

基于异构特征聚合的图神经网络算法设计

侯浩民¹, 陈昌奉², 郭建邦¹

(1. 吉首大学 计算机科学与工程学院; 2. 吉首大学 通信与电子工程学院, 湖南 吉首 416000)

摘要: 异构网络表示学习主要用于探索复杂网络的特征和性质, 从而提高网络的安全性和社交网络分析效率。然而, 现有对异构网络的研究容易忽略网络结构及其节点的互动效果, 导致异构内容对学习结果产生一定影响。因此, 提出一种异构特征聚合的深度神经网络模型 HAGNN。首先, 利用 RWR 机制提取异构邻节点; 其次, 采用 Bi-GRU 学习复杂的结构信息, 对样本的邻近节点进行特征聚合; 最后, 采用图上下文丢失法和小量梯度下降算法对模型进行端到端训练, 并加入注意力机制进行优化。在应用实验中, HAGNN 节点预测的 F1 值能够达到 0.831, 场地推荐的 Recall 值达到 0.631, 相比其他图神经网络的应用结果有明显提升。

关键词: 异构网络表示学习; 异构特征聚合; 图神经网络

DOI: 10.11907/rjdk.241571

中图分类号: TP391

文献标识码: A

开放科学(资源服务)标识码(OSID):

文章编号: 1672-7800(2025)008-0049-10



Design of Graph Neural Network Algorithm Based on Heterogeneous Feature Aggregation

HOU Haomin¹, CHEN Changfeng², GUO Jianbang¹

(1. School of Computer Science and Engineering, Jishou University;

2. School of Communication and Electronic Engineering, Jishou University, Jishou 416000, China)

Abstract: Heterogeneous network representation learning is important for exploring the features and properties of complex networks, thereby improving network security and the efficiency of social network analysis. However, existing research on heterogeneous networks tends to ignore the impact of network structure and its node heterogeneous content on the learning results, so this paper proposes a deep neural network algorithmic model HAGNN for heterogeneous feature aggregation. The method first extracts heterogeneous neighbouring nodes using the RWR mechanism; furthermore, Bi-GRU is used to learn complex structural information so as to perform feature aggregation on the neighbouring nodes of the sample; finally, the model was trained end-to-end using graph context loss method and small amount of gradient descent algorithm, and the attention mechanism was added for optimization. In the application experiment, the F1 value predicted by HAGNN in the node can reach 0.831, and the venue recommendation Recall value can reach 0.631, which is significantly improved compared with the relevant application results of other graph neural networks.

Key Words: heterogeneous network representation learning; heterogeneous feature aggregation; graph neural networks

0 引言

随着互联网和信息技术的迅猛发展,网络中产生了大量异构数据,因而异构网络(Heterogeneous Network, HN)

成为研究热点之一。异构网络由不同类型的节点和关系组成,存在于社交网络、推荐系统、生物信息学和知识图谱等各个领域。异构网络的复杂性和多样性给传统的数据分析与挖掘带来了挑战,传统方法往往难以直接处理异构网络数据。因此,如何从异构网络中提取有用信息成为一

收稿日期: 2024-07-02

扫描二维码阅读全文:



基金项目: 国家自然科学基金项目(62266019); 湖南省自然科学基金项目(2024JJ7412); 湖南省教育厅优秀青年项目(24B0496); 国家级大学生创新训练计划项目(202210531009)

作者简介: 侯浩民(2001-), 男, 吉首大学计算机科学与工程学院学生, 研究方向为异构网络与医学成像技术; 陈昌奉(1993-), 男, 博士, 吉首大学通信与电子工程学院讲师, 研究方向为图神经网络、自然语言处理、计算机辅助医学诊断; 郭建邦(2001-), 男, 吉首大学计算机科学与工程学院学生, 研究方向为异构神经网络。本文通讯作者: 陈昌奉。

项重要的研究任务。

异构网络表示学习旨在将异构网络中的节点和关系映射到低维向量空间中,以便于后续的数据挖掘与分析任务。通过将异构网络中的节点和关系映射为连续向量,可以捕捉到节点之间的语义和结构信息,从而揭示异构网络中的隐藏模式和关联关系。异构网络表示学习的研究在社交网络分析、推荐系统、生物信息学和知识图谱等领域取得了显著进展,并得到了学术界和工业界的广泛关注。

近年来,随着深度学习和图神经网络的兴起,基于图神经网络的异构网络表示学习方法成为了研究热点。Schlichtkrull 等^[1]提出图卷积网络(Graph Convolutional Networks, GCN)模型,用于建模关系型数据,包括异构图数据。该模型能够捕捉节点之间的关系和局部结构信息,适用于异构图数据的表示学习和预测任务,但其对于高维节点特征的处理能力相对较弱。高俊艳等^[2]提出异构图注意力网络模型(Heterogeneous Graph Attention Network, HGAN),使用注意力机制建模异构图中不同类型节点和边的关系。HGAN能够自适应地捕捉重要节点和边的信息,提高模型的表达能力和准确性,但其对于大规模异构图的扩展性有限。Vaswani 等^[3]引入 Transformer 模型的思想,提出 Heterogeneous Graph Transformer (HGT),在异构图数据上使用多头注意力机制进行节点建模和信息传递。其能够处理异构图数据中的复杂关系和属性信息,提高模型的灵活性,但其对于大规模异构图的计算复杂度较高。Verma 等^[4]提出 GraphMix 方法用于半监督学习中的图神经网络训练,通过数据增强和正则化技术提高模型的鲁棒性。其能提高模型对未标记数据的利用效率,减轻对标记数据的依赖性。但在标记数据较少的情况下,模型性能可能下降,而且对于大规模图数据的计算复杂度较高。Zhao 等^[5]提出基于元图的融合方法,用于 Web 服务推荐中的异构图数据融合,通过元图将不同类型的节点关联起来。其能够融合不同类型的信息,提高推荐的准确性和个性化程度。

Dong 等^[6]对异构网络(HN)的嵌入进行研究,并将其运用于节点或图的分类问题中。Li 等^[7]研究并解决了节点聚类方式相关问题。在传统方法中,矩阵分解技术常用于 HN 生成隐含特征^[8]。然而,分解大型矩阵的计算成本通常极高,且存在统计性能不佳的问题^[9]。为了解决以上问题,近年来提出了广受关注的异质网络嵌入方法,其中最核心的步骤是将输入空间映射到低维空间,同时保留异质结构和语义信息。

此外,Chen 等^[10]将各种 HN 嵌入方法进行对比。从技术组成的角度来看,HN 使用的模型主要有两类:浅层模型与深层模型。浅层模型随机初始化节点嵌入,之后通过优化一些精心设计的目标函数来学习节点嵌入,以保留异质结构和语义,而深层次的深度模型旨在使用深度神经网络(Deep Neural Networks, DNNs)从节点属性中学习嵌入或相

互作用的特征。其中的异质图神经网络(Heterogeneous Graph Neural Networks, HGNNs)表现突出,也是本文的研究重点。HN 嵌入技术在现实世界中的应用十分成功,包括 Hu 等^[11]构建的推荐系统、刘亚妹等^[12]提出的恶意软件检测系统等。

本文在文献分析的基础上,总结了异构图神经网络相关研究,发现其中仍存在如下问题:

(1)异构网络中心节点的嵌入可能受到关联程度较低的邻居影响而受损,同时因信息有限,边缘节点的嵌入也无法充分体现其邻居信息。

(2)在异构图中,节点可能具有多种不同类型的异质信息,如属性、图像等节点可能关联着不同类型的内容。在此情况下,传统的图神经网络方法可能无法有效地处理这些内容之间的深度互动。

(3)各种类型的邻居在异构图中会对嵌入节点产生不同程度的影响。然而,目前大多数图神经网络(Graph Neural Networks, GNNs)主要集中在同质图中,没有充分考虑节点类型对嵌入的影响。

针对上述问题,本文提出一种基于 GRU(Gated Recurrent Unit)和多头注意力机制的异构网络聚合神经网络模型(Graph Neural Network Algorithm Based on Heterogeneous Feature Aggregation, HAGNN)。首先,利用附有重启的随机游走机制,从各结点中提取固定大小且具有高度关联性的异构邻居,并将其按照不同的结点分类。其次,聚集不同邻近群体的内容嵌入,并将每一个簇群的影响综合起来,以得到最终的节点嵌入。最后,进行模型消融。

1 相关工作

1.1 异构图

异构图(Heterogeneous Graph, HG)通常包含了大量且类型多样的节点和边。相比同质图,异构图可以表示更复杂的关系。在异构图中,每个节点和边都有一个类型。例如,在知识图谱中,节点可以表示实体(如人、公司、城市等),边可以表示关系(如居住、工作等)。

异构图和同构图的一个重要区别是其组成。如图 1 所示,同构图由相同类型的节点组成,并由相同类型的边连接;异构图由不同类型的节点组成,并由不同类型的边连接。另一个区别是,与同质图相比,异质图可以代表更复杂和多样化的数据结构。

Cui 等^[13]进行了大量工作致力于同质图嵌入,主要基于深层模型并结合图的属性学习节点或边的嵌入。例如,Perozzi 等^[14]提出的 DeepWalk 算法结合了随机行走和跳格模型,Tang 等^[15]提出的 LINE(Large-scale Information Network Embedding)利用一阶和二阶相似性学习区分节点嵌入的方法,Wang 等^[16]提出的 SDNE(Structural Deep Network Embedding)使用深度自动编码器提取图的非线性

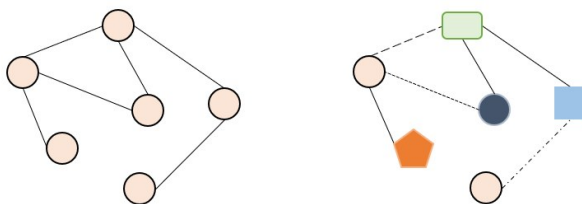


Fig. 1 Isomorphic and heterogeneous graph

图 1 同构图与异构图

特征。

1.2 DeepWalk 游走策略

随机游走是指在图上按照一定概率规则进行的连续行为。一个随机游走 W 由节点序列 $V_1, V_2, V_3, \dots, V_n$ 构成,表示从节点 V_1 开始,沿着边 (V_1, V_2) 到达节点 V_2 。以此类推,最终到达节点 V_n 。随机游走的长度定义为节点序列的长度 n 。

Zhang 等^[17]发现随机游走过会在图中产生一些节点序列。因此,该方法被广泛应用于图的表达学习中,以抽查节点的相邻关系,并捕捉局部图中的结构。从当前节点转移到下一节点的概率由转移概率矩阵进行描述,具体来说, $P(i, j)$ 表示从节点 i 到节点 j 的转移概率,满足 $0 \leq P(i, j) \leq 1$ 和 $\sum P(i, j) = 1$ 。随机游走示意图如图 2 所示。

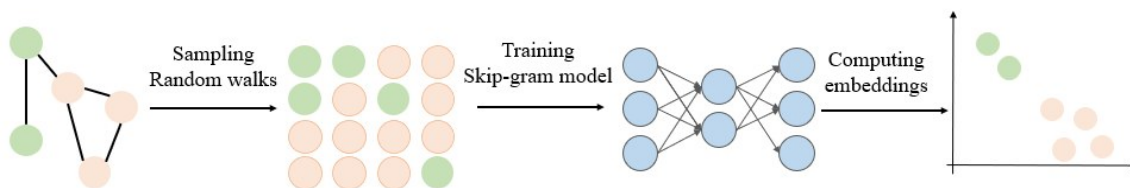


Fig. 2 Random walk diagram

图 2 随机游走示意图

在 HG 中,由于节点和边的类型限制,通常采用元路径引导的随机行走。因此,生成的节点序列不仅包含结构信息,而且包含语义信息。Dong 等^[6]发现,保留序列结构,节点嵌入可以同时保留一阶和高阶特性。同时,其提出 MetaPath2vec 算法,使用元路径引导的随机行走捕获多个节点的语义信息。

随机游走算法主要包括简单随机游走、附有重启的随机游走 (Random Walk Restart, RWR) 以及有偏随机游走。简单随机游走利用从当前节点转移到下一节点的概率与邻居节点个数成反比的特性,能够捕捉图的基本结构特性,但可能受到冗余路径和陷阱节点的影响,而 RWR 算法是为了解决简单游走问题而引入的方法。在 RWR 算法中,当前节点转移到下一节点的概率为 α ,返回初始节点的概率为 $1-\alpha$ 。该方法可以有效避免冗余路径和陷阱节点,提高随机游走的稳定性和收敛性。

1.3 图神经网络相关技术

图神经网络 (Graph Neural Network) 是一种能对图数据进行深度学习的方法。相较于线性表和树,图的结构更为复杂。在图结构中,节点之间的关系可以是任意的,图中任意两个数据元素之间都可能相关,使用图神经网络能更好地表达出蕴含在代码背后丰富的语义和结构信息。

作为普遍使用的图神经网络模型,GCN^[1]和 Graph-

SAGE^[18]常被用于完成基础图任务,GatedGraphConv^[19]应用门控神经网络 GRU^[20]的思想进行消息更新。异构图中不同特性的节点与边增加了图的复杂性。HetGNN^[21]尽量考虑了节点异构的内容,异构的节点与边采用 Bi-LSTM^[22]对无序列性质的内容进行编码。HAN (Heterogeneous Graph Attention Network) 根据不同元路径提取不同的语义信息为每个元路径分配权重,得到语义级注意力^[23]。本文主要对 LSTM 与 GRU 进行对比。

1.4 注意力机制

在图算法中,对于目标节点,邻居的重要性由其边缘权重决定。注意力机制 (Attention Mechanism) 在自然语言处理、计算机视觉等领域有着重要的应用价值,其核心思想是在处理序列数据时,可根据不同的上下文关系自适应地为每个输入赋予不同权重,从而使神经网络能够专注于关键信息,提高模型的表达能力和准确性。

注意力机制最早起源于视觉系统的研究,这种分布加权的处理方式能够在有限的计算资源下快速、有效地处理大量视觉信息。图注意力机制如图 3 所示。注意力机制主要分为自注意力和全局注意力两大类型。自注意力机制涉及在同一序列内部计算权重,而全局注意力机制则在两个不同序列之间计算权重。注意力机制的核心包括如下部分:

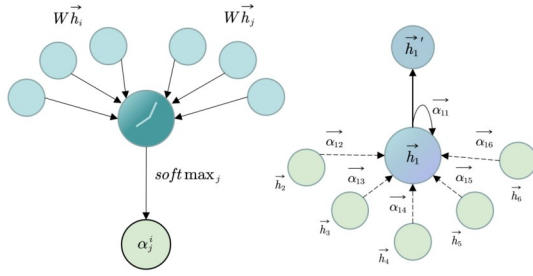


Fig. 3 Graph attention mechanism

图3 图注意力机制

(1)图注意层。图注意层解释了如何将 $k-1$ 层的隐藏节点表示 $H^{k-1} \in \mathbb{R}^{N \times F}$ 转移到新的节点表示 $H^k \in \mathbb{R}^{N \times F'}$ 。为了保证有足够的表达能力将低级节点表示转换为高级节点表示,对每个节点进行共享线性变换,记为 $W \in \mathbb{R}^{F \times F'}$ 。然后,在节点上转移自注意,通过共享注意力机制 $\alpha: \mathbb{R}^{F'} \times \mathbb{R}^{F'} \rightarrow \mathbb{R}$ 测量任意一对节点的注意系数。

$$e_{ij} = \alpha(WH_i^{k-1}, WH_j^{k-1}) \quad (1)$$

e_{ij} 表示节点 i 与节点 j 之间的关系强度,使用 H_i^{k-1} 表示列向量。对于每个节点,在理论上可以允许其关注图上的任意其他节点,但会忽略图的结构信息。在一般的应用中,通常只使用包括其自身的最近邻居。此外,通常采用Softmax函数对注意力系数进行归一化处理,以确保节点之间是可比的。

$$\alpha_{ij} = \text{Softmax}(\{e_{ij}\}) = \frac{\exp(e_{ij})}{\sum_{l \in N(i)} \exp(e_{il})} \quad (2)$$

(2)多头注意力。在实践中也可以使用多头注意力机

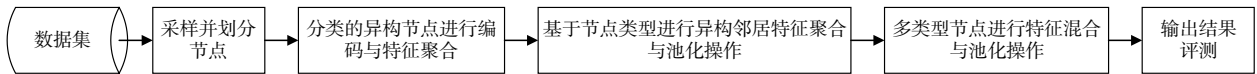


Fig. 4 Flow of HAGNN algorithm

图4 HAGNN算法流程

2.1 Word2Vec采样

大部分图神经网络的核心思路都是从一阶近邻网络中提取特征信息。如GraphSAGE^[24]或GAT^[25]。但在异构图中,此类方法往往存在无法获取不同区域特征、部分邻域未聚合、不同特征的异构域不适用性等问题。因此,本文设计了一个基于RWR的异质邻居采样策略,其包含两个连续的步骤:

(1)对一段时间内固定长度的RWR进行采样。首先从一个节点 v 开始,在该结点处进行随机行走,步行迭代到当前节点的邻居,或者以概率 p 返回。RWR运行至搜集任务完成则终止,将其表示为RWR(v)。在第一步中,本文着重介绍了Word2Vec在编码抽样中的功能,如图5所示。

Mikolov等^[26]提出CBOW、Skip-gram模型,其中CBOW适用于小体量数据集,而Skip-Gram在大型语料库中表现更佳。CBOW模型如图6左侧所示,其主要使用目标周围的其他单词,并将其作为单词向量的输入,通过加权处理

制。每个注意力决定了节点上不同的相似函数。多头注意力是一种同时利用多个注意力子空间的方法。对于每个注意力,可以独立地得到一个新的节点表示。具体公式如下:

$$H_i^k = \left\| \sum_{t=1}^T \sigma \left(\sum_{j \in N(i)} \alpha_{ij}^t W^t H_j^{k-1} \right) \right\| \quad (3)$$

其中, T 是注意头的总数, α_{ij}^t 是从第 t 个注意力头计算出的注意系数, W^t 是第 t 个注意头的线性变换矩阵。

2 基于异构特征聚合的神经网络算法

本文提出的基于GRU和多头注意力机制的异构网络聚合神经网络模型(HAGNN)是图神经网络拓展方式的一种,其核心思想是图嵌入提取特征,利用图注意力机制进行特征增强,最后由神经网络相关算法对图挖掘应用进行评估,并与其他模型进行对比。该模型能够有效表征代码的深层语义和结构信息。

HAGNN算法框架如图4所示。本文在相关异构模型基础上创新性地提出了游走算法与神经网络嵌入机制,并加入自注意力方法进行权重分配。HAGNN模型主要包括4个步骤:首先,对相关数据中的内容进行预处理采样;其次,利用神经网络对同邻居编码,并进行特征聚合与平均池化处理,强化深度语义和结构信息表达;再次,利用图神经网络实现特征矩阵聚合归一,得到特征组;最后,利用图注意力网络实现对结点权重的自适应分配。

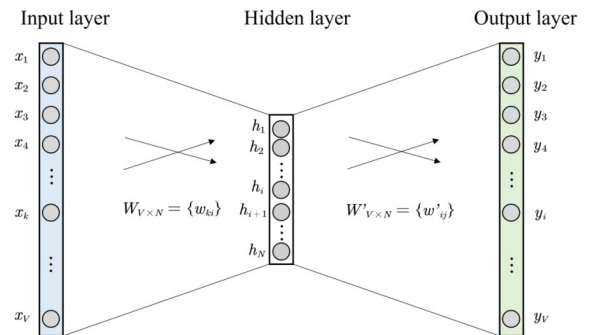


Fig. 5 Word2Vec training model

图5 Word2Vec训练模型

后,最后经过映射层输出目标。与CBOW不同,Skip-gram模型由单输入进行多词概率预测,如图6右侧所示。

(2)对不同的邻居类型进行分组。对于各种给定的节点类型,可以从RWR中选择排名前 k 的高频节点作为节点类型相关邻居集合。具体操作如下:①RWR对每个节点

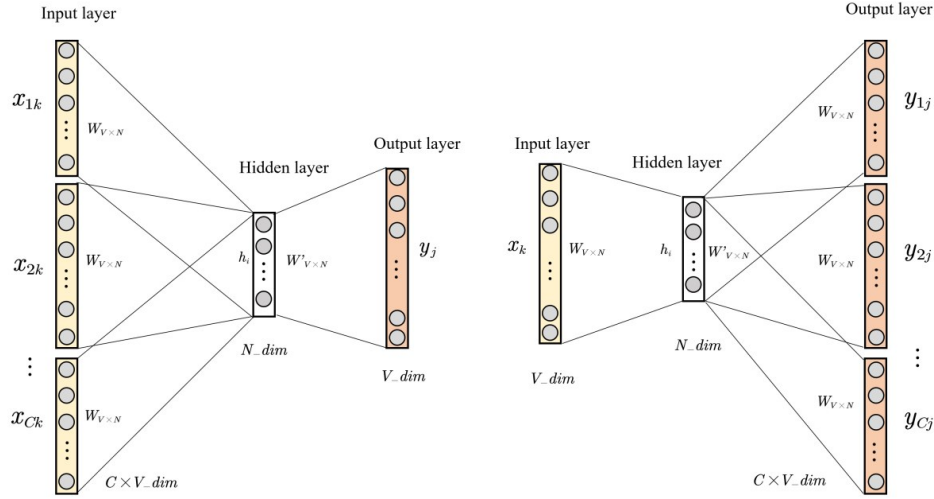


Fig. 6 Flow of two Word2Vec algorithms

图 6 Word2Vec 两种算法流程

各种类型的近邻进行采集;②对节点中取样的近邻数目进行计数,并选取被最频繁访问的近邻;③将同一类型(具有同一内容特性)的邻近节点进行分组,使得能够实现基于该类型的聚集。在此基础上,提出一种基于两层结构的网络模型,对网络中各节点采集到的异质性邻域信息进行汇总。

2.2 异构内容嵌入

为了解决编码问题,本文构建一种基于网络的非均匀性信息抽取模型,利用 f_1 实现对非均匀性信息的编码。具体来说,是将 C_v 中第 i 个内容的特征表示为 $x_i \in \mathbb{R}^{d_f \times l}$ 。

其中, d_f 是内容特征维度。 x_i 可以被事先训练,使用不同技巧来处理不同内容。例如,可以利用 Par2Vec^[27] 来预训练文本内容或采用 CNNs^[28] 对图像内容进行预训练。在以往的应用中,大多是将不同的内容特征直接进行连接,或将其线性地转换成一个统一的向量^[25]。与上述不同的是,本文利用一个基于循环神经网络的 GRU 新架构来捕捉深层特征的相互作用,并获得更强的表达能力。嵌入 v 内容的计算公式如下:

$$f_1(v) = \frac{\sum_{i \in C_v} [GRU\{\mathcal{F}C_{\theta_i}(x_i)\}]}{|C_v|} \quad (4)$$

式中, $\mathcal{F}C_{\theta_i}$ 表示能被确认的特征转换器,使用参数 θ_i 进行网络连接。长短时记忆网络(LSTM)和门控循环单元(GRU)在处理数据序列与捕捉长距离依赖方面相较于传统的 RNN 表现更优异。虽然这两种网络结构在许多任务中都取得了成功,但 GRU 具有一些突出优点,使其在某些情况下成为深度学习更合适的选择。GRU 具有以下优点:

(1) 参数数量。GRU 相较于 LSTM 的一个显著优点是其参数数量较少。GRU 仅有更新门和重置门两个门控单元,能够减轻计算负担,减少训练时间,使 GRU 在资源有限的条件下,处理较大规模数据集时更具优势。

(2) 训练速度。由于 GRU 的参数数量较少,其训练速度通常优于 LSTM,使得 GRU 在需要进行快速迭代和实验的项目中成为更理想的选择。此外,较少的参数也意味着简化的结构,有助于降低模型的复杂性,从而提高模型的可解释性。

(3) 适用性。虽然 LSTM 和 GRU 在处理长序列及捕捉长期依赖方面的性能相近,但 GRU 在某些任务上的表现可能优于 LSTM。

综上所述,GRU 由于具有参数较少、训练速度更快、结构简单、适用性广泛以及更容易调整等优点,使其在处理大规模数据集、快速迭代项目以及提高模型可解释性等方面更具有竞争力。然而,选择 LSTM 还是 GRU 仍取决于特定任务和性能的需求。

GRU 包括两部分:重置门用于控制先前瞬时隐藏层状态向当前目标隐藏层状态的更新情况;更新门用于控制前一时刻的隐藏层状态,确定其更新情况,并进入到激活函数中进行运算,使启动的结果被压缩到 0~1 之间。

具体而言,上述采用的架构首先使用不同的层转换不同的内容特征,接着采用 GRU 捕捉更深层特征的相互作用,并积累特征。此外,因为所需聚合的内容各不相同,本文使用不同的 GRU 组合计算公式来聚合不同类型节点的内容特征。

2.3 同型与异构节点聚合

为聚合每个节点异质邻居的内容嵌入,并解决各种类型的邻居在异构图中对嵌入节点影响不一致的问题,本文设计了另一个耦合模块。其包括两个连续的步骤:①聚集相同类型的邻居特征,生成网络;②使用相关算法组合不同类型结构的节点。

此前,基于 RWR 的策略,本文针对每个节点的不同类型,对邻居集进行固定后采样。因此,将 $v \in V$ 的 t 型抽样邻居集表示为 $N_t(v)$,之后使用神经网络 f_2 聚合 $v \in N_t(v)$ 的内容嵌入。 t 型抽样邻居集计算公式如下:

$$f_2(v) = \mathcal{AO}_{v \in N_t(v)}^t \{f_1(v)\} \quad (5)$$

其中, $f_1(v) \in R^{d \times t}$ 是来自 v 的内容嵌入函数, $\mathcal{AO}^t$ 是类型 t 的邻居集采样器, 能够充分链接神经网络。在该工作中, 使用 GRU 的效果更好。

HAGNN 算法使用不同的 GRU 区分不同节点类型, 并进行邻居聚合。受 GraphSAGE 的启发, GRU 在无序的邻居集上进行操作。

以上步骤为节点 v 生成 $|O_v|$ (图中节点类型的集合) 的聚合嵌入。在此基础上, 本文引入注意力机制, 使之能够更好地融合到 v 型的邻域嵌入中。主要原因在于, 对于 v

$$\alpha^{v,i} = \frac{\exp\{\text{LeakyReLU}(u^T[f_i \oplus f_1(v)])\}}{\sum_{f_j \in \mathcal{F}(v)} \exp\{\text{LeakyReLU}(u^T[f_j \oplus f_1(v)])\}} \quad (8)$$

其中, LeakyReLU 是整流线性单元激活函数, u 是注意力参数。在上述框架中, 为了实现嵌入维度的一致性和模型调整的便利性, 本文使用的输出嵌入维度相同。

为了进行异质图表示学习, 本文定义下列带有参数 Θ 的目标:

$$o = \underset{\Theta}{\operatorname{argmax}} x \prod_{v \in V} \prod_{t \in O_v} \prod_{v_c \in CN_v^t} p(v_c|v; \Theta) \quad (9)$$

其中, CN_v^t 是 v 的 t 型上下文节点集合, 邻居在图中以短随机游走的方式储存。条件概率 $p(v_c|v; \Theta)$ 被定义为异构网络的 Softmax 函数:

$$p(v_c|v; \Theta) = \frac{\exp\{\varepsilon_{v_c} \cdot \varepsilon_v\}}{\sum_{c_k \in V_t} \exp(\varepsilon_{v_k} \cdot \varepsilon_v)}, \quad \varepsilon_v = \mathcal{F}_{\Theta}(v) \quad (10)$$

其中, V_t 是图中 t 型节点的集合, ε_v 是指定的输出节点嵌入, 所有神经网络的参数 θ 均使用负采样技术来优化目标。具体来看, 通过负采样技术构建激活函数如下:

$$\log \sigma(\varepsilon_{v_c} \cdot \varepsilon_v) + \sum_{m=1}^M E_{v_c \sim P_t(v_c)} \log \sigma(-\varepsilon_{v_c} \cdot \varepsilon_v) \quad (11)$$

其中, M 是负样本量, $P_t(v_c)$ 是预先定义的噪声分布节点。在模型中定义 $M=1$, 因为当 $M>1$ 时, 其影响很小。因此, 公式可简写为交叉熵损失:

$$\log \sigma(\varepsilon_{v_c} \cdot \varepsilon_v) + \log \sigma(-\varepsilon_{v_c} \cdot \varepsilon_v) \quad (12)$$

具体来说, 首先在异质图中均匀地生成一组随机走动的 S 。然后, 对于 $S_i \in S$ 中的每个节点 v , 收集满足 $\text{dist}(v, v_c) \leq \tau$ 的上下文节点, 即与 v 在 S_i 中的距离为 τ 。最后, 利用 Adam 优化器对模型参数进行修正。通过反复学习, 直至两次迭代间的差异足够小, 利用所获得的模型参数推导出各类图挖掘任务的节点表达式。

的最后表现形式, 根据不同种类节点的邻近情况会产生不同的权重贡献。因此, 输出嵌入被表述为:

$$\varepsilon_v = \alpha^{v,v} f_1(v) + \sum_{t \in O_v} \alpha^{v,t} f_2(v) \quad (6)$$

其中, $\alpha^{v,*}$ 表示不同嵌入情况的重要性, f_1 、 f_2 分别是关于 v 的内容嵌入以及不同类型的聚合内容嵌入。输出的嵌入公式为:

$$\varepsilon_v = \sum_{f_i \in \mathcal{F}(v)} \alpha^{v,i} f_i \quad (7)$$

3 实验设计与分析

3.1 实验环境

为了验证 HAGNN 算法的可行性与有效性, 本文将其运用于各种图挖掘任务中, 如链接预测、个性化推荐、节点分类与聚类等。图神经网络(GNN)也在节点分类与聚类任务中发挥着重要作用。通过学习图中节点的嵌入表示来捕捉节点之间的复杂依赖关系和图结构信息, 从而提高分类和聚类性能。在 HAGNN 模型实现过程中, 本文使用高级语言 Python 编写代码, 同时采用框架 PyTorch 以更好地进行神经网络生成与模型训练。PyTorch 提供了强大的张量计算能力以及对动态计算图的支持, 使构建和训练神经网络变得更加简单、高效。考虑到所使用模型的复杂情况, 采用 PyTorch-GPU 版本进行操作, 以减少模型训练时间、提高计算效率。实验环境相关参数如表 1 所示。

Table 1 Experimental environment related parameters

表 1 实验环境相关参数

名称	参数
处理器	Intel(R) Xeon(R) CPU E5-2680 v4
图形处理器	NVIDIA 3090-24G
操作系统	Linux
开发语言	Python3.8
深度学习框架	PyTorch-GPU 1.11.0, Cuda 11.3.0
开发平台	Vscode

本文采用 Aminer 数据集, Aminer (ArnetMiner) 广泛应用于图挖掘、推荐系统、自然语言处理、知识图谱等领域, 其中主要包括论文数据集、作者数据集、会议和期刊数据集、机构数据集等。本文选用部分论文、作者与会议数据集, 并根据相关评测指标进行对比。首先, 实现 HAGNN 算法, 获取评测数据; 其次, 优化 HAGNN 细节参数, 对改进前后的指标进行对比; 最后, 与其他基准算法模型进行对比。

3.2 评估指标

(1) 损失张量 (Loss Tensor)。损失张量是表示损失函

数计算结果的一个张量,其通常是一个标量,表示模型在整个批次(batch)数据上的平均损失。在某些情况下,损失张量也可以是一个向量,表示模型在每个样本上的损失。本文采用交叉熵损失(Cross-Entropy Loss)计算方式:

$$L(Y|f(x)) = - \sum_{i=1}^N Y_i \log f(x_i) \quad (13)$$

在实际应用中,有时需要根据问题的特性自定义损失函数。损失函数的设计需要平衡各种因素,如易于优化、对不同错误类型的敏感性等。在训练深度学习模型时,损失张量作为优化目标,可通过梯度下降等优化算法更新模型参数,以最小化损失。

(2) AUC (Area Under Curve)。AUC 是指 ROC 曲线下方的区域,用于评估二分类模型在不同阈值条件下的表现。ROC 展示了模型在各个分类极限下的真阳性率(True Positive Rate, TPR)和假阳性率(False Positive Rate, FPR)。

AUC 取值介于 0~1 之间,该值越接近 1,说明模型的分类性能越好;反之,说明模型的分类性能较差。当 AUC 等于 0.5 时,说明模型的预测能力相当于随机猜测。在实际应用中,AUC 常用于评估不平衡数据集的分类性能。

(3) 准确率 (Accuracy)。准确率表示一个模型分类正确时,样本数占比较对象总数的比例。准确率计算公式如下:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

其中,TP(True Positive)表示本身为阳性,且预测为阳性的样本数;FP(False Positive)表示本身为阴性,但最终预测为阳性的样本数;TN(True Negative)表示本身为阴性,且预测为阴性的样本数;FN(False Negative)表示本身为阳性,但最终预测为阴性的样本数。

(4) 精确率 (Precision)。精确率表示最终被预测为阳性的样本中实际为阳性的比例。计算公式如下:

$$Precision = \frac{TP}{TP + FP} \quad (15)$$

(5) 召回率 (Recall)。召回率表示本身为阳性的样本中最终被预测为阳性的比例。计算公式如下:

$$Recall = \frac{TP}{TP + FN} \quad (16)$$

(6) F1 分数 (F1-score)。F1 分数是统计学中用来衡量二分类(或多任务二分类)模型精确度的一种指标。F1 分数越高,表示模型的精确率和召回率越高。F1 分数计算公式如下:

$$F1_score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (17)$$

在某些情况下,准确率可能并不是一个合适的评估指标,特别是在数据集不平衡的情况下。此时,精确率、召回

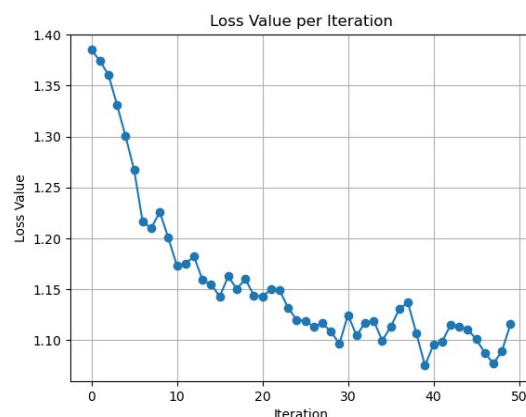
率以及 F1 分数可能更适合对模型性能进行评估,从而更好地理解模型在不同类别上的表现,以便对模型进行调整和优化。在 Python 中,可以使用 Scikit-learn 库的 precision_score、recall_score 和 F1_score 函数计算这些指标。

3.3 实验结果分析

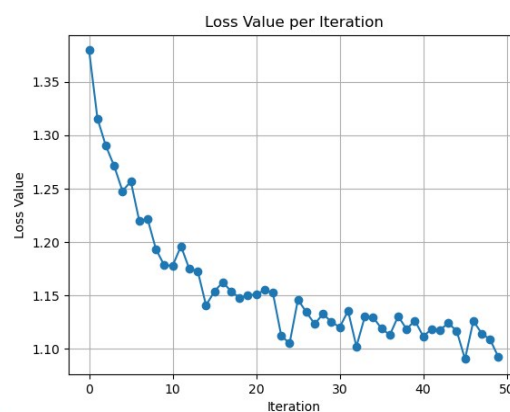
3.3.1 DeepWalk 优化对比

由上述分析可知,DeepWalk 算法具有两种计算模式。在实验过程中,首先估计运行 Skip-gram 算法与 CBOW 算法对整体模型产生的影响,以进一步优化模型。之后,对比 Skip-gram 算法与 CBOW 算法计算方式的区别,并针对应用进行评估。

图 7 对比了 50 次迭代中 Loss tensor 的下降情况,发现使用 CBOW 算法造成的精度损失更少,效果更佳。



(a) Skip-gram



(b) CBOW

Fig. 7 Comparison of Loss values of Skip-gram and CBOW algorithm

图 7 Skip-gram 与 CBOW 算法 Loss 值比较

链接预测作为一项重要的下游应用,被广泛应用于图挖掘中,用于对图中可能存在的边(关系)进行预测。其算法通常基于节点间的相似度度量,例如共享邻居、Jaccard 系数、Adamic-Adar 指数等。在链接预测指标评估中,本文进行了两方面的比较。一个是作者关联预测 (Author Collaboration Link Prediction, Ac),另一个是作者—论文预测

(Author Paper Citation Link Prediction, Ap)。评估指标选用 AUC、Accuracy、Precision、Recall 以及 F1 值。由表 2 的结果可知,使用 CBOW 计算模式的评测值明显高于 Skip-gram。

Table 2 DeepWalk link prediction results

表 2 DeepWalk 链接预测结果

算法	AUC	Accuracy	Precision	Recall	F1
Skip-gram (Ac)	0.657	0.657	0.790	0.360	0.512
CBOW (Ac)	0.682	0.682	0.870	0.431	0.574
Skip-gram (Ap)	0.723	0.723	0.813	0.548	0.665
CBOW (Ap)	0.731	0.731	0.820	0.566	0.679

个性化场地推荐根据使用者的历史行为、兴趣和需求,为用户提供定制化的内容。在图挖掘中,个性化推荐可看作是作者与项目之间的链接预测问题。在进行此项评估时选用 Recall、Precision 以及 F1 值进行比较。同时,为了更精确地显示结果,本文选取了 top3 和 top5 两个范围。通过表 3 的结果可知,CBOW 有着更好的性能。

Table 3 DeepWalk recommendation comparison

表 3 DeepWalk 推荐对比

算法	Rec@top3	Pre@top3	F1@top3	Pre@top5	F1@top5
Skip-gram	0.195	0.129	0.156	0.129	0.185
CBOW	0.197	0.132	0.173	0.159	0.187

3.3.2 网络层数及 LSTM 与 GRU 优化对比

考虑到神经网络层数以及不同构建方法会有不同的效果,本文选用长短期记忆递归神经网络(LSTM)以及门控循环单元(GRU)进行对比,试图探索出在 HAGNN 中效果更好的神经网络。同时,为了保证数据的精准度,对隐藏层中的网络深度进行了增添。算法 1 采用 LSTM 神经网络,层数为 5;算法 2 采用 GRU 神经网络,层数为 50。对比结果如图 8 所示。

由图 8 可知,迭代过程中算法 2 的 Loss tensor 值下降更为平稳,最终得到的值更小,表明构建的模型损失更少。算法 1 与算法 2 链接预测对比如表 4 所示,算法 2 的 AUC 值、Precision 值明显高于算法 1。因此,在 HAGNN 模型构建中采用算法 2。

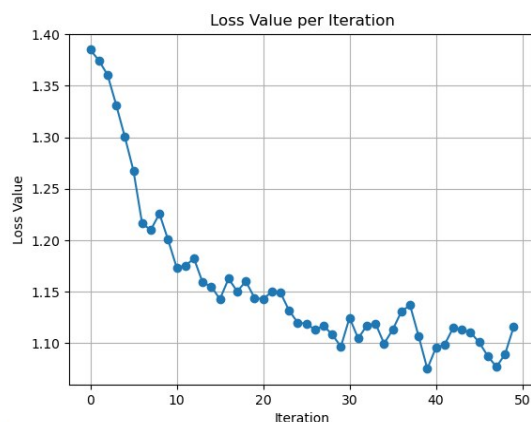
算法 1 与算法 2 场地推荐对比如表 5 所示,结果显示算法 2 也有着更好的性能。

3.3.3 激活函数优化对比

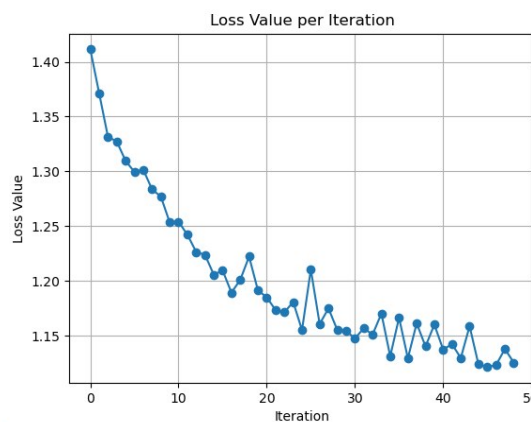
LeakyReLU 激活函数的应用非常广泛,常用于 CNN 以及深度前馈神经网络中,可替代标准的 ReLU 激活函数。LeakyReLU 在图像识别、自然语言处理、语音识别等任务中都具有良好表现。由于 LeakyReLU 是 ReLU (Rectified Linear Unit) 激活函数的一种变体,本文将使用 LeakyReLU 激活函数的算法命名为算法 3,使用 ReLU 激活函数的算法命名为算法 4,并将两者进行对比。

由图 9 可以看出,ReLU 的 Loss tensor 下降过程略优于 LeakyReLU。

由表 6 可知,不论是在 Author Collaboration Link Pre-



(a) Algorithm 1



(b) Algorithm 2

Fig. 8 Comparison of Loss values of algorithm 1 and algorithm 2

图 8 算法 1 与算法 2 的 Loss 值对比

Table 4 Link prediction comparison of algorithm 1 and algorithm 2

表 4 算法 1 与算法 2 链接预测对比

算法	AUC	Accuracy	Precision	Recall	F1
算法 1 (Ac)	0.657	0.657	0.790	0.360	0.512
算法 2 (Ac)	0.695	0.695	0.875	0.456	0.600
算法 1 (Ap)	0.723	0.723	0.813	0.548	0.665
算法 2 (Ap)	0.738	0.738	0.830	0.610	0.702

Table 5 Site recommendation comparison of algorithm 1 and algorithm 2

表 5 算法 1 与算法 2 场地推荐对比

算法	Rec@top3	Pre@top3	F1@top3	Pre@top5	F1@top5
算法 1	0.195	0.129	0.156	0.129	0.185
算法 2	0.201	0.140	0.165	0.140	0.195

diction 还是在 Author Paper Citation Link Prediction 中,算法 4 都有着更大的优势。

算法 3 与算法 4 场地推荐对比如表 7 所示,结果显示算法 4 的效果更好。

至此,本文完成了模型消融。在前文中对 Word2Vec 算法进行了选择,并优化了网络结构及深度,最后选用了合适的激活函数 ReLU。

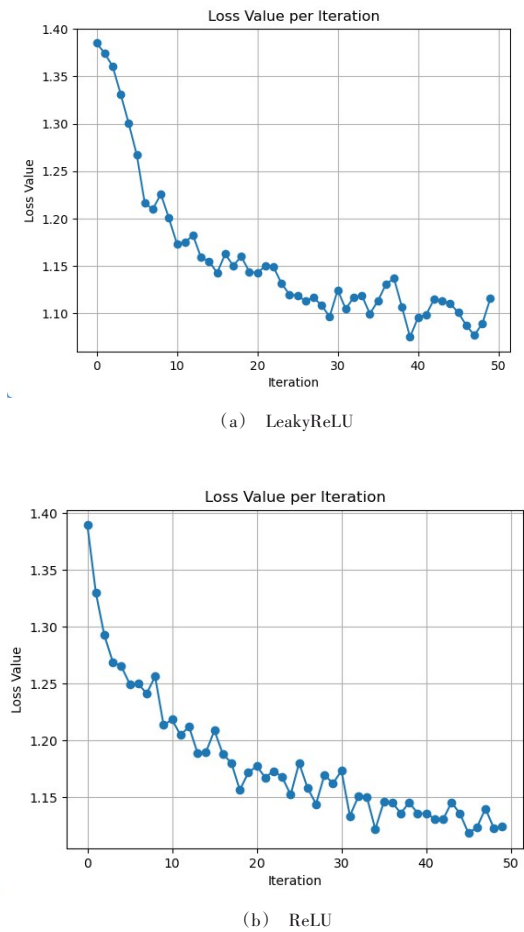


Fig. 9 Comparison of Loss values of LeakyReLU and ReLU

图9 LeakyReLU与ReLU的Loss值对比

Table 6 Link prediction comparison of algorithm 3 and algorithm 4

表6 算法3与算法4链接预测对比

算法	AUC	Accuracy	Precision	Recall	F1
算法3(Ac)	0.657	0.657	0.790	0.360	0.512
算法4(Ac)	0.687	0.687	0.866	0.434	0.579
算法3(Ap)	0.723	0.723	0.813	0.548	0.665
算法4(Ap)	0.725	0.725	0.815	0.584	0.681

Table 7 Site recommendation comparison of algorithm 3 and algorithm 4

表7 算法3与算法4场地推荐对比

算法	Rec@top3	Pre@top3	F1@top3	Pre@top5	F1@top5
算法3	0.195	0.129	0.156	0.129	0.185
算法4	0.210	0.133	0.133	0.161	0.188

3.3.4 模型对比

本文采用5个模型进行比较,包括异构图的嵌入模型MetaPath2vec、归属图的模型ASNE与SHNE、神经网络模型GraphAGE和GAT,以及较为流行的HetGNN模型^[29]。

MetaPath2vec作为一种异构图嵌入的典型模型,通过元路径引导的随机游走和Skip-gram模型学习节点嵌入;ASNE采用归属图嵌入方法,对节点的潜在特征进行归属后进一步学习嵌入;SHNE优化了图结构和文本语义的相

关性;GraphAGE通过对不同的神经网络进行融合,进而聚合邻居的特征信息;GAT作为常见的图注意力网络模型,采用自注意力机制对邻居节点信息进行聚合;HetGNN作为较为完善的异构图网络,初步融合了上述模型的优点。

不同模型预测比较如表8所示,发现应用了图嵌入或图神经网络算法的模型具有更理想的评测数值。HAGNN在两个主要的评测指标中效果最佳,说明所提出的异构聚合图神经网络模型是有效的。

Table 8 Prediction comparison of different models

表8 不同模型预测比较

指标	MP2V	ASNE	SHNE	GSAGE	GAT	HetGNN	HAGNN
AUC	0.596	0.689	0.683	0.695	0.678	0.717	0.738
F1	0.348	0.643	0.639	0.615	0.613	0.669	0.831

不同模型推荐比较如表9所示,发现HAGNN也具有较好的性能。

Table 9 Recommendation comparison of different models

表9 不同模型推荐比较

指标	MP2V	ASNE	SHNE	GSAGE	GAT	HetGNN	HAGNN
Rec	0.516	0.419	0.608	0.540	0.568	0.625	0.631
Pre	0.207	0.174	0.241	0.219	0.230	0.252	0.261
F1	0.295	0.333	0.345	0.312	0.327	0.359	0.374

4 结语

本文提出一个异构图的表示学习框架及相应实现方法,成功解决了传统推荐算法中非直接相邻的图节点信息组合问题,从而提升了推荐模型的性能。该研究为推荐系统的发展提供了新的思路和方法。在实际应用中,利用图结构化数据及复杂结构信息,能够更好地刻画用户偏好及商品属性,进而提高推荐精度。此外,用户行为偏好还受其周边邻居的影响,也即是说,用户行为偏好依赖于其本身的属性和情境。因此,建立使用者与情境环境的联合模型,也不失为一个行之有效的方法。

参考文献:

[1] SCHLICHTKRULL M S , CAO N D , TITOV I. Interpreting graph neural networks for NLP with differentiable edge masking[DB/OL]. <https://arxiv.org/abs/2010.00577>.

[2] GAO J Y , LIU W Y , YANG Z G. Combined with the feature of attention and fusion of target tracking [J]. Journal of Guangdong University of Technology, 2019, 36 (4): 18-23.

高俊艳,刘文印,杨振国. 结合注意力与特征融合的目标跟踪[J]. 广东工业大学学报, 2019, 36(4):18-23.

[3] VASWANI A , SHAZEER N , PARMAR N , et al. Attention is all you need [DB/OL]. <https://arxiv.org/abs/1706.03762>.

[4] VERMA V , QU M , KAWAGUCHI K , et al. Graphmix: improved training of GNNs for semi-supervised learning [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 10024-10032.

[5] ZHAO H , YAO Q , LI J , et al. Meta-graph based recommendation fusion over heterogeneous information networks [C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data

- Mining, 2017: 635–644.
- [6] DONG Y X, CHAWLA N V, SWAMI A. Metapath2vec: scalable representation learning for heterogeneous networks [DB/OL]. <https://doi.org/10.1145/3097983.3098036>.
- [7] LI X, KAO B, REN Z, et al. Spectral clustering in heterogeneous information networks [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 4221–4228.
- [8] NEWMAN M E J. Modularity and community structure in networks [J]. Proceedings of the National Academy of Sciences, 2006, 103(23): 8577–8582.
- [9] WANG S. User behavior pattern mining based on matrix decomposition and neural network [D]. Beijing: Beijing University of Posts and Telecommunications, 2017.
王珊. 基于矩阵分解和神经网络方法的用户行为模式挖掘 [D]. 北京: 北京邮电大学, 2017.
- [10] CHEN H, YIN H, WANG W, et al. PME: projected metric embedding on heterogeneous networks for link prediction [C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018: 1177–1186.
- [11] HU B, SHI C, ZHAO W X, et al. Leveraging meta-path based context for top-N recommendation with a neural co-attention model [C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018: 1531–1540.
- [12] LIU Y S, HOU Y R, YAN H B. Malicious code detection based on heterogeneous information network [J]. Journal of Beijing university of Aeronautics and Astronautics, 2022, 48(2): 258–265.
刘亚妹, 侯跃然, 严寒冰. 基于异质信息网络的恶意代码检测 [J]. 北京航空航天大学学报, 2022, 48(2): 258–265.
- [13] CUI P, WANG X, PEI J, ZHU W. A survey on network embedding [J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(5): 833–852.
- [14] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: online learning of social representations [DB/OL]. <https://arxiv.org/abs/1403.6652>.
- [15] TANG J, QU M, WANG M, et al. Line: large-scale information network embedding [DB/OL]. <https://arxiv.org/abs/1503.03578>.
- [16] WANG D, CUI P, ZHU W. Structural deep network embedding [C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 1225–1234.
- [17] ZHANG Y, PAL S, COATES M, et al. Bayesian graph convolutional neural networks for semi-supervised classification [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 5829–5836.
- [18] HAMILTON W L, YING R, LESKOVEC J. Inductive representation learning on large graphs [DB/OL]. <https://arxiv.org/abs/1706.02216>.
- [19] LI Y J, TARLOW D, BROCKSHMIDT M, et al. Gated graph sequence neural networks [DB/OL]. <https://arxiv.org/abs/1511.05493>.
- [20] CHUANG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [DB/OL]. <https://arxiv.org/abs/1412.3555>.
- [21] ZHANG C, SONG D, HUANG C, et al. Heterogeneous graph neural network [C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019: 793–803.
- [22] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735–1780.
- [23] WANG X, JI H, SHI C, et al. Heterogeneous graph attention network [C]//The World Wide Web Conference, 2019: 2022–2032.
- [24] HAMILTON W L, YING R, LESKOVEC J. Inductive representation learning on large graphs [DB/OL]. <https://arxiv.org/abs/1706.02216>.
- [25] VELIKOVI P, CUCURULL G, CASANOVA A, et al. Graph attention networks [DB/OL]. <https://arxiv.org/abs/1710.10903>.
- [26] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [DB/OL]. <https://arxiv.org/abs/1301.3781>.
- [27] KANG M J. Application of natural semantic analysis and machine learning to big data security [J]. Electronic Technology and Software Engineering, 2022(18): 202–207.
康缪建. 自然语义分析与机器学习在大数据安全中的应用 [J]. 电子技术与软件工程, 2022(18): 202–207.
- [28] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3431–3440.
- [29] ZHANG C, SONG D, HUANG C, et al. Heterogeneous graph neural network [C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019: 793–803.

(责任编辑: 黄健)