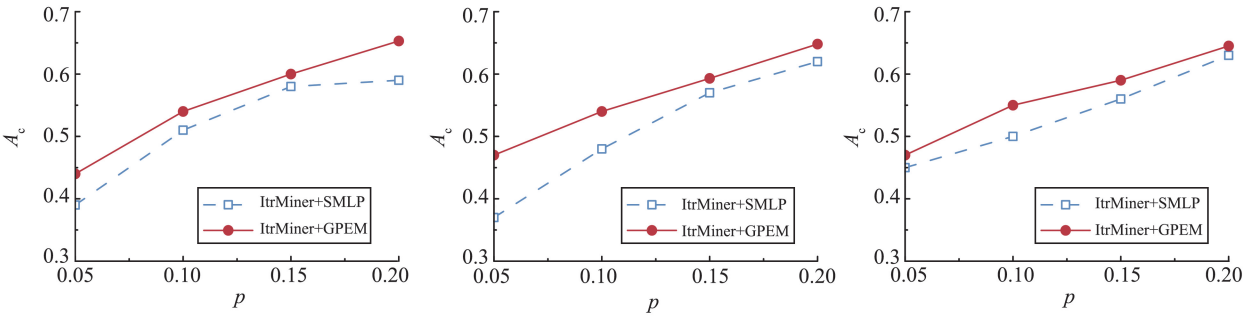


图 9　消融试验准确率

Fig.9　Accuracy of ablation experiments





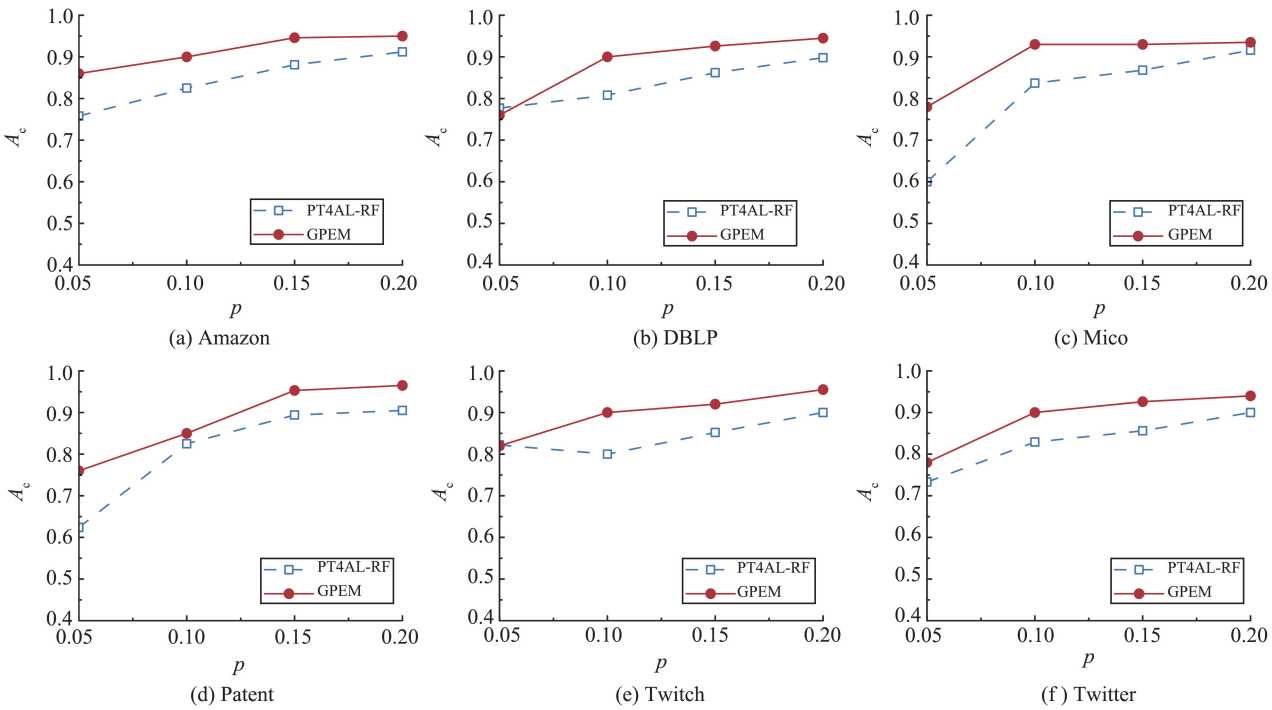
(d) Patent (e) Twitch (f) Twitter

图 10 $\alpha=0.5$ 时融合主客观评价的消融试验准确率

Fig.10 The accuracy of the ablation experiment integrating subjective and objective evaluation at $\alpha=0.5$

4.4.3 主观评价方法对比

对比 GPEM 与 PT4AL-RF[26] 的主观评价效果。在 6 个真实数据集上,预测模式数量比例 p 从 0.05 起,每次增 0.05 至 0.20,两种方法准确率变化情况如图 11 所示。由图 11 可见,GPEM 主观偏好预测性能总体更优,仅 DBLP 数据集 $p=0.05$ 时,因样本不足其准确率低于 PT4AL-RF,随着样本数量增加,准确率不断提升。



(a) Amazon (b) DBLP (c) Mico

(d) Patent (e) Twitch (f) Twitter

图 11 主观方法评价准确率

Fig.11 Subjective method evaluation accuracy

4.4.4 模型稳定性

利用 3 种挖掘算法 ItrMiner、AprTopK、BMiner,得到 3 个相同规模模式集合。根据公式[10]得模式客观分数,分别与 GPEM 结合得主观分,再经兴趣度函数 I 融合主客观分数排名,评估模型进行 Top-k 频繁模式挖掘稳定性。主观分数权重 α 设为 0.5,使主客观权重均等,在 6 个真实数据集上,以 100 为增量,模式数量 k 从 100 增至 500,测试 3 种方法准确率变化情况,试验结果如图 12 所示。由图可见,随着图规模增加,所有算法准确率上升,模型稳定性较好。

4.4.5 α 对模式评估的影响

本试验利用兴趣度指标函数对模式进行评估,探究参数 α 对兴趣度的影响及图规模增大时准确率变化情况。先将 k 值分别设为 100、200、300、400、500,在 6 个真实数据集上,α 以 0.1 步长从 0.6 增至 1.0,测试其对 GPEM 模式评估的影响,试验结果如图 13 所示。由图可见:GPEM 准确率随 α 增大呈上升趋势,因 α 越大,兴趣度在主观偏好上收益越大,模型更倾向于主观性强的模式。

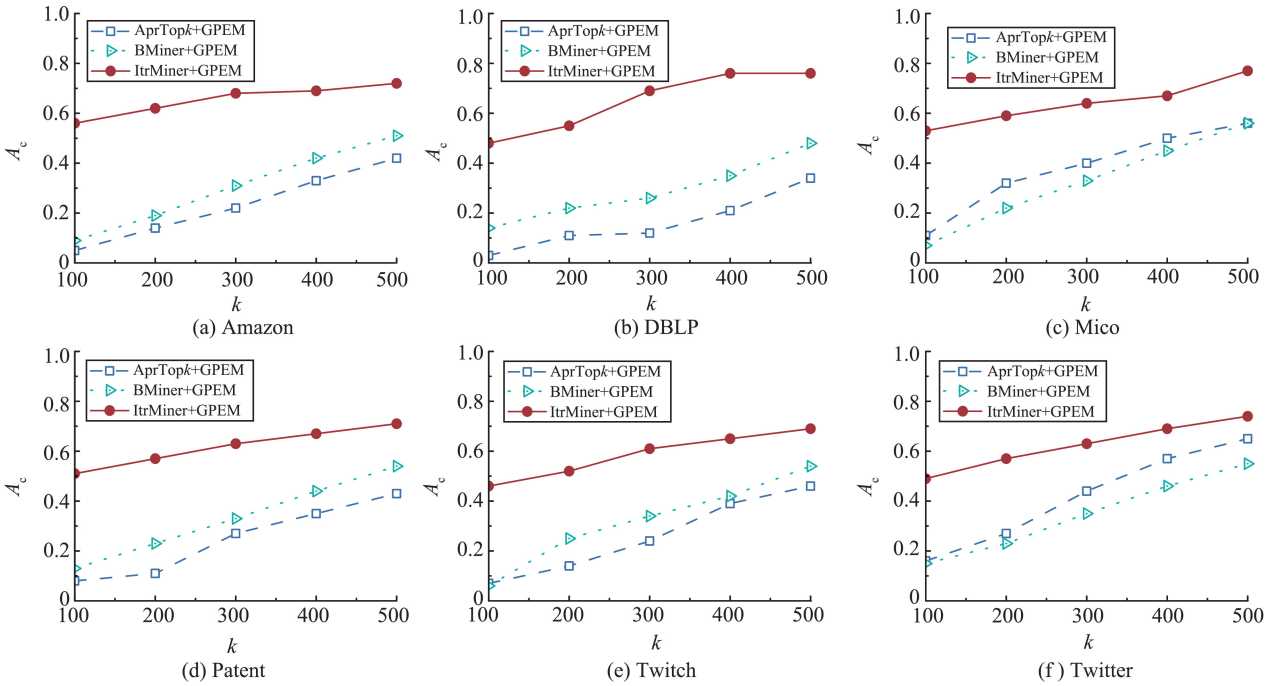


图 12　$\alpha=0.5$ 时不同客观挖掘算法下模型稳定性测试

Fig.12　Model stability test under different objective mining algorithms at $\alpha=0.5$

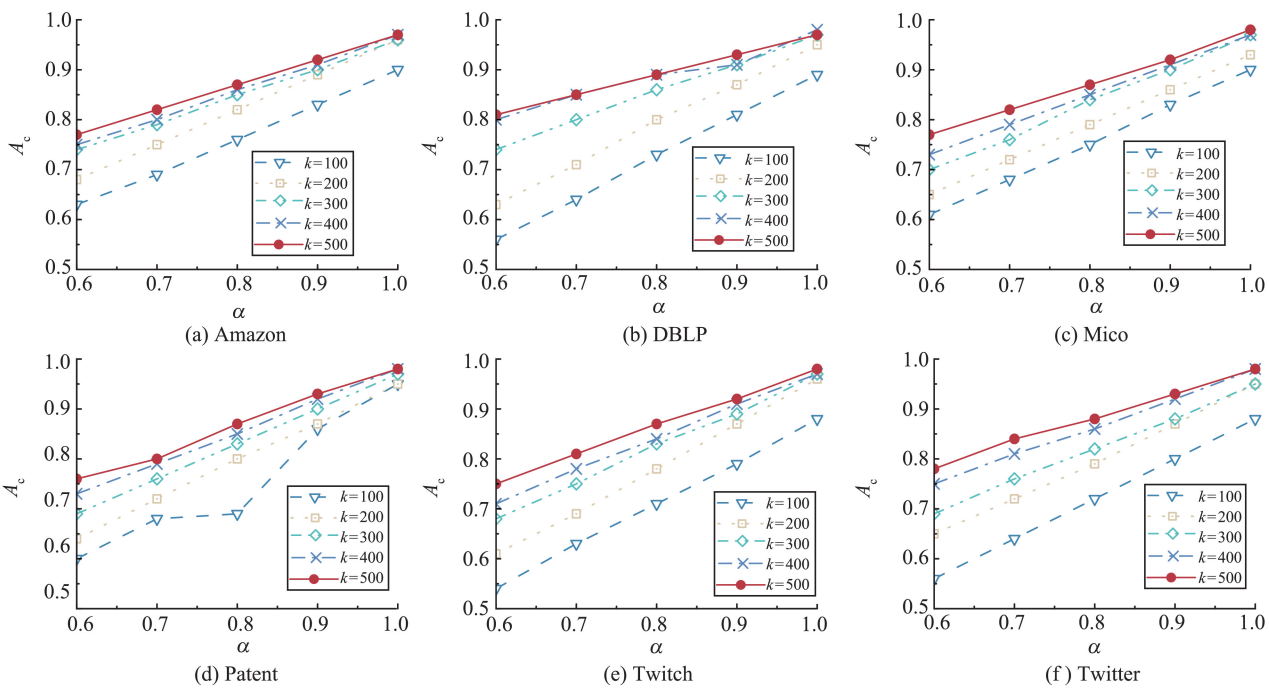


图 13　变化 α 对 GPEM 兴趣度函数准确率的影响

Fig.13　The effect of varying α on the accuracy of the GPEM model′s interest function

5　结论

针对频繁模式挖掘中传统客观评估方法无法反映用户主观兴趣的问题,本研究设计模式主观评价模型 GPEM,并使用真实图数据集验证 GPEM 性能。试验结果表明,GPEM 相较 MLP、CNN+MLP 模型,在多项指标上表现更优,Spearman 等级相关系数和 kendall 等级相关系数更大,余弦相似度和准确率更高。本研究提出同时考虑用户主观偏好和模式客观信息的兴趣度度量指标,旨在挖掘主客观融合的 Top-k 频繁模式。此外,引入 BMiner 和 AprTop$k$ 算法,验证了 GPEM 在兴趣度指标上的稳定性。

参考文献:

[1] INGALALLI V, IENCO D, PONCELET P. Mining frequent subgraphs in multigraphs [J]. Information Sciences, 2018, 451: 50-66.

[2] WANG X, LAN Z, HE Y A, et al. A cost-effective approach for mining near-optimal Top-*k* patterns [J]. Expert Systems with Applications, 2022, 202: 117262.

[3] PENG H, ZHANG D F. CFGM: an algorithm for closed frequent graph patterns mining[J]. Information Sciences, 2023, 625: 327-341.

[4] ZENG J, U L H, YAN X, et al. Fast core-based Top-*k* frequent pattern discovery in knowledge graphs[C]//2021 IEEE 37th International Conference on Data Engineering (ICDE). Chania, Greece: IEEE, 2021: 936-947.

[5] WANG X, XIANG M Y, ZHAN H Y, et al. Distributed Top-*k* pattern mining[M]// Cham: Springer International Publishing, 2021: 203-220.

[6] LE T, VO B, HUYNH V N, et al. Mining Top-*k* frequent patterns from uncertain databases [J]. Applied Intelligence, 2020, 50(5): 1487-1497.

[7] NATARAJAN D, RANU S. A scalable and generic framework to mine Top-*k* representative subgraph patterns [C]//2016 IEEE 16th International Conference on Data Mining (ICDM). Barcelona, Spain: IEEE, 2016: 370-379.

[8] SEMERTZIDIS K, PITOURA E. Top-*k* durable graph pattern queries on temporal graphs[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(1): 181-194.

[9] PRATEEK A, KHAN A, GOYAL A, et al. Mining Top-*k* pairs of correlated subgraphs in a large network [J]. Proceedings of the VLDB Endowment, 2020, 13(9): 1511-1524.

[10] 邹杰军, 王欣, 石俊豪, 等. 面向大图的 Top-Rank-K 频繁模式挖掘算法[J].南京大学学报(自然科学版), 2024, 60(1): 38-52.
ZOU Jiejun, WANG Xin, SHI Junhao, et al. Top-Rank-K frequent pattern mining algorithm for large graphs[J]. Journal of Nanjing University(Natural Science), 2024, 60(1): 38-52.

[11] BELMECHERI N, ARIBI N, LAZAAR N, et al. Boosting the learning for ranking patterns[J]. Algorithms, 2023, 16(5): 218.

[12] DAVASHI R. ITUFP: a fast method for interactive mining of Top-*k* frequent patterns from uncertain data [J]. Expert Systems with Applications, 2023, 214: 119156.

[13] LEHEMBRE E, CREMILLEUX B, ZIMMERMANN A, et al. WaveLSea: helping experts interactively explore pattern mining search spaces [J]. Data Mining and Knowledge Discovery, 2024, 38(4): 2403-2439.

[14] WANG X, SHI J H, ZOU J J, et al. Supports estimation via graph sampling[J]. Expert Systems with Applications, 2024, 240: 122554.

[15] FIRMANSYAH F, NURDIAWAN O. Penerapan data mining menggunakan algoritma frequent pattern-growth untuk menentukan pola pembelian produk chemicals[J]. Jurnal Mahasiswa Teknik Informatika, 2023, 7(1): 547-551.

[16] YAN X F, HAN J W. gSpan: graph-based substructure pattern mining[C]//2002 IEEE International Conference on Data Mining, 2002 Proceedings. Maebashi City, Japan: IEEE, 2002: 721-724.

[17] LI Y K, WU Z Y, LIN S, et al. Walking with perception: efficient random walk sampling via common neighbor awareness [C]//2019 IEEE 35th International Conference on Data Engineering (ICDE). Macao, China: IEEE, 2019: 962-973.

[18] YE S J, WANG Z, XIONG P B, et al. Multi-stage few-shot micro-defect detection of patterned OLED panel using defect inpainting and multi-scale Siamese neural network[J]. Journal of Intelligent Manufacturing, 2024, 35(6): 2653-2669.

[19] ROZEMBERCZKI B, ALLEN C, SARKAR R, et al. Multi-Scale attributed node embedding [J]. Journal of Complex Networks, 2021, 9(1): 1-22.

[20] YANG J, LESKOVEC J. Defining and evaluating network communities based on ground-truth [J]. Knowledge and Information Systems, 2015, 42(1): 181-213.

[21] ELSEIDY M, ABDELHAMID E, SKIADOPOULOS S, et al. GraMi: frequent subgraph and pattern mining in a single large graph [J]. Proceedings of the VLDB Endowment, 2014, 7(7): 517-528.

[22] ABDELHAMID E, ABDELAZIZ I, KALNIS P, et al. ScaleMine: scalable parallel frequent subgraph mining in a single large graph [C]//SC′16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. Salt Lake City, USA: IEEE, 2016: 716-727.

[23] LESKOVEC J, MCAULEY J. Learning to discover social circles in ego networks[C]// Proceedings of the 26 International Conference on Neural Information Processing Systems. Nevade, USA: ACM, 2012: 539-547.

[24] KLIE J C, DE CASTILHO R E, GUREVYCH I. Analyzing dataset annotation quality management in the wild [J]. Computational Linguistics, 2024, 50(3): 817-866.

[25] SINGH R H, MAURYA S, TRIPATHI T, et al. Movie recommendation system using cosine similarity and KNN [J]. International Journal of Engineering and Advanced Technology, 2020, 9(5): 556-559.

[26] YI J S K, SEO M, PARK J, et al. PT4AL: using self-supervised pretext tasks for active learning [M]// Cham: Springer Nature Switzerland, 2022: 596-612.

（编辑：管昕昉）