山西大学学报(自然科学版)47(5):993—1003,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2024106

利用邻域k元节点组信息的节点结构相似性判定方法

杨贵[1],韦兴宇[1],郑文萍[1,2,3]

(1.山西大学 计算机与信息技术学院,山西 太原 030006;
2.计算智能与中文信息处理教育部重点实验室(山西大学),山西 太原 030006;
3.山西大学 智能信息处理研究所,山西 太原 030006)

摘 要:复杂网络中的节点往往形成一些频繁出现且具有特定局部连接模式的高阶子结构,利用这些高阶子结构可以更好地刻画网络的拓扑特征及相关功能模块。通过度量节点间的结构相似性,有助于研究拓扑结构中节点之间的交互模式,理解复杂网络的局部结构和功能。为更充分利用节点邻域的高阶结构信息,提出了一种利用节点邻域内的k元节点组标签信息的结构相似性判定方法GANNLI(Group-based Aggregated Neighborhood Label Information)。该方法首先将k元节点组形成的非同构子图作为其组标签,再利用WL方法对k元节点组的邻域组标签信息进行聚合和更新,统计节点所构成的不同k元组的标签信息以得到节点表示,并利用余弦相似度计算节点间的结构相似性。与仅考虑节点度、接近中心性等低阶信息的方法相比,本方法利用高阶k元组结构信息更有效地度量了节点间的结构相似性。在真实网络数据集上的实验结果表明,所提出的GANNLI算法能更有效地计算节点间的结构相似性,在节点分类任务中的性能相比Struc2vec提高了2%至6%,相比Node2vec提高了8%至14%。

关键词:复杂网络;结构相似性;k元组;高阶结构;Weisfeiler-Lehman方法
中图分类号:O436 **文献标志码**:A **文章编号**:0253-2395(2024)05-0993-11

A Structure Similarity Determination Method for Aggregation Label Information of *k*-tuple Group in Node Neighborhood

YANG Gui[1], WEI Xingyu[1], ZHENG Wenping[1,2,3]

(1. *School of Computer and Information Technology, Shanxi University, Taiyuan* 030006*, China;*
2. *Key Laboratory of Computation Intelligence and Chinese Information Processing of Ministry of Education* (*Shanxi University*), *Taiyuan* 030006*, China;*
3. *Institute of Intelligent Information Processing, Shanxi University, Taiyuan* 030006*, China*)

Abstract: In complex networks, nodes often form higher-order substructures with specific local connectivity patterns that frequently appear. These higher-order substructures can better characterize the network's topological features and related functional modules. Measuring the structural similarity between nodes aids in studying the interaction patterns within the network's topology and understanding the local structures and functions of complex networks. To fully utilize the higher-order structural information in node neighborhoods, we propose a method for determining structural similarity using the label information of *k*-tuple groups in node neighborhoods, called GANNLI (Group-based Aggregated Neighborhood Label Information). This method first forms non-isomorphic subgraphs as group labels for *k*-tuple node groups, and then uses the Weisfeiler-Lehman (WL) method to aggregate and update the neighborhood group label information of *k*-tuple node groups. It counts the label information of different *k*-tuples formed by nodes to obtain node representations and calculates the structural similarity between nodes using cosine similarity. Compared to

收稿日期:2024-03-07;**接受日期**:2024-06-10
基金项目:国家自然科学基金(62072292);山西省1331工程项目
作者简介:杨贵(1975-),男,山西大同人,高级实验师,主要研究方向为机器学习。E-mail:gyang@sxu.edu.cn
引文格式:杨贵,韦兴宇,郑文萍.利用邻域k元节点组信息的节点结构相似性判定方法[J].山西大学学报(自然科学版),2024,47(5):993-1003. DOI:10.13451/j.sxu.ns.2024106

methods that only consider low-order information such as node degree and closeness centrality, our approach leverages higher-order *k*-tuple structural information to more effectively measure structural similarity between nodes. Experimental results on real network datasets demonstrate that the proposed GANNLI algorithm can more effectively calculate structural similarity between nodes, thereby improving the performance of node classification tasks. Specifically, the GANNLI shows a performance improvement of 2% to 6% over Struc2vec algorithm and an improvement of 8% to 14% over Node2vec algorithm. These results indicate that the GANNLI's ability to incorporate higher-order structural information into the analysis of node neighborhoods allows for more accurate and insightful modeling of complex networks, leading to enhanced understanding and better performance in practical applications.

Key words: complex network; structural similarity; *k*-tuple; higher-order structure; Weisfeiler-Lehman

0 引言

通常用$G=(V,E)$来表示节点集V,边集E的图G。节点作为图数据的基本元素,网络中功能相似的节点在结构和属性等方面有较高的相似性。节点相似性的研究在识别社交网络中心节点和分类交通网络中的枢纽节点重要性等方面应用广泛。基于节点自身的属性信息来度量节点间的相似性通常不考虑节点间的拓扑连接,两个节点属性越相似,其相似性越高。在社交网络中,可以根据两个人的年龄和兴趣等信息判断两个人是否可能成为朋友;在交通网络中,可以通过客运量和行程车速等节点属性的相似性识别关键枢纽节点。

实际上,节点邻域的拓扑结构在很大程度上决定了该节点在网络中所行使的功能,如社交网络中不同社区的中心节点具有相似的邻域拓扑结构;无标度网络中的Hub节点与网络中其他节点紧密连接;生物网络中代谢通路关键蛋白质节点倾向于大度节点。可以从结构相似性和同质相似性[1-3]两个角度对节点间的邻域拓扑结构相似性进行描述。同质相似性从共同邻居角度计算节点间相似性,如果两个节点共享较多的邻居节点,则它们的同质相似性较高。同质相似性高的节点位于网络中同一社区的可能性更大。

结构相似性利用更丰富的邻域拓扑性质定义节点间的相似性,如节点度、邻居节点度分布、接近中心性、聚集系数等信息。结构相似性较高的节点可能位于网络的不同社区,甚至位于不同的连通分量。如在社交网络不同社区中角色相似的节点往往有较高的结构相似性[4];在交通网络中不同区域的枢纽节点间结构相似性较高。早期的节点结构相似性计算方法如REGE和面向分类型数据的REGE算法(CATegorical REGE, CATREGE)迭代搜索节点邻居的最优匹配[5],时间复杂度较高,无法应用于大规模网络。基于表示学习的结构相似性计算方法Struc2vec[6]基于节点邻域内度分布构造相似图,在其上随机游走获取节点邻域上下文,以学习得到节点表示。然而,随机游走过程带来的随机性,可能导致结构相似的节点得到不同的邻域上下文,无法准确度量节点间的结构相似性。

实际上,网络中的节点间往往成组作用,形成一些频繁出现且具有特定局部连接模式的高阶子结构。利用这些高阶子结构可以更好地刻画网络的拓扑特征及其相关的功能模块[7]。例如在社会学中,通过网络中三角形的相对数量来描述网络的聚类特性;在分子化学中,官能团和环与物质的化学性质密切相关;在蛋白质研究中,类簇结构在构建蛋白质相关作用网络中起到关键作用[8]。利用节点成组结构特征度量节点间的结构相似性,在社交网络节点角色识别、交通网络枢纽节点重要性分类等方面具有广泛的应用和重要的实用价值。

为更充分利用网络高阶结构信息,本文提出了一种利用节点邻域内的k元节点组标签信息的结构相似性判定方法GANNLI(Aggregation Label Information of *k*-tuple Group in Node Neighborhood)。该方法首先按k元节点组形成的非同构子图作为其组标签,再利用WL方法对k元节点组的邻域组标签信息进行聚合,并更新组标签;统计节点所构成的不同k元组的标签信息得到节点表示;并利用余弦相似度计算节点间的结构相似性。与仅考虑节点度、接近中心性等低阶信息的方法相比,本方法利用高阶k元组结构信息更有效地度量了节点间的

结构相似性。在真实网络数据集上实验结果表明,利用节点组信息的结构相似性提高了节点分类任务的精确度。

1 相关工作

节点间相似性度量是进行图数据分析以及进行节点分类、节点聚类、链路预测、角色分析等下游任务的基础。对节点间结构相似性的度量,利用更丰富的邻域拓扑性质,在社交网络节点角色识别、交通网络枢纽节点重要性分类等方面应用广泛,具有重要的实用价值。如在Facebook、微博等社交网络中,可以发现不同社区中相同角色的人,从而提高对用户的针对性推荐和广告投放等工作的准确度;在交通网络中,识别重要枢纽节点,对重要枢纽节点予以保护。

图数据中节点相似性的度量,可以从同质相似性和结构相似性两方面进行描述。同质相似性通常根据节点之间的路径连通信息计算节点间的相似性。公共邻居相似性(CN)[9]、杰卡德相似性(Jaccard)[10]、Adamic-Adar相似性(AA)[11]等利用节点间的公共邻居信息计算相似性。其中公共邻居相似性指标[9]利用节点间公共邻居绝对数量计算相似性,杰卡德相似性指标[10]利用节点间公共邻居数量在两者所有邻居中的占比计算相似性。Zhou等基于网络资源分配的思想提出了资源分配相似性指标(RA)[12],以节点间的公共邻居作为传播媒介,将自身所拥有的资源平均分配给它的邻居,利用邻居节点可以接收到的资源数作为二者间的相似性值。Adamic等考虑到图数据度分布的偏态性影响提出Adamic-Adar相似性指标(AA)[11]。基于公共邻居的相似性度量本质上利用了节点间的2-路径信息,若两节点间拥有较多的2-路径,则它们拥有较多的公共邻居。为了更充分地利用节点间的路径连通信息以计算相似性,研究者根据节点间不同长度的路径数、节点间随机游走概率等定义了节点间的相似性。Katz指标利用网络中节点之间的所有路径进行相似性计算,越短的路径赋予越高的权重。Zhou等根据网络中节点间长度为2和3的路径数定义节点间的局部路径相似性度量(Local Path Similarity Index,LP)指标[12-13]。2015年Chen等根据网络中节点间的长度不超过3的路径数定义节点间的局部相似性指标(Local Similarity Index,LS)[14]。随着表示学习技术的发展,有很多的学者提出了一系列的对图结构进行表示学习的算法。2014年,Perozzi等提出DeepWalk[15]算法,在节点局部邻域上采用随机游走的思想进行节点采样得到节点路径序列,将路径作为语言模型的输入得到节点的学习表示。2016年,Grover等[16]针对DeepWalk算法中节点之间游走概率相等的缺点提出了Node2vec算法,通过引入参数p和q,控制随机游走的广度和深度,可以更加灵活地捕捉节点的上下文信息。基于路径的相似性度量扩大了邻域节点的探索范围,可以在更广泛邻域内对节点间相似性进行度量。然而因为大度节点与其他节点的关系紧密,度越大的节点在生成路径时被访问的概率越大,使得许多节点最相似的是大度节点[17]。

节点间的结构相似性度量了节点及周围邻域的拓扑结构信息,如节点度[18]、邻居节点度分布、接近中心性、节点间距离分布、聚集系数等信息。Zhang等基于一阶邻域度特征信息提出了局部相对方法(Local Relative Entropy,LRE)[19],利用节点一阶邻域度分布描述其邻域拓扑结构,利用相对熵计算节点一阶邻域度分布差异以得到节点间的相似性。Mu等提出了基于节点间距离分布的结构相似性度量方法DDRE(Distance Distribution and Relative Entropy-based Similarity)[20],通过节点之间的最短路径定义节点的距离分布,DDRE需要遍历网络中所有节点,计算代价较高。Zou等将节点相似性矩阵与平均聚合机制相结合对图神经网络的邻域聚合机制进行改进[21]。Ribeiro等基于语言模型提出了算法Struc2vec[6],利用节点邻域内度序列得到节点间的相似性,并作为节点间路径权重利用随机游走得到节点在拓扑结构相似图上的邻域上下文,最终学习得到反映节点结构相似性的潜在表示。然而随机游走过程带来的随机性,导致结构完全相同的两个节点得到的邻域上下文会存在不同,从而使得结构完全相同的两个节点得到的嵌入表示存在一定的

距离。

2 背景知识

2.1 基本概念

用$G=(V,E)$来表示图数据,其中$V(G)=\{v_1,\cdots,v_n\}$表示图数据中实体的集合,$|V(G)|=n$表示图数据中的实体数量,$E(G)$表示图数据中实体间关系的集合,$|E(G)|=m$代表实体间关系的数量。在图G中,节点v_i的一阶邻域表示为$N(v_i)=\{v_j:(v_i,v_j)\in E(G)\}$,节点$v_j\in N(v_i)$称为节点$v_i$的一阶邻居。节点$v_i$的度为$d(v_i)=|N(v_i)|$表示图$G$中与节点$v_i$有连边的节点数,简记$d_i$。假设图$S$为图$G$的一个节点数为$k$的非空子图,其节点集$V(S)=\{w_1,\cdots,w_k\}\subseteq V(G)$且边集$E(S)=\{(w_i,w_j):w_i,w_j\in V(S)\}\subseteq E(G)$,用$\left[V(G)\right]^k$表示所有节点数为$k$的非空子图的集合,将节点数为$k$的非空子图称为图$G$的一个$k$元节点组。

2.2 Weisfeiler-Lehman图核

常见的图核方法以图的分解方式进行分类,主要有基于路径的图核方法[22]、基于子图的图核方法[23]以及基于子树的图核方法[24]。这三类方法分别将图分解为路径、子图、子树等不同的组件,使用R-卷积核[25]对比每对组件对图的相似度进行计算[26],被广泛应用于药物研发[27]、生物网络链路预测[28]中。Weisfeiler-Lehman(WL)图核方法[29-30]是使用频率最高的基于子树的图核方法,它将WL算法与核函数相结合以有效进行图结构相似性分析。设待比较的两个图分别为G和G'($|V(G)|=|V(G')|$),WL算法迭代的为图中的每个顶点赋予标签,若在某次迭代中两图对应的标签序列不同,则可判定G和G'不同构;若算法迭代$|V(G)|$次,对应的标签序列仍相同,则两图同构。每一次迭代主要分为四个步骤:标签初始化、邻域标签表示、标签压缩、重标签,分别对应于图1中①、②、③、④。对网络中的节点v,将其所有邻居节点的标签排序得到邻域标签字符串,合并v标签与排序后的邻域标签集,并利用hash函数压缩合并后的标签集合为一个新标签号以更新v的标签。WL算法为图G赋予标签的T步迭代过程如算法1所示。

算法1 WL算法为图G节点赋予标签

输入:图$G=(V,E)$,最大迭代次数T;

输出:图G的标签向量。

1. 令$t=-1$,对任意节点$v_i\in V(G)$初始化其标签为v_i的度,即$l^{(0)}(v_i)=d_i$;
2. 令$t=t+1$,若$t>T$,则转步骤6;
3. 对于每个节点$v_i\in V(G)$,对v_i的邻居节点标签按照字典序排列,得到排序后的邻居节点标签序列为$l^{(t)}(v_{j_1}),l^{(t)}(v_{j_2}),\cdots,l^{(t)}(v_{j_{d_i}})$,赋予$v_i$的一个临时标签,定义为$\ell(v_i)=l^{(t)}(v_i)+l^{(t)}(v_{j_1})+\cdots+l^{(t)}(v_{j_{d_i}})$,其中'+'代表字符串拼接;
4. 使用hash函数将节点v_i的临时标签$\ell(v_i)$映射为一个压缩标签$l^{(t+1)}(v_i)$;
5. 若新标签的种类与迭代前相同,即$|\{l^{(t+1)}(v_i):v_i\in V\}|\neq|\{l^{(t)}(v_i):v_i\in V\}|$,或对应类别的节点数不同,则转步骤2;
6. 结束,返回图G中的节点标签$\{l^{(t)}(v_i):v_i\in V\}$。

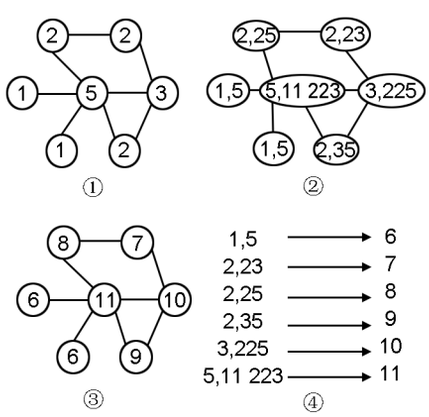

图1 WL方法赋予节点标签的一次迭代过程

Fig. 1 An iterative process of the WL algorithm for assigning labels to nodes

当网络中节点信息较少时,WL算法通常采用节点度作为初始标签,这导致算法不能很好地区分一些非同构图[31],如图2所示的两个图由于具有相同的度序列,且每个节点的邻居度也完全相同,导致WL算法无法区分。这是由于算法1中仅利用单个节点及其邻域的度信息进行图同构检测,无法反映图中节点在高阶结构方面的差异。针对此,Babai[30]提出了利用网络中由k个节点组成的高阶结构进行图同构判定的方法,称为k-Weisfeiler-Lehman方法,简称k-WL方法[31],迭代地为网络中的k元节点组赋予标签,对比两个网络的k元组标签在迭代过程的差异进行同构判定,用户可根据网络特征,选择合适的k来增强算法区分非同构图的能力。通

常,k越大则算法的图同构判定能力越强。显然,WL算法是k-WL方法在k=1时的特例。

尽管WL算法是对两个图是否同构进行判定的,实际上,其迭代的赋予节点或节点组标签的过程,也是对节点或节点组进行编码的过程。局部邻域范围内拓扑结构相似的节点通常会被赋予相似的标签,本文利用WL算法的编码过程对节点的局部邻域结构相似性进行判定。

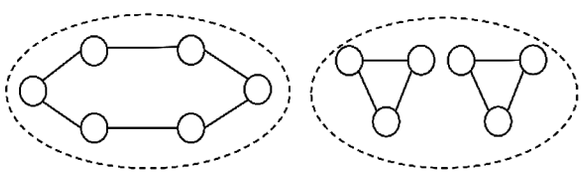

图2 WL算法无法区分的两个图

Fig. 2 Two graphs that WL algorithm cannot distinguish

3 节点间结构相似性度量方法GANNLI

3.1 k元节点组的邻域

为了提取k元节点组的邻域结构,对k元节点组P的邻域进行了定义,如下式所示:

$$\Gamma(P)=\{p_1,\cdots,p_{i-1},q,p_{i+1},\cdots,p_k:\\ q\in N_1(p_i)\wedge q\notin P\},$$

k元节点组的邻域$\Gamma(P)$是通过用被替换节点p_i的邻域节点$q\in N(p_i)$来替换的,即被替换节点与替换节点之间有连边。如图3所示,k元节点组$\{1,2,5\}$的邻域k元节点组有$\{1,2,4\}$和$\{1,5,6\}$,其中$\{1,2,4\}$是节点5的邻域节点4替换节点5得到的,$\{1,5,6\}$是节点2的邻域节点6替换节点2得到的。

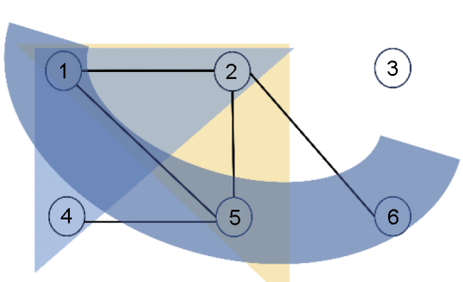

图3 k元组{1,2,5}的邻域

Fig. 3 The neighborhood of the k-tuple {1, 2, 5}

3.2 GANNLI方法

GANNLI方法包括生成k元节点组、k元节点组组标签初始化、k元节点组邻域聚合、特征统计、节点间结构相似性计算等5个主要过程。

首先,对于图中所有的节点$V(G)$,生成k元节点组的集合$[V(G)]^k$。判断k元节点组P组成的子结构类型作为k元节点组的组标签。在k元节点组邻域聚合部分利用WL方法对k元节点组的邻域组标签信息进行聚合,并更新组标签。在特征统计过程中,对包含节点v的k元节点组的不同标签类型及对应数量进行统计,生成关于节点v的特征向量$F(v)$。在节点间结构相似性计算部分,使用余弦相似度,如式(1)所示。

$$\cos(F(u),F(v))=\frac{F(u)^TF(v)}{\|F(u)\|\cdot\|F(v)\|}。\quad(1)$$

对节点v和节点u的特征向量$F(u)$和$F(v)$进行余弦相似度计算,从而得到节点v和节点u之间的相似性。

算法2给出了GANNLI算法框架。

4 实验分析

4.1 k元节点组表达能力分析

为了验证不同大小的k元节点组对网络结构的表达能力,设计了一个验证实验。如图4所示图中有4个连通分量,其节点可分为7类(用不同颜色代表)。

图5展示了不同k元节点组对图4的分类结果。当k=1,图中所有节点被识别为同一类。当k=2时,节点将按度分类,因此图中节点被分为2度点和3度点;当k=3时,由于图中黄色和粉色节点的所有3元组构成情况相同,因此被错误识别为同一类;而当k=4时,所有类别不同的节点的4元组构成情况各异,因此将所有节点都划分到正确类别。可以看出,不同规模的k元节点组对网络中节点具有不同的表达能力,通常k越大,不同类别节点的k元组构成情况差别越明显,因此算法的分类能力也增强。

4.2 真实网络上的节点相似性比较

本节实验采用网络是由两个完全相同的跆拳道网络(Karate)构成的镜像Karate网络,如图6所示,其中边框颜色相同的节点互为镜像节点。分别利用本节所提算法GANNLI和传统方法Struc2vec得到镜像Karate网络中节点的表示,其中将GANNLI算法得到的最终标签向量作为节点表示。图7给出了用t-SNE对两种算法所得表示进行降维可视化的效果图,其中红

算法2 算法GANNLI(G, T, k)

输入:图$G=(V, E)$,节点组大小k,邻域聚合次数T;

输出:节点结构相似性矩阵$\boldsymbol{S}$。

1. 令$t=-1$,生成所有k元节点组集合$[V(G)]^k$;
2. 按k元节点组的子结构类型,初始化每个k元节点组P的组标签$l^{(0)}(P)$;
3. 令$t=t+1$,若$t>T$,则转步骤8;
4. 令$L^{(t)}([V(G)]^k)=\{l^{(t)}(P):P\in[V(G)]^k\}$表示$[V(G)]^k$的标签集合,令$\tau^{(t)}=|L^{(t)}([V(G)]^k)|$;
5. 对任意节点$v_i\in V(G)$:
5.1. 令$[V(G)]_i^k=\{P:v_i\in P\wedge P\in[V(G)]^k\}$表示包含节点$v_i$的$k$元节点组;
5.2. 令$f^{(t)}(v_i)=[l_1(v_i),\cdots,l_{\tau^{(t)}}(v_i)]$表示$v_i$的第$t$轮标签向量,其中$l_j(v_i)$表示$[V(G)]_i^k$中第$j(1\leqslant j\leqslant\tau^{(t)})$类标签的出现次数;
6. 对任意节点组$P=\{p_1,\cdots,p_i,\cdots,p_k\}\in[V(G)]^k$:
6.1. 根据式(1)得到其邻域节点组
$\Gamma(P)=\{p_1,\cdots,p_{i-1},q,p_{i+1},\cdots,p_k:$
$q\in N_1(p_i)\wedge q\notin P\}\subseteq[V(G)]^k$;
6.2. 将$\Gamma(P)$中元素按照其k元组标签字典非递减排序,即$l^{(t)}(P_1)\leqslant l^{(t)}(P_2)\leqslant\cdots\leqslant l^{(t)}(P_{|\Gamma(P)|})$;
6.3. 赋予P的一个临时标签,定义为$\ell(P)=l^{(t)}(P)+l^{(t)}(P_1)+\cdots+l^{(t)}(P_{\|N_k(P)\|})$,其中'+'代表字符串拼接;
6.4. 使用hash函数将P的临时标签$\ell(P)$映射为一个压缩标签$l^{(t+1)}(P)$;
7. 若标签种类与迭代前相同,即$|\{l^{(t+1)}(P):P\in[V(G)]^k\}|\neq|\{l^{(t)}(P):P\in[V(G)]^k\}|$,或对应类别的节点数不同,则转步骤3;
8. 对每个节点$v_i\in V$,拼接各轮标签向量得到最终标签向量$F(v_i)=[f^{(0)}(v_i),f^{(1)}(v_i),\cdots,f^{(T)}(v_i)]$;
9. 对任意节点对$v_i,v_j\in V(G)$,使用式(1)计算v_i和v_j标签向量的相似性$S_{ij}=\cos(F(v_i),F(v_j))$;
10. 算法结束,返回相似性矩阵$\boldsymbol{S}$。

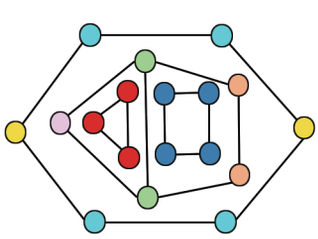

图4 包含4个连通分量的图

Fig. 4 A graph with 4 connected components

色结果展示本文算法的可视化结果,蓝色结果展示Struc2vec算法的可视化结果,坐标系上方与右方的箱体分布图分别表示横纵方向上节点的分布结果。

从图7可以看出,互为镜像的节点通过Struc2vec算法所得到节点表示并不完全相同,而通过本文所提的GANNLI方法可以得到完全相同的标签向量。这是由于Struc2vec利用随机游走得到节点特征序列以产生节点表示,而游走过程的随机性可能使两个结构完全相同的节点得到不同的节点特征序列,从而导致它们最终的节点表示产生差异。除镜像节点外,结构相似性高的节点利用GANNLI算法可以得到更相似的节点表示,如节点组{15, 16, 19, 21, 23}在网络中都仅与节点33和34相连,因此GANNLI对这5个节点给出了完全相同的标签向量,而节点1和34分别为真实社区中的中心节点,它们的标签向量也距离更近。

图8展示了GANNLI和Struc2vec算法对镜像Karate俱乐部网络的节点表示与节点度之间的分布关系。由图可知,两种方法节点表示与节点度之间的分布相似,验证了本文所提方法GANNLI利用k元节点组对节点间结构相似性计算的正确性。

图9展示了在镜像Karate网络上,GANNLI与Struc2vec方法得到节点表示间的距离分布对比情况,包括镜像节点对间的距离分布以及所有节点对间的距离分布,其中标记"+"的曲线为镜像对之间的距离,标记"×"的曲线为所有对之间的距离。由于本文所提算法GANNLI赋予了镜像节点对相同的标签向量,因此镜像节点间的平均距离为0,而Struc2vec镜像节点间的平均距离为0.87。此外,GANNLI方法所得表示的所有节点对的平均距离为9.34,而Struc2vec方法所有节点对的平均距离为8.06,这表明所提方法得到的节点表示对网络节点的区分性优于Struc2vec。

4.3 在分类任务上的比较结果

节点分类任务是一个常见节点表示的下游任务。当节点的标签与节点的结构有关而不是与邻居节点的标签有关时,分类时更应该关注节点间的结构相似性。本节使用空中交通网络数据集来验证所提算法GANNLI对节点间结构相似性的判别能力。数据集包含巴西空中交通网络(Brazil_airports)、美国空中交通网络(USA_airports)和欧洲空中交通网络(Europe_airports)等3个无权无向网络,其中节点代

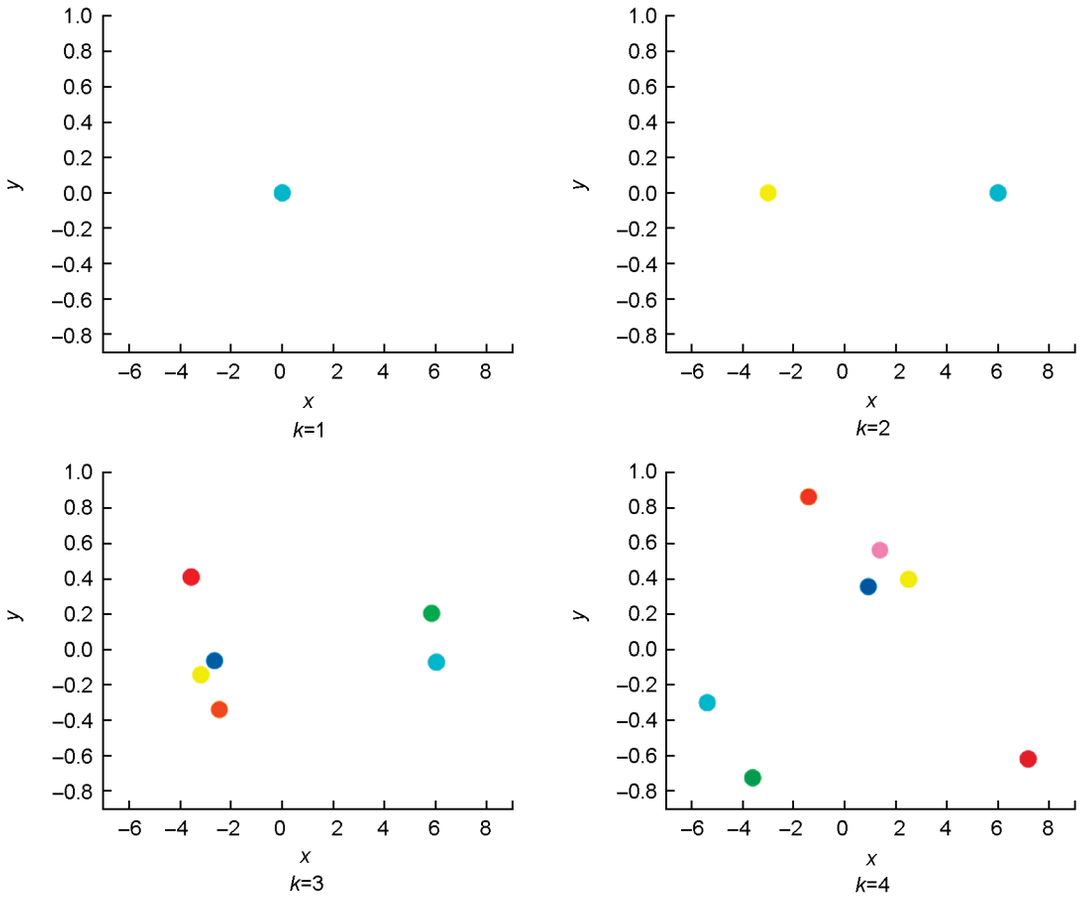

图5 不同大小的k元节点组分类结果展示

Fig. 5 Classification results of different k-tuples

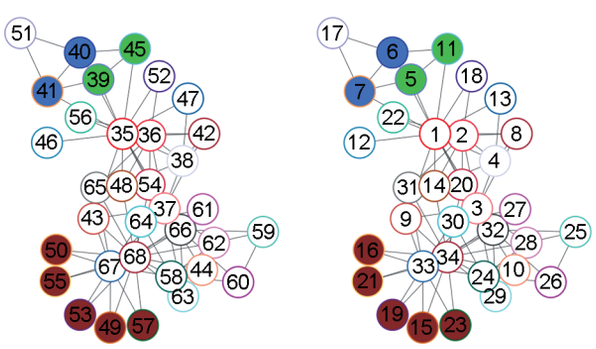

图6 镜像Karate网络,边框颜色相同的节点表示对应的镜像节点

Fig. 6 Mirror Karate network, nodes with the same border color represent corresponding mirrored nodes

表机场,边表示机场间存在直达航班。根据机场的航班或人员的活动水平为机场分配标签。对每个数据集,按照机场活动水平将机场分为四组,按机场活动水平排名前25%、25%~50%、50%~75%、后25%赋予4类不同标签。

• 巴西空中交通网络(Brazil_airports):从巴西国家民航局(ANAC)收集2016年1月至12月的数据,有131个节点,1 038个边(直径为5)。机场活动水平通过对应时期的航班起降总数来衡量。

• 美国空中交通网络(USA_airports):数据收集自美国交通统计局2016年1月至10月,有1 190个节点,13 599个边(直径为8)。机场活动水平通过对应时期通过机场客流量来衡量。

• 欧洲空中交通网络(Europe_airports):从欧盟统计局(Eurostat)收集的2016年1月至11月的数据,有399个节点,5 995个边(直径为5)。机场活动水平通过对应时期的航班起降总数来衡量。

本节将所提方法GANNLI所得的节点表示用于分类任务,在空中交通网络数据集上,与Struc2vec、Node2vec算法所得的节点表示进行对比实验。对Node2vec,采用节点度作为节点的特征信息,由采样节点的度信息构成采样路径信息。GANNLI算法采用节点的3元组标签信息,即k=3。采用支持向量机(SVM)、决策

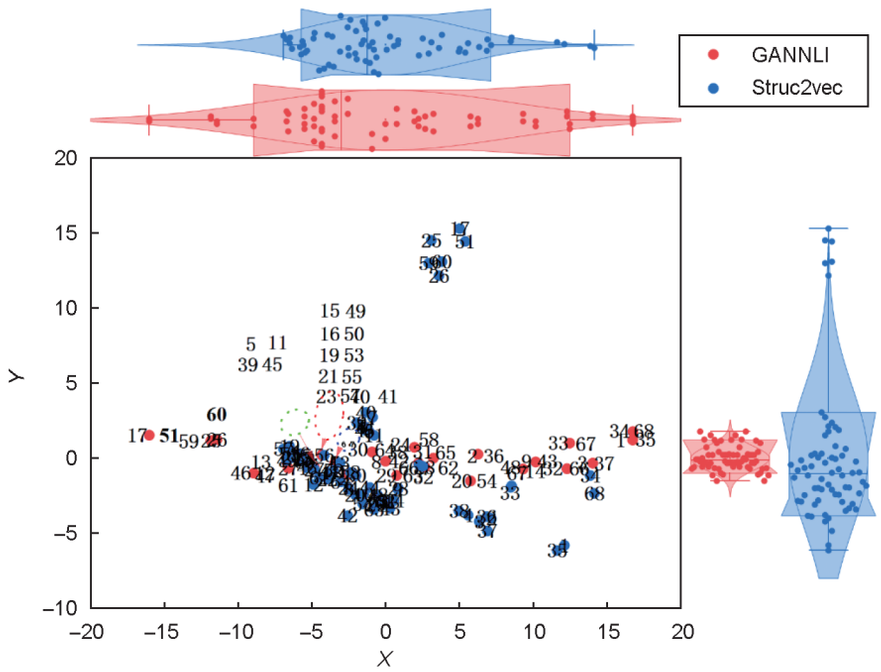

图7 GANNLI和Struc2vec算法对镜像Karate网络的节点表示

Fig. 7 Nodes representations of GANNLI and Struc2vec algorithms in mirrored Karate network

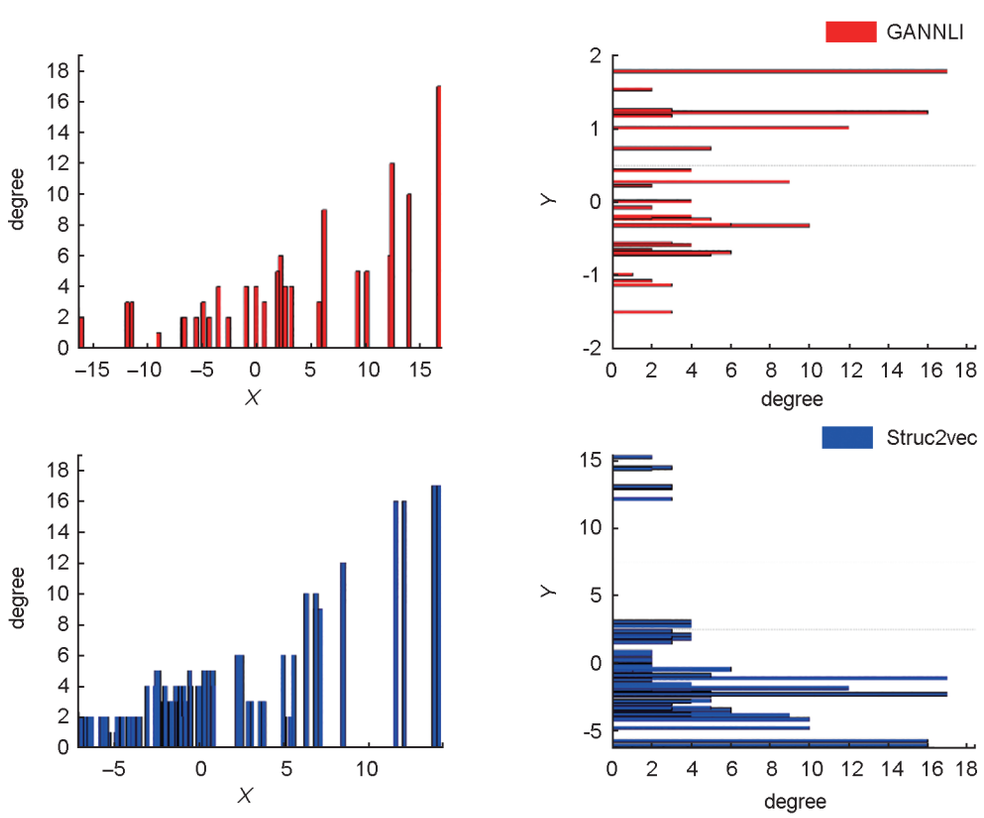

图8 GANNLI和Struc2vec算法对镜像Karate网络的节点表示与节点度之间的分布关系

Fig. 8 Nodes representations of GANNLI and Struc2vec algorithms for mirrored Karate network and the distribution relationship with node degree

树(DTree)、逻辑回归(LR)等3种分类器评价节点表示的分类性能,实验比较结果如图10所示。可以看到,GANNLI算法在巴西空中交通网络(Brazil_airports)、美国空中交通网络

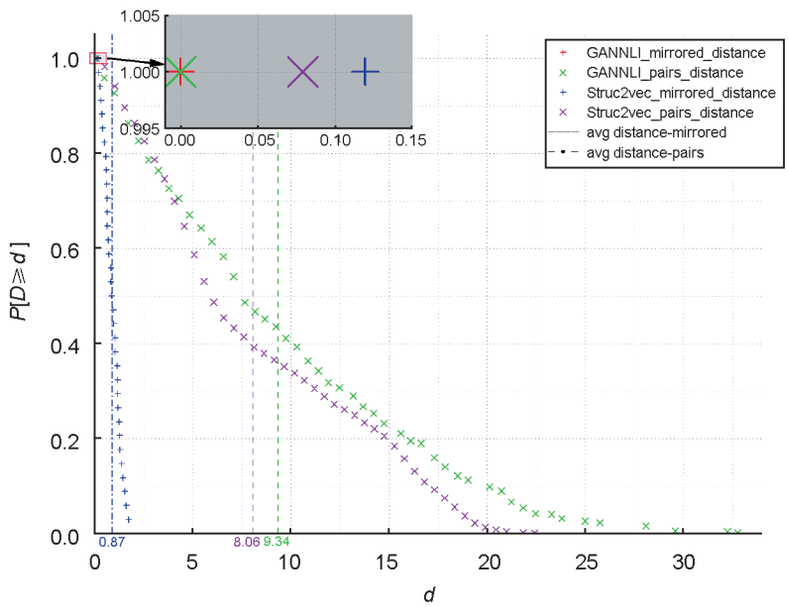

图9 GANNLI与Struc2vec对镜像Karate网络的节点表示

Fig. 9 GANNLI and Struc2vec representation of nodes in a mirrored Karate network

(USA_airports)和欧洲空中交通网络(Europe_airports)等数据集上的分类性能相比Struc2vec提高了2%至6%,相比Node2vec提高了8%至14%。

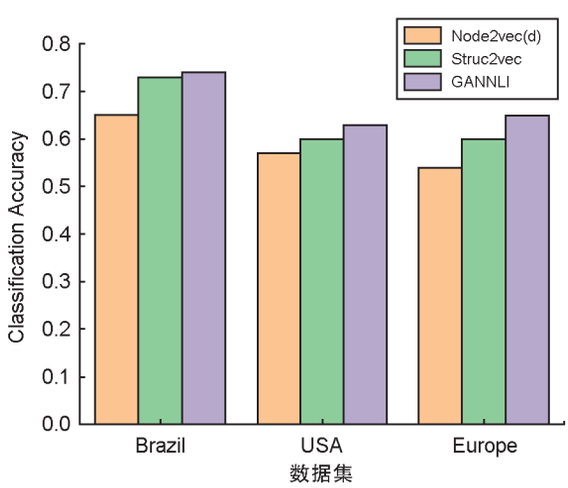

图10 利用Node2vec(d)、Struc2vec、GANNLI方法在Brazil_airports、USA_airports、Europe_airports数据集上的分类精度比较

Fig. 10 Comparison of classification accuracy using Node2vec(d), Struc2vec and GANNLI methods on Brazil_airports, USA_airports and Europe_airports datasets

5 总结

本文提出一种利用节点邻域内的k元节点组标签信息的结构相似性判定方法GANNLI,首先按k元节点组形成的非同构子图作为其组标签,再利用WL方法对k元节点组的邻域组标签信息进行聚合,并更新组标签;统计节点所构成的不同k元组的标签信息得到节点表示;利用余弦相似度计算节点间的结构相似性。在真实网络数据集上实验结果表明,本文算法GANNLI能更有效地计算节点间的结构相似性,从而提高节点分类任务的性能。

下一步,GANNLI算法可以进一步优化,以适应更大规模的网络数据和更加复杂的网络结构。此外,将该方法应用于其他类型的图分析任务,如社区发现、链路预测和图嵌入等,可能会带来更多有价值的发现。结合其他先进的图神经网络技术,GANNLI有望在更广泛的应用领域中发挥更大的作用,提高复杂网络分析的精度和效率。进一步的研究还可以探索如何在动态网络环境中有效地应用和扩展GANNLI,以应对实时数据变化带来的挑战。

参考文献:

[1] FORTUNATO S. Community Detection in Graphs[J]. *Phys Rep*, 2010, **486**(3/4/5): 75-174. DOI: 10.1016/j.physrep.2009.11.002.

[2] HENDERSON K, GALLAGHER B, ELIASSI-RAD T, *et al*. RolX: Structural Role Extraction & Mining in

Large Graphs[C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2012: 1231−1239. DOI: 10.1145/2339530.2339723.

[3] YANG J, LESKOVEC J. Overlapping Communities Explain Core-Periphery Organization of Networks[J]. *Proc IEEE*, 2014, **102**(12): 1892−1902. DOI: 10.1109/JPROC.2014.2364018.

[4] ROSSI R A, AHMED N K. Role Discovery in Networks [J]. *IEEE Trans Knowl Data Eng*, 2015, **27**(4): 1112−1131. DOI: 10.1109/TKDE.2014.2349913.

[5] TU K, CUI P, WANG X, *et al*. Deep Recursive Network Embedding with Regular Equivalence[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2018: 2357−2366. DOI: 10.1145/3219819.3220068.

[6] RIBEIRO L F R, SAVARESE P H P, FIGUEIREDO D R. Struc2vec: Learning Node Representations from Structural Identity[EB/OL]. arXiv Preprint: 1704, 03165, 2017. https://arxiv.org/abs/1704.03165.

[7] ARVIND V, FUHLBRÜCK F, KÖBLER J, *et al*. On Weisfeiler-leman Invariance: Subgraph Counts and Related Graph Properties[J]. *J Comput Syst Sci*, 2020, **113**: 42−59. DOI: 10.1016/j.jcss.2020.04.003.

[8] BAEK M, DIMAIO F, ANISHCHENKO I, *et al*. Accurate Prediction of Protein Structures and Interactions Using a Three-track Neural Network[J]. *Science*, 2021, **373** (6557): 871−876. DOI: 10.1126/science.abj8754.

[9] LÜ L Y, ZHOU T. Link Prediction in Complex Networks: A Survey[J]. *Phys A Stat Mech Appl*, 2011, **390** (6): 1150−1170. DOI: 10.1016/j.physa.2010.11.027.

[10] ERTL O. ProbMinHash-A Class of Locality-sensitive Hash Algorithms for the (Probability) Jaccard Similarity [J]. *IEEE Trans Knowl Data Eng*, 2022, **34**(7): 3491−3506. DOI: 10.1109/TKDE.2020.3021176.

[11] ADAMIC L A, ADAR E. Friends and Neighbors on the Web[J]. *Soc Netw*, 2003, **25**(3): 211−230. DOI: 10.1016/s0378-8733(03)00009-1.

[12] ZHOU T, LÜ L Y, ZHANG Y C. Predicting Missing Links via Local Information[J]. *Eur Phys J B*, 2009, **71** (4): 623−630. DOI: 10.1140/epjb/e2009-00335-8.

[13] BASTAMI E, MAHABADI A, TAGHIZADEH E. A Gravitation-based Link Prediction Approach in Social Networks[J]. *Swarm Evol Comput*, 2019, **44**: 176−186. DOI: 10.1016/j.swevo.2018.03.001.

[14] CHEN Z Q, XIE Z, ZHANG Q. Community Detection Based on Local Topological Information and Its Application in Power Grid[J]. *Neurocomputing*, 2015, **170**: 384−392. DOI: 10.1016/j.neucom.2015.04.093.

[15] PEROZZI B, AL-RFOU R, SKIENA S. DeepWalk: Online Learning of Social Representations[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014: 701−710. DOI: 10.1145/2623330.2623732.

[16] GROVER A, LESKOVEC J. Node2vec: Scalable Feature Learning for Networks[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 855−864. DOI: 10.1145/2939672.2939754.

[17] ZHENG W P, CHEN C H, QIAN Y H, *et al*. A Graph Clustering Algorithm Based Paths Between Nodes in Complex Networks[J]. *Chin J Comput*, 2020, **43**(7): 1312−327. DOI: 10.1016/j.phy sa.2016.11.015.

[18] BURT R. Structural Holes and Good Ideas[J]. *Am J Sociol*, 2004, **110**(2): 349−399. DOI: 10.1086/421787.

[19] ZHANG Q, LI M Z, DENG Y. Measure the Structure Similarity of Nodes in Complex Networks Based on Relative Entropy[J]. *Phys A Stat Mech Appl*, 2018, **491**: 749−763. DOI: 10.1016/j.physa.2017.09.042.

[20] MU J F, LIANG J Y, ZHENG W P, *et al*. Node Similarity Measure for Complex Networks[J]. *J Front Comput Sci Tech*, 2020, **14**(5): 749−759. DOI: 10.3778/j.issn.1673-9418.1905015.

[21] ZOU M H, GAN Z X, CAO R Z, *et al*. Similarity-navigated Graph Neural Networks for Node Classification[J]. *Inf Sci*, 2023, **633**: 41−69. DOI: 10.1016/j.ins.2023.03.057.

[22] SHERVASHIDZE N, SCHWEITZER P, VAN LEEUWEN E J, *et al*. Weisfeiler-Lehman Graph Kernels[J]. *J Mach Learn Res*, 2011, **12**: 2539−2561.

[23] MAHÉ P, VERT J P. Graph Kernels Based on Tree Patterns for Molecules[J]. *Mach Learn*, 2009, **75**(1): 3−35. DOI: 10.1007/s10994-008-5086-2.

[24] GÄRTNER T, FLACH P, WROBEL S. On Graph Kernels: Hardness Results and Efficient Alternatives[C]//Learning Theory and Kernel Machines. Berlin, Heidelberg: Springer, 2003: 129−143. DOI:10.1007/978-3-540-45167-9_11.

[25] MAHÉ P, UEDA N, AKUTSU T, *et al*. Extensions of Marginalized Graph Kernels[C]//Twenty−first international conference on Machine learning−ICML '04. New York: ACM, 2004: 70−77. DOI: 10.1145/1015330.1015446.

[26] ABURIDI M, MARCIA R. Wasserstein Distance-based Graph Kernel for Enhancing Drug Safety and Efficacy Prediction[C]//2024 IEEE First International Conference on Artificial Intelligence for Medicine, Health and Care (AIMHC). New York: IEEE, 2024: 113−119. DOI:

10.1109/AIMHC59811.2024.00029.

[27] LI M, WANG Z, LIU L, *et al*. Subgraph-Aware Graph Kernel Neural Network for Link Prediction in Biological Networks[J]. *IEEE J Biomed Health*, 2024, **28**(7): 4373−4381. DOI: 10.1109/JBHI.2024.3390092.

[28] CAI J Y, FURER M, IMMERMAN N. An Optimal Lower Bound on the Number of Variables for Graph Identification[C]//30th Annual Symposium on Foundations of Computer Science. New York: IEEE, 1989: 612−617. DOI: 10.1109/SFCS.1989.63543.

[29] KIEFER S, SCHWEITZER P. Upper Bounds on the Quantifier Depth for Graph Differentiation in First Order Logic[C]//Proceedings of the 31st Annual ACM/IEEE Symposium on Logic in Computer Science. New York: ACM, 2016:287−296. DOI: 10.1145/2933575.2933595.

[30] BABAI L. Graph Isomorphism in Quasipolynomial Time[C]//Proceedings of the Forty−eighth Annual ACM Symposium on Theory of Computing. New York: ACM, 2016: 684−697. DOI: 10.1145/2897518.2897542.

[31] MORRIS C, RITZERT M, FEY M, *et al*. Weisfeiler and Leman Go Neural: Higher-order Graph Neural Networks[J]. *Proc AAAI Conf Artif Intell*, 2019, **33**(1): 4602−4609. DOI: 10.1609/aaai.v33i01.33014602.