计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(10): 2733−2751, 2026

基于对比学习的图结构规则学习方法

董博文[1] 刘雪莉[2] 冯潇洋[3] 麦珍珍[3] 王文俊[1]
[1](天津大学人工智能学院 天津 300350)
[2](天津大学网络安全学院 天津 300350)
[3](天津大学计算机科学与技术学院 天津 300350)
(1020201107@tju.edu.cn)

Graph-Structured Rule Learning Method via Contrastive Learning

Dong Bowen[1], Liu Xueli[2], Feng Xiaoyang[3], Mai Zhenzhen[3], and Wang Wenjun[1]
[1](*School of Artificial Intelligence*, *Tianjin University*, *Tianjin* 300350)
[2](*School of Cybersecurity*, *Tianjin University*, *Tianjin* 300350)
[3](*School of Computer Science and Technology*, *Tianjin University*, *Tianjin* 300350)

Abstract Knowledge graph reasoning has attracted extensive attention because of its ability to alleviate the inherent incompleteness of knowledge graphs. However, existing approaches mainly rely on embedding models or chain-based rules, but they still suffer from limitations in interpretability and expressiveness. Existing rule learning methods are often limited to linear dependency modeling and struggle to capture multi-path interactions and hierarchical semantic relations, while traditional rule-mining methods usually follow the "enumeration–verification" paradigm, resulting in high computational costs and insufficient modeling of deep semantic correlations among rules. To address these challenges, we propose CGRL (contrastive graph rule learning), a novel method for learning graph-structured rules through contrastive learning. CGRL enhances rule learning through a threefold mechanism. First, it applies an adaptive clustering strategy to generate soft type labels for entities, introducing semantic priors that improve rule generalization. Second, it designs a multi-hop neighborhood sampling method to construct semantic subgraphs and to generate high-quality positive and negative rule instances. Third, it employs a Transformer-based encoder and a contrastive learning objective to map rule pairs into a unified embedding space, thereby jointly capturing structural dependencies and semantic consistency. Extensive experiments on six public knowledge graph completion benchmarks demonstrate that CGRL achieves substantial improvements over state-of-the-art baselines. On datasets with relatively complex relational structures, such as FB15k-237 and WN18RR, CGRL yields average gains of 6.04% in MRR and 16.26% in Hits@1, confirming its effectiveness in discovering high-quality graph-structured rules and significantly enhancing reasoning performance.

Key words knowledge graph completion; knowledge graph reasoning; rule learning; graph-structured rules; contrastive learning

摘 要 知识图谱推理因其能够有效缓解知识图谱中的数据不完整问题而受到广泛关注。现有方法多依赖于嵌入模型或链式规则,但在可解释性与表达能力方面仍存在明显局限。一方面,基于闭合路径的规则学习仅能建模线性链式依赖,难以捕捉图结构中存在的多路径交互与层次化语义关系;另一方面,基于

收稿日期: 2025-11-24; **修回日期**: 2026-05-29
基金项目: 国家自然科学基金项目(62572345)
This work was supported by the National Natural Science Foundation of China (62572345).
通信作者: 刘雪莉(xueli@tju.edu.cn)

统计指标的传统方法通常采用“穷举+验证”策略,不仅在大规模图谱中面临高昂计算成本,也难以充分挖掘规则间的深层语义关联,从而制约了推理性能的进一步提升。为此,提出一种基于对比学习的图结构规则学习方法 CGRL(contrastive graph rule learning),以同时提升推理的表达能力与可解释性。CGRL 通过三重机制实现高效规则学习:首先,采用自适应聚类策略为节点生成软标签,引入类别语义以增强规则泛化能力;其次,设计多跳邻域采样方法构建语义子图,并生成高质量正负规则样本;最后,利用基于 Transformer 的编码器学习规则表示,并通过对比学习将规则对映射至统一向量空间以优化语义一致性判别。在 6 个公开知识图谱补全任务上的实验结果表明,CGRL 取得了显著性能提升。在 FB15k-237 和 WN18RR 这种关系结构相对复杂的数据集下,其 MRR 指标平均提升 6.04%,Hits@1 指标平均提升 16.26%,显著优于现有基线方法,验证了其在学习高质量图结构规则与提升推理效果方面的有效性。

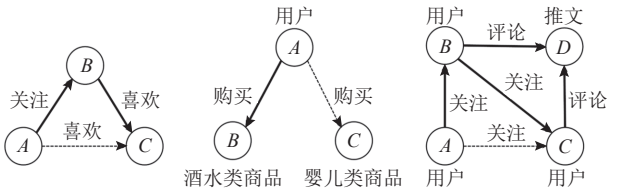
关键词 知识图谱补全;知识图谱推理;规则学习;图结构规则;对比学习

中图法分类号 TP311.13

DOI: 10.7544/issn1000-1239.202550845 **CSTR:** 32373.14.issn1000-1239.202550845

知识图谱(knowledge graph, KG)将知识表示为结构化三元组,通过图结构展现实体间的丰富关联,广泛支撑智能问答和语义搜索等任务[1-2]。然而,由于数据采集与构建过程的局限性,构建的知识图谱往往具有不完备性,这在一定程度上限制了下游应用的性能[1]。在开放世界假设下,未被收录于图谱中的事实并不能直接视为错误或不存在,这使得知识补全成为知识图谱应用中的关键任务[2]。围绕这一问题,知识图谱推理(knowledge graph reasoning, KGR)被广泛研究,其核心在于根据已有三元组及其结构语义信息推断潜在的缺失事实[3-6]。

当前的知识图谱推理方法主要包括基于嵌入的方法[1,7-11]与基于规则的方法[12-20]。前者通过低维向量捕捉语义与结构信息[21],广泛应用于链路预测[21]和实体分类[22]等任务,但在统一建模对称、非对称与组合关系时存在局限,且依赖“黑盒”计算,缺乏可解释性;后者则通过逻辑规则进行显式推理[21],具有较强的可解释性,但大多局限于链式规则,难以刻画复杂图模式。事实上,图中的关联结构远不止链式,还包括树结构与更一般的图结构:链式规则通常只能刻画如图 1(a)所示的简单路径推导关系,例如由用户之间的关注关系和用户对对象的喜欢关系,推断新的潜在喜欢关系,而树结构规则可揭示典型的协同模式(如图 1(b)所示“酒水类商品与婴儿类商品”的共购关系),图结构规则还能捕捉社交网络中“共同关注和评论”等多方交互模式(图 1(c)所示)。在实际应用中,这类复杂规则具有重要价值:推荐系统中,用户关注行为往往由共同好友和共同兴趣共同驱动;药物发现领域,药物—靶点—疾病之间的作用关系往往构成多靶点协同网络;社交平台上,交互模式更是天然呈现树状或图状结构。因此,仅依赖链式规则不仅导致推理结果片面、补全能力受限,还限制了下游任务的准确性与可解释性,凸显了开展图结构规则学习的现实必要性。



(a) 链式规则 (b) 树结构规则 (c) 图结构规则

注:实线关系及其所连接的节点共同表示规则体,虚线关系及其所连接的节点共同表示规则头,链式规则中的节点不作标签标注。

Fig. 1 Examples of rule pattern

图 1 规则形式示例

在技术层面上,现有图结构规则学习多依赖支持度与置信度的统计框架,通常通过先枚举候选规则,再进行有效性验证的方式,从图中挖掘潜在模式。随着图谱规模的扩大,这类方法的计算复杂度呈指数级增长;同时,由于仅能孤立处理单条规则,难以刻画规则间的语义关联,从而在准确性与效率上都受到限制。

为应对上述挑战,本文提出了一种基于对比学习的图结构规则学习模型 CGRL(contrastive graph rule learning)。该模型摒弃了传统方法对预定义支持度等统计指标的依赖,利用神经网络直接从图数据中学习规则模式。具体而言,CGRL 首先通过采样获取潜在子图模式,并构建“规则体-规则头”对作为正样本,同时设计定制化的负采样策略以生成无效规则对。随后,模型在编码器与投影头形成的向量空间中,利用对比损失函数最大化有效规则对的相似

度、最小化无效规则对的相似度，从而有效捕捉图结构中的内在语义特征。本文的贡献总结如下：

1）面向知识图谱中复杂的图结构规则形式，提出了一种基于对比学习的图结构规则学习方法CGRL，克服了传统链式规则在表达能力与应用范围上的局限性。

2）为应对数据稀疏下低频但高价值规则难以学习的问题，CGRL引入自适应节点聚类生成软标签以增强语义约束，并结合多跳邻域采样构建语义子图与规则样本；同时采用结构感知Transformer编码器与对比学习框架，联合建模规则的结构特征与语义信息。

3）在6个公开数据集上进行了实验，验证了CGRL在图结构规则学习中的有效性。实验结果表明，CGRL在多个复杂数据集上表现优异，并在小规模数据集上达到与最先进方法相当的效果，充分证明了高质量图结构规则在知识图谱推理中的潜力以及CGRL学习此类规则的能力。

1 相关工作

1.1 基于统计特征的规则挖掘方法

传统规则挖掘最早发展在关系数据库领域，通过统计特征发现频繁模式与关联规则[23]。随着图数据研究的深入，关联规则挖掘方法逐步扩展至图数据领域，当前图关联规则方法通常先对搜索空间进行剪枝过滤，再枚举候选图模式，最后利用子图同构方法[24]进行图模式匹配，并通过剪枝和并行优化提升挖掘效率。

在知识图谱规则学习领域，现有研究大多聚焦于霍恩规则[25]。典型代表如AMIE[19]，采用自底向上的搜索策略，从单一关系逐步扩展到复杂规则，并通过支持度与置信度评估规则质量，分别量化模式的普遍性与预测可靠性。其优化版本AMIE+[20]针对大规模知识图谱引入动态剪枝机制与并行计算框架，并提出基于局部完整性假设的PCA置信度模型，从而在提升效率的同时显著增强了评估精度。另一代表性方法AnyBURL[26]同样采用自底向上的生成策略，并结合启发式搜索与候选筛选机制，能够在有限时间内高效挖掘高质量规则。

尽管上述方法在知识图谱规则挖掘中取得了一定进展，但仍存在明显局限性。首先，它们过于聚焦传统的链式规则，难以有效建模图中的复杂结构关联。虽然已有部分工作尝试引入图关联规则[27]，其方法仍主要基于显式的图模式匹配或频繁子图挖掘，而非通过图神经网络等图学习范式来获取结构表征，因此难以捕捉复杂结构特征并支持可微的规则学习。其次，现有方法往往过度依赖支持度与置信度等统计指标，难以保证规则的语义合理性；其特征设计主要集中于实例频次，忽视了规则本身的逻辑与语义结构，从而难以识别那些虽在数据中罕见但在逻辑上成立的潜在规则。

1.2 基于神经网络的规则学习方法

随着神经网络技术的发展，越来越多的研究尝试将神经网络引入规则学习领域，以提升效率与表达能力。这类方法多采用基于查询三元组的监督学习范式：给定目标三元组，以头实体与关系为输入，模型通过参数化生成候选规则体，并利用矩阵连乘组合路径推理尾实体。训练阶段则通过最小化预测分布与真实分布间的损失进行参数更新。Neural-LP[12]首创端到端可微分规则学习框架，通过Tensorlog操作符结合注意力机制学习可变长度规则，从而提升规则搜索效率，但在原子共享情况下容易生成高置信度的无意义规则。为此，DRUM[28]采用双向RNN对潜在错误规则体进行剪枝，更有效地捕捉规则体中的上下文信息，从而过滤无效规则。NLIL[18]提供了更灵活的规则表示形式，可发现更复杂的逻辑关系。Ruleformer[16]在学习过程中引入头实体信息及子图上下文，利用Transformer特征提取能力捕捉实体间复杂关系，但其模型复杂性限制了大规模图谱的应用。IterE[29]引入迭代增强机制，提升稀疏实体嵌入质量。RNNLogic[13]提出2阶段学习范式：规则生成器产生候选规则，推理预测器评估其质量，并结合强化学习策略进行优化，以提高规则搜索效率。DiffLogic[30]则提出神经符号推理框架，通过三元组过滤器选择必要事实，并结合概率软逻辑实现端到端可微分学习。

近年来，RLogic[14]和NCRL[15]提出新范式，将规则置信度定义为规则体与规则头的语义一致性，均通过闭合路径采样和交叉熵优化。RLogic基于语义相似度避免传统频率统计缺陷；NCRL引入滑动窗口机制整合规则，增强规则间关联建模。MPRM[31]将规则推理建模为马尔可夫链，通过聚合路径概率高效计算置信度，显著降低计算开销。

尽管上述方法在知识图谱推理中取得一定进展，但它们依赖有限监督信号（如查询三元组或闭合路径），在知识稀疏或不完备情况下难以充分捕捉语义特征，限制规则学习的全面性。随着大语言模型的

发展, ChatRule[32] 利用大语言模型构建规则生成与排序模块, 有效挖掘逻辑规则并提升推理性能与可解释性, 但当前大语言模型更倾向于生成简单链式规则, 而非复杂图结构规则。

1.3 基于对比学习的知识图谱嵌入方法

对比学习(contrastive learning, CL)通过最大化正样本相似度和最小化负样本相异度, 学习强区分性特征表示。SimKGC[33] 提出了包括 In-Batch, Pre-Batch, Self-Negative 在内的 3 种负采样策略, 以增强嵌入表示的判别性与鲁棒性; KGCL[2] 构建拓扑去噪框架, 利用跨视图对比目标过滤噪声边缘, 提升推荐系统性能。总体来看, 对比学习不仅优化局部表示, 还通过结构对齐强化全局语义建模, 为推理任务提供更可靠、泛化性更强的特征基础。

2 预备知识

2.1 知识图谱

知识图谱是一种结构化的知识表示形式(如图 2 所示), 形式化定义为

$$G=\{(v,r,v')|v,v'\in\mathcal{E},r\in\mathcal{R}\},$$

其中$\mathcal{E}$和$\mathcal{R}$分别代表实体集合和关系集合。三元组(v,r,v')表示头实体v与尾实体v'间存在关系r。

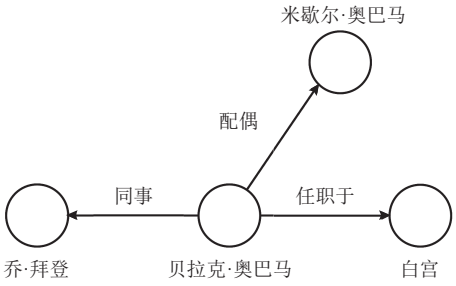


Fig. 2 Illustration of knowledge graph

图 2 知识图谱示意图

由于知识图谱中的实体通常缺乏显式标签, 本文设计了软标签机制: 通过训练映射函数L, 为每个节点$v\in\mathcal{E}$分配标签$L(v)$, 从而在规则学习中显式引入语义约束。

2.2 逻辑规则

2.2.1 谓 词

谓词是逻辑表达式中的基本单元, 用以表述实体间关系或属性。本文聚焦于边谓词, 其表示形式为(x,r,y): 表示实体x与实体y存在关系r。

2.2.2 霍恩规则

知识图谱中的逻辑规则通常以霍恩规则[25](Horn rules)的形式进行建模, 霍恩规则一般由规则体和规则头 2 部分组成, 其中规则体表示推理成立所需满足的前提条件, 规则头表示在规则体条件满足时可以推出的目标关系。定义将第 i 个规则体谓词表示为$r_{B_i}(x_i)$, 相应地, 规则头谓词表示为$r_H(x)$, 霍恩规则的一般形式化定义为

$$r_{B_1}(x_1)\wedge\cdots\wedge r_{B_m}(x_m)\rightarrow r_H(x),$$

其中规则头是一个单一正谓词, 规则体由若干不含否定或析取的谓词构成。在知识图谱推理场景中, 本文进一步将霍恩规则限定为仅由二元关系谓词(边关系)构成的规则, 即规则头为一个二元关系, 规则体由若干二元关系构成。

2.2.3 链式逻辑规则

链式逻辑规则是霍恩规则在知识图谱推理中的一种常见特例, 因易于理解和实现而被广泛应用。其规则体通常由多个二元关系谓词通过共享变量依次连接, 形成一条关系路径, 规则头则为单一目标关系。类似地, 链式规则$\rho(r_H,r_B)$形式化表示为

$$r_H(X,Z)\leftarrow r_{B_1}(X,Y_1)\wedge r_{B_2}(Y_1,Y_2)\cdots\wedge r_{B_m}(Y_{m-1},Z),$$

其中r_H是特定关系, 称为规则头; $r_B=\bigwedge_{i=1}^{m}r_{B_i}$表示规则体中的谓词合取形式, 由 m 个按序连接的谓词构成; $r_{B_i}(\cdot,\cdot)$表示规则体中的第 i 个关系谓词; $X,Z,Y_i\in\mathcal{E}$表示具体实体的实体变量。链式规则的推理过程通过顺序应用规则体中的关系谓词, 实现由起点实体到目标实体的多跳推导。例如, 图 1(a)中的链式规则可表示为: 若用户 A 关注用户 B, 且用户 B 喜欢物品 C, 则可推断用户 A 对物品 C 存在潜在兴趣。

2.2.4 图结构逻辑规则

图结构逻辑规则是一种用于捕捉知识图谱中复杂关系和约束的规则形式。本研究关注单一目标关系的图结构规则, 其规则体由一个带标签的模式子图构成, 规则头为从该模式中可推理出的目标边。

一个图结构规则$\rho(Q,e)$的形式化定义为:

$$\rho(Q,e)\colon Q\rightarrow e,$$

其中$Q=\{\mathcal{E}_Q,\mathcal{R}_Q,L_Q\}$,表示带标签的图模式, $\mathcal{E}_Q$与$\mathcal{R}_Q$分别为模式中的节点集合和边集合, L_Q为节点标签映射函数, 将实例节点映射到模式节点。规则体Q中的每条边谓词可定义为五元组

$$(u_i,r_i,v_i,L_Q(u_i),L_Q(v_i)),$$

其包含头节点u_i及其标签$L_Q(u_i)$、关系r_i、尾节点v_i及其标签$L_Q(v_i)$。

当Q在知识图谱G中存在满足标签约束与结构

约束的匹配时(即存在图同态或子图匹配)，即可根据规则ρ推理出目标边e的存在。例如，图 1(c)中的图结构规则表示：若用户 A 关注用户 B，用户 B 关注用户 C，且用户 B 与用户 C 均与推文 D 发生评论交互，则可推断用户 A 对用户 C 存在潜在关注关系。

相比只能描述单一路径关系序列的链式规则，图结构规则允许规则体中包含分叉或汇合的关系结构，使多条关联路径可以共同约束同一规则结论，从而刻画多个实体关系联合影响推理结果的情形。

本文使用的主要符号定义对照见表 1。

Table 1 Definitions of the Main Symbols Used in Our Paper

表 1 本文主要符号定义

符号	含义	符号	含义
G	知识图谱	$\mathcal{E}$	实体集合
$\mathcal{R}$	关系集合	u,v	实体节点变量
r	关系	$e=(u,r,v)$	由头实体 u、关系 r 和尾实体 v 构成的三元组
m	最大边数	$\boldsymbol{R}_r$	关系 r 的嵌入表示
n	节点最大度	l	图深度
k	规则数量	h	跳数限制
H	规则头	B	规则体
$\boldsymbol{R}_H$	规则头嵌入表示	$\boldsymbol{R}_B$	规则体嵌入表示
$\boldsymbol{R}_v$	节点v嵌入表示	$\boldsymbol{R}_e$	三元组e嵌入表示
τ	温度参数	L_Q	标签映射函数
DC_v	度中心性	CV	聚类变异系数
Q_{sub}	模式子图	ψ	实例到模式的映射函数

3 基于对比学习的图结构规则学习模型

3.1 问题定义

图结构规则学习旨在从知识图谱中提取满足结构约束的高质量规则集合。设规则的候选空间为$\Phi=\{\rho_i\}$，为避免搜索空间爆炸，本文对规则结构施加以下约束：规则体最大边数m、规则体最大节点度n、规则体最长路径深度l。$\mathcal{E}(\rho)$表示规则ρ中涉及的实体集合，节点$v\in\mathcal{E}(\rho)$表示规则体结构模式中的实体节点。即本文仅考虑满足

$$|\mathcal{R}(\rho)|\leqslant m,\max_{v\in\mathcal{E}(\rho)} deg(v)\leqslant n,depth(\rho)\leqslant l \tag{1}$$

的约束条件下，学习一个参数化的置信度函数

$$\alpha_\theta:\rho\to[0,1],$$

用于评估每条规则的可靠性，直观目标是在满足结构约束的候选集合上，使高质量规则获得更高的置信度，从而支持后续规则排序与筛选。该目标可表示为

$$\max_\theta \sum_{\rho\in\Phi}\alpha_\theta(\rho),$$

其中θ为模型参数，$\mathcal{R}(\rho)$和$\mathcal{E}(\rho)$分别为规则ρ的边集与节点集，$deg(v)$表示节点度，$depth(\rho)$表示规则中最深路径的长度。式(1)仅用于刻画规则置信度学习的总体优化方向，实际训练过程通过构造正负规则样本并优化对比损失来实现。

为了优化求解，本文通过特征映射将规则体B与规则头H投影到统一语义空间，记规则体和规则头的嵌入表示分别为$\boldsymbol{R}_B$和$\boldsymbol{R}_H$，目标是通过规则体表示与规则头表示之间的语义一致性度量来量化$\alpha_\theta(\rho)$。具体训练时，通过构造对比损失函数L_θ来优化参数θ，从而刻画规则体和规则头的语义一致性：

$$\min_\theta L(\theta)=\min_\theta\sum_{\rho\in\Phi}\ell(\alpha_\theta(\boldsymbol{R}_B(\rho),\boldsymbol{R}_H(\rho))),$$

其中$\ell(*)$是使用的对比损失函数，$\boldsymbol{R}_B(\rho)$和$\boldsymbol{R}_H(\rho)$分别表示候选规则ρ的规则体与规则头在统一语义空间中的嵌入表示。由此，模型能够通过衡量规则体与规则头在结构和语义层面的相似性，引导二者在表征空间中形成一致对齐关系。

3.2 模型描述

当前图结构规则学习面临着以下 3 个主要挑战：首先，规则学习通常依赖节点的类型来区分其在结构中的角色，但现实中大多数知识图谱缺乏显式的类型标注，仅提供实体 ID 与关系三元组，导致节点语义标签稀缺；其次，与仅需采样路径的链式规则不同，图结构规则需要构建包含多个节点和关系的语义子图，而传统随机采样在效率、覆盖率与信噪比方面均难以满足需求；最后，图结构规则的表示学习需要同时捕捉拓扑结构和上下文语义，然而现有方法多采用简单表示形式进行建模，难以捕捉高阶交互与复杂依赖。

针对以上挑战，本文提出了 CGRL 模型，其整体架构如图 3 所示。首先，CGRL 在节点分类阶段，对于未标注的三元组数据，利用 Node2Vec[34] 将节点映射到低维向量空间，并通过自适应聚类方法进行无监督聚类，自动确定类别数并生成软标签；对于已标注的数据，则直接使用原有标签信息。随后，在规则样本构建阶段，模型采用多跳邻域采样策略获取满足约束条件的子图，将实例级子图抽象为模式级子图，通过遮挡目标边构造规则体-规则头的正样本，并结合关系替换与结构扰动生成语义与拓扑负样本。

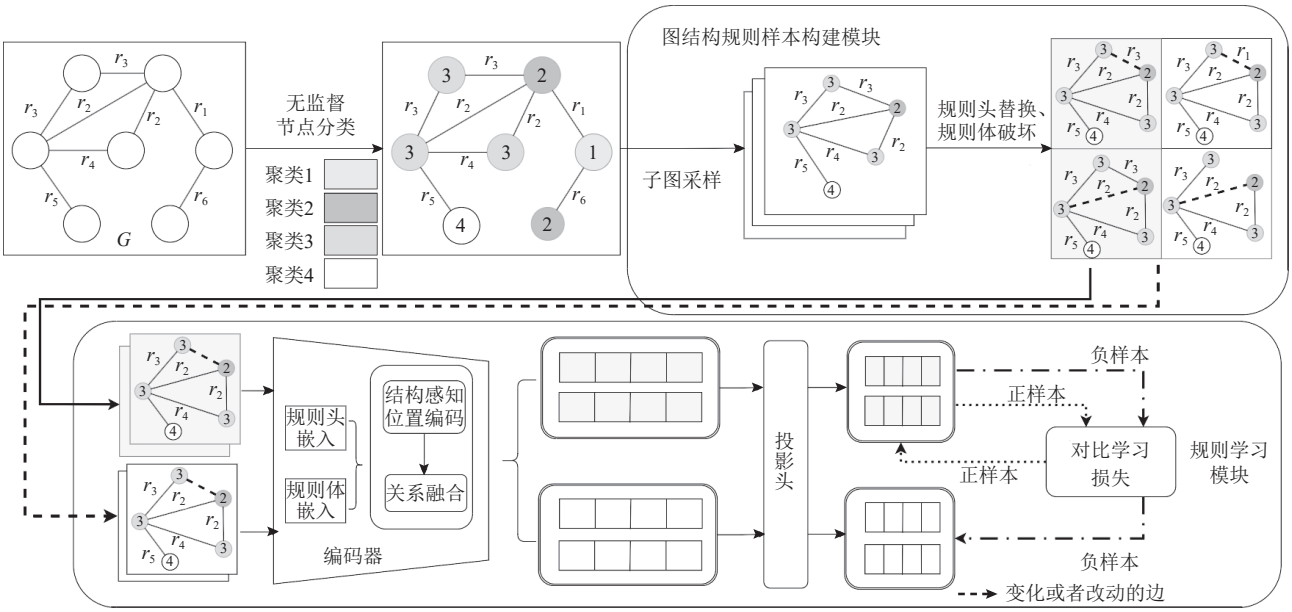


Fig. 3 Framework of CGRL model

图 3 CGRL 模型框架

最后, 在规则表示学习阶段, CGRL 设计了基于 Transformer 的编码框架, 分别建模规则体与规则头的向量表示, 并通过对比学习最小化正样本对距离以及最大化负样本对距离, 从而实现对规则置信度的量化估计。完成训练后, 模型根据置信度对候选规则排序, 筛选出前k条高质量规则用于后续推理任务。

3.3 无监督节点分类机制

在知识图谱规则学习中, 节点表示质量直接影响规则匹配与推理性能, 但节点普遍缺乏明确的类型标签, 不同图谱在规模、结构复杂度和关系类型上的差异也使固定聚类方案难以适配, 从而削弱了语义一致性并限制了规则挖掘的可解释性。为解决这一问题, CGRL 提出融合自适应聚类与软标签生成的无监督节点分类机制, 通过自适应方式确定类别并为节点生成鲁棒的语义标签, 从而提升规则学习的泛化能力与解释性。

在具体实现中, CGRL 首先针对不同知识图谱动态确定聚类数量。直接设定固定聚类数往往会造成过拟合或欠拟合, 因此 CGRL 采用基于肘部法的 *K*-means 策略[35] 对节点嵌入进行聚类, 其目标函数为

$$\sum_{j=1}^{k}\sum_{i=1}^{n}|x_i^{(j)}-c_j|^2,$$

其中k为簇的数量, n为数据点数量, $x_i^{(j)}$表示属于第j个簇的第i个数据点, c_j为第j个簇的质心。为了确定最优聚类数, CGRL 通过分析不同k下的组内平方和 *WSS*(within sum of square)变化趋势来识别拐点:

$$WSS(k)=\sum_{i=1}^{k}\sum_{x\in C_i}\|x-\mu_i\|^2,$$

其中C_i表示第i个簇, μ_i为该簇质心, $\|x-\mu_i\|^2$为样本点到簇质心的欧氏距离平方。

完成最优聚类数的确定后, CGRL 进一步评估聚类均衡性。其采用变异系数(coefficient of variation, *CV*)来衡量簇大小的均衡性:

$$CV=\frac{\sigma}{\mu},$$

其中σ和μ分别为簇大小的标准差和平均值。理想情况下$CV<0.5$, 若 *CV* 过高, 则说明聚类结果存在严重不均衡, CGRL 会重新调整聚类数量并执行 *K*-means, 直至获得均衡的聚类划分。

通过上述机制, CGRL 在无需人工标注的条件下为节点生成语义软标签, 这些软标签被引入规则表示学习的训练目标中, 作为额外的语义约束信号, 不仅提升了模型对节点类别信息的感知能力, 还增强了规则推理的泛化性与一致性。

3.4 图结构规则样本构建模块

在无监督节点分类模块中, CGRL 已通过语义聚类获得节点类型标签, 这些软标签L_Q将作为本模块的关键输入, 用于实现构建图结构规则样本。然而, 直接从大规模知识图谱中学习图结构规则面临以下挑战: 1)计算效率瓶颈, 在全图上进行模式挖掘计算开销巨大, 难以扩展; 2)模式泛化需求, 需要从具体实例中抽象出通用规则模式, 而非记忆特定三元组。为解决这些问题并提升规则学习的泛化能力, CGRL

设计了一个高效的 2 阶段规则样本构建流程：子图采样提取局部结构，规则数据生成抽象模式并构造正负训练样本。

3.4.1 子图采样

为克服全图计算的高昂开销，CGRL 在设计上借鉴了 GraphSAGE[36]，采用基于采样的局部学习策略。其核心思想是通过采样代表性子图作为训练单元，在保留关键结构信息的同时显著降低计算复杂度。然而，简单的随机采样容易遗漏重要结构模式。为此，CGRL 引入基于节点重要性的加权采样机制，并辅以并行化加速策略。

在节点重要性评估方面，CGRL 使用 PageRank[37] 计算每个节点 v 的全局重要性 $PR(v)$，形式化定义为

$$PR(v)=\frac{1-d}{N}+d\sum_{u\in\mathcal{N}(v)}\frac{PR(u)}{deg(u)},$$

其中 d 为阻尼因子（通常取值为 0.85），N 为节点总数，$\mathcal{N}(v)$ 为节点 v 的邻居集合，$deg(u)$ 为节点 u 的度。基于重要性分数，CGRL 定义加权邻居采样概率：对于任意起始节点 v 及其邻居 $u\in\mathcal{N}(v)$，其采样概率定义为

$$P(u|v)=\frac{PR(u)/deg(u)}{\sum_{x\in\mathcal{N}(v)}PR(x)/deg(x)}。$$

在节点重要性相近的情况下，该策略优先选择低度节点，以提升对稀疏区域及潜在结构模式的覆盖能力。

此外，为解决大规模数据下图计算瓶颈，CGRL 还设计了并行化采样方法。具体做法是将节点集合划分为若干子集，并在多个独立处理单元上并行执行采样。由于子图采样任务之间相互独立，仅需在最终阶段合并结果，因此无需频繁通信，就能够高效并行。整个过程由主控节点协调，负责数据分区与任务分配，各处理单元完成采样后将结果返回主控节点进行汇总，从而生成完整的采样结果集。

3.4.2 图结构规则训练数据生成

CGRL 通过采样子图得到实例级子图集合 $\{G_{\text{sub}}=(\mathcal{E}_{\text{sub}},\mathcal{R}_{\text{sub}})\}$，但直接基于实例进行学习会导致规则过拟合且缺乏泛化性，难以迁移到未见结构。因此需要从每个 G_{sub} 中抽象出其对应的模式子图 Q_{sub}，以实现结构级泛化。

具体来说，CGRL 首先将实例子图转化为通用模式，使得不同子图的相似结构能映射到同一模式，支撑规则泛化。具体而言，CGRL 利用无监督分类模块的节点标签，通过标签映射函数 L_Q 将 $\mathcal{E}_{\text{sub}}$ 映射到语义标签空间。然后对节点和关系进行匿名化重编号，消除实例 ID 的干扰，得到模式子图 $Q=(\mathcal{E}_Q,\mathcal{R}_Q,L_Q)$。通过该模式化转换，相似结构将被归一到共享模式，以支持规则的泛化学习。

在获得模式子图后，CGRL 设计随机规则头替换策略和随机规则体破坏策略，生成具区分度的正负样本：正样本通过从模式 Q 中随机移除一条边 $e=(v,r,v')$ 得到规则体 Q'，并由此构建规则 $\rho^+=Q'\to e$，其中规则体 Q' 编码模式约束，规则头 e 代表待预测关系。该过程可重复 k 次以增强样本多样性；负样本则通过多种方式生成，包括将规则头替换为不存在的三元组、对规则体进行结构扰动（如增删边）以及替换节点语义标签。生成负样本候选集后，通过冲突检测与过滤以剔除无效负例，从而保证样本质量。

最终，每个子图 G_{sub} 通过映射函数 ψ 转化为规则训练集：

$$Q_{\text{sub}}=\psi(G_{\text{sub}})=\{\{\rho^+\},\{\rho^-\}\}。$$

该方法为每个采样子图构造规则样本，其中规则头由一条边表示，规则体由子图中其余结构组成。由此生成的规则样本为规则学习模块提供了结构化训练数据。

3.5 规则学习模块

尽管前文通过子图采样与规则生成构建了训练样本，但这些规则仅停留在符号模式层面，缺乏可解释的量化指标来支撑推理与补全决策。传统的链式规则方法往往依赖频率统计计算置信度，难以捕捉复杂图结构与语义依赖，从而限制了规则在实际知识图谱中的适用性。因此，如何在保留规则结构可解释性的同时，为规则分配合理的置信度估计，以提升规则推理的可靠性。

为应对上述挑战，CGRL 设计了基于 Transformer 架构的规则学习模块，其目标是为每条规则 ρ 估计置信度 $c_\rho\in[0,1]$，量化规则体 Q' 蕴涵规则头 e 的概率。该模块主要由 3 个部分组成。首先，编码器通过结构感知位置编码与 Transformer 建模规则体和规则头；其次，投影头将嵌入映射到低维对比学习空间；最后，损失函数通过 InfoNCE 对齐规则体与规则头，从而实现置信度估计与概率解释。

3.5.1 编码器

编码器将规则 $\rho=Q'\to(u,r,v,L_Q(u),L_Q(v))$ 中的子图 Q' 与规则头 $(u,r,v,L_Q(u),L_Q(v))$ 转换为嵌入表示，用于支持规则置信度的计算。本文令 $L_Q(x)=x$，即直接使用节点替代其标签作为初始特征向量，从而在规则编码过程中引入节点类别信息。

传统的正余弦位置编码是针对线性序列设计的,在知识图谱场景下存在2方面不足:1)序列位置编码无法区分标签相同但局部结构不同的节点,导致结构信息丢失;2)将图强行线性化会引入虚假的位置关系,可能破坏原有拓扑结构的语义一致性。为此,本文提出结构感知位置编码,即在位置表示中引入图的拓扑特征,以增强节点区分能力。其目标在于使规则表示不仅捕捉语义信息,还能够反映细粒度的结构差异,从而提高规则推理的准确性与鲁棒性。

1)节点级结构编码。对于给定的节点v,提取其局部拓扑特征,包括节点度、局部聚类系数和归一化度中心性,然后将这些结构特征输入多层感知机进行非线性映射,生成结构感知位置向量:

$$d_v=\sum_{u\in\mathcal{E}}a_{v,u},\ c_v=\frac{2E_v}{d_v(d_v-1)},\ DC_v=\frac{d_v}{\sum_i d_i},$$

$$\boldsymbol{v}^{\mathrm{pos}}=ReLU(\boldsymbol{W}[d_v,c_v,DC_v])+\boldsymbol{v},$$

其中d_v表示节点的度且$d_v>2$(当d_v为0或1时c_v=0),$a_{v,u}$是邻接矩阵中的第v行、第u列的元素,表示节点v与节点u之间的连接关系,当节点v与节点u存在边连接时取值$a_{v,u}=1$,否则$a_{v,u}=0$。E_v是节点v的邻居节点之间实际存在的边数,c_v是局部聚类系数,DC_v是归一化度比例,$\boldsymbol{W}$是可学习权重矩阵,$\boldsymbol{v}$是节点的初始特征向量,$\boldsymbol{v}^{\mathrm{pos}}$是节点的结构编码。该方式流程是将$d_v$, c_v, DC_v拼接成特征向量$[d_v,c_v,DC_v]$,通过权重矩阵$\boldsymbol{W}$进行线性变换后,使用$ReLU$激活函数引入非线性表达能力,最终与节点原始特征$\boldsymbol{v}$相加得到结构编码$\boldsymbol{v}^{\mathrm{pos}}$,最终的结构编码$\boldsymbol{v}^{\mathrm{pos}}$在保留原始语义信息的基础上引入局部拓扑差异,使得具有相同标签但结构不同的节点能够被有效区分。

2)关系语义融合。将关系信息融入实体中,得到单条边的嵌入表示,并分别对规则头H和规则体B进行嵌入。具体来说,对于每个三元组$e=(u,r,v)$,记头实体u和尾实体v的结构编码分别为$\boldsymbol{u}^{\mathrm{pos}}$, $\boldsymbol{v}^{\mathrm{pos}}$,关系$r$的嵌入表示为$\boldsymbol{r}$,则三元组的嵌入表示$\boldsymbol{e}$由头实体与尾实体的结构编码$\boldsymbol{u}^{\mathrm{pos}}$, $\boldsymbol{v}^{\mathrm{pos}}$和关系嵌入$\boldsymbol{r}$相加得到:

$$\boldsymbol{e}=\boldsymbol{u}^{\mathrm{pos}}+\boldsymbol{r}+\boldsymbol{v}^{\mathrm{pos}}。$$

3)规则体语义建模。规则体$B=Q$是由多个边组成的图模式。对于规则体的语义建模,CGRL按照子图采样阶段确定的拓扑顺序,将规则体中的边表示组织成序列矩阵$\hat{\boldsymbol{E}}=(\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n)\in\mathbb{R}^{n\times d}$,并将边表示序列输入Transformer编码器进行结构语义建模。其中,$\boldsymbol{e}_i$表示第i条边的嵌入表示。

CGRL的Transformer编码器由多个堆叠的编码层组成,每个编码层主要由多头自注意力机制和前馈神经网络构成。具体来说,给定模式规则体输入序列$\hat{\boldsymbol{E}}=(\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n)\in\mathbb{R}^{n\times d}$,每个注意力头对输入序列进行线性映射得到查询矩阵$\hat{\boldsymbol{Q}}$、键矩阵$\hat{\boldsymbol{K}}$和值矩阵$\hat{\boldsymbol{V}}$:

$$\hat{\boldsymbol{Q}}=\hat{\boldsymbol{E}}\boldsymbol{W}_{\hat{Q}},\ \hat{\boldsymbol{K}}=\hat{\boldsymbol{E}}\boldsymbol{W}_{\hat{K}},\ \hat{\boldsymbol{V}}=\hat{\boldsymbol{E}}\boldsymbol{W}_{\hat{V}},$$

其中$\boldsymbol{W}_{\hat{Q}},\boldsymbol{W}_{\hat{K}},\boldsymbol{W}_{\hat{V}}\in\mathbb{R}^{d\times d_k}$是可学习的权重矩阵,$d_k$是每个注意力头的维度。

因此,每个注意力头的输出为

$$\boldsymbol{head}=softmax\left(\frac{\hat{\boldsymbol{Q}}\hat{\boldsymbol{K}}^{\mathrm{T}}}{\sqrt{d_k}}\right)\hat{\boldsymbol{V}}。$$

为了更好地理解复杂图结构中节点和关系之间的多层次关系,提高模型的表达能力和泛化性能,CGRL采用多头自注意力机制,通过并行计算多个注意力头,并将它们的输出拼接在一起,得到多头注意力结果:

$$\boldsymbol{MultiHead}(\hat{\boldsymbol{Q}},\hat{\boldsymbol{K}},\hat{\boldsymbol{V}})=(\boldsymbol{head}_1\oplus\boldsymbol{head}_2\oplus\cdots\oplus\boldsymbol{head}_m)\boldsymbol{W},$$

其中m是注意力头的数量,$\oplus$是连接操作符,$\boldsymbol{W}\in\mathbb{R}^{md_k\times d}$是可学习的权重矩阵。

为防止网络退化并缓解梯度消失,使模型更易训练,多头自注意力机制的输出进行残差连接[38]和层归一化(LN)处理:

$$\hat{\boldsymbol{E}}'=LN\left(\hat{\boldsymbol{E}}+\boldsymbol{MultiHead}\left(\hat{\boldsymbol{Q}},\hat{\boldsymbol{K}},\hat{\boldsymbol{V}}\right)\right)。$$

随后,前馈网络结合线性变换、$ReLU$激活、残差连接和层归一化,其计算过程表示为

$$\boldsymbol{Z}=LN\left(\hat{\boldsymbol{E}}'+ReLU\left(\hat{\boldsymbol{E}}'\boldsymbol{W}_1+\boldsymbol{b}_1\right)\boldsymbol{W}_2+\boldsymbol{b}_2\right),$$

其中$\boldsymbol{W}_1,\boldsymbol{W}_2\in\mathbb{R}^{d\times d}$为可学习权重矩阵,$\boldsymbol{b}_1,\boldsymbol{b}_2\in\mathbb{R}^d$为偏置项。将上述结构堆叠$L$层,构成编码器,可获得规则体的深层语义编码。

为了将可变长度的规则体序列统一为固定维度表示,CGRL对规则体中各谓词的上下文嵌入表示进行平均池化。记$\boldsymbol{Z}_i$为规则体中第i个谓词经过编码器得到的上下文表示,则规则体整体表示$\boldsymbol{R}_B$定义为

$$\boldsymbol{R}_B=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i,$$

其中n表示规则体中谓词的数量。平均池化将可变长度序列聚合为固定维度向量,在保留全局结构语义的同时,能够适配不同规模的规则体。

对于规则头三元组(u,r,v),CGRL将三元组(u,r,v)的实体嵌入$\boldsymbol{u}^{\mathrm{pos}}$, $\boldsymbol{v}^{\mathrm{pos}}$和关系嵌入$\boldsymbol{r}$相加,得到规则头表示$\boldsymbol{R}_H$:

$$\boldsymbol{R}_H = \boldsymbol{u}^{\text{pos}} + \boldsymbol{r} + \boldsymbol{v}^{\text{pos}} \text{。}$$

至此，模型编码器输出规则体表示$\boldsymbol{R}_B$与规则头表示$\boldsymbol{R}_H$均为d维向量。

3.5.2 投影头设计

CGRL中的投影头由线性层与激活函数构成，用于将Transformer编码器输出的高维表示映射到低维语义空间。通过特征空间变换，它可将规则体与规则头的嵌入有效压缩到目标任务所需的低维空间，使编码结果更好地适配后续规则学习。

在CGRL中，将规则体和规则头的嵌入表示$\boldsymbol{R}_B$和$\boldsymbol{R}_H$通过线性层、*ReLU*激活函数和Dropout层进行处理，得到投影后的表示$\boldsymbol{P}_{\boldsymbol{R}}$，投影过程定义为：

$$\boldsymbol{P}_{\boldsymbol{R}} = Dropout(ReLU((\boldsymbol{R})\boldsymbol{W}_p + \boldsymbol{b}_p), p)\text{，}$$

其中$\boldsymbol{W}_p \in \mathbb{R}^{d \times d_p}$是可学习的权重矩阵，$\boldsymbol{R}$为$\boldsymbol{R}_B$或$\boldsymbol{R}_H$（对应投影后的表示分别为$\boldsymbol{P}_B$与$\boldsymbol{P}_H$），$\boldsymbol{b}_p \in \mathbb{R}^{d_p}$是偏置向量，$d_p$是投影后的维度，$p$是Dropout的概率。

3.5.3 损失函数设计

经过结构感知位置编码、Transformer编码器和投影头处理后，分别得到规则体和规则头投影后的向量表示$\boldsymbol{P}_B$与$\boldsymbol{P}_H$。为实现规则体投影后表示和规则头投影后表示在语义空间中的对齐，CGRL采用InfoNCE对比损失：对于一条真实规则，其规则体投影表示与对应规则头投影表示构成正样本对；对于不匹配的规则体与规则头组合，则构成负样本对。给定正样本对$(\boldsymbol{P}_B^+, \boldsymbol{P}_H^+)$及$N$条负样本对集合$\{(\boldsymbol{P}_B^+, \boldsymbol{P}_H^{j-})\}_{j=1}^N$，定义正负样本对相似度分别为$s^+ = sim(\boldsymbol{P}_B^+, \boldsymbol{P}_H^+)$，$s_j^- = sim(\boldsymbol{P}_B^+, \boldsymbol{P}_H^{j-})$，则损失函数定义为

$$\mathcal{L}_{\text{InfoNCE}} = -\ln \frac{\exp(s^+/\tau)}{\exp(s^+/\tau) + \sum_{j=1}^{N} \exp\left(s_j^-/\tau\right)}\text{，}$$

其中τ为温度系数，N为负样本数量。从互信息视角，InfoNCE可视作$I(\boldsymbol{P}_B; \boldsymbol{P}_H)$的一个下界估计，其期望形式近似满足：

$$I(\boldsymbol{P}_B; \boldsymbol{P}_H) \geqslant \ln(N+1) - \mathcal{L}_{\text{InfoNCE}}\text{，}$$

其中$I(\boldsymbol{P}_B; \boldsymbol{P}_H)$是$\boldsymbol{P}_B$和$\boldsymbol{P}_H$之间的互信息，$N$是负样本数量。最小化该损失能够提高正样本对的相似度，降低负样本对相对于正样本对的相似度权重，从而增强规则体与规则头之间的语义一致性和匹配区分能力。

进一步地，规则体与规则头之间的相似度得分可用于刻画规则置信度。经过温度系数缩放与指数变换后，用以表示给定规则体B时规则头H成立的未归一化置信度：

$$p(H|B) \propto \exp\left(\frac{sim(\boldsymbol{P}_B, \boldsymbol{P}_H)}{\tau}\right)\text{。}$$

因此，CGRL通过规则学习模块捕捉复杂的图结构与关系模式，并为规则分配概率化置信度，以区分不同规则在推理中的贡献程度。

3.6 模型训练和时间复杂度分析

3.6.1 模型训练流程

CGRL模型训练流程如下：首先，以知识图谱G为输入，借助Node2Vec进行节点表示学习，再通过自适应聚类方法对节点完成分类。接着，针对每个节点v，从其开始进行节点级子图采样，得到子图G_{sub}。之后，对每个子图G_{sub}，提取模式Q_{sub}以生成正样本$\{\rho^+\}$，同时通过破坏规则体和替换规则头关系来生成负样本$\{\rho^-\}$。最后，对于每个正样本对(B^+, H^+)和负样本对(B^-, H^-)，利用编码器计算嵌入$\boldsymbol{R}_B$和$\boldsymbol{R}_H$，计算InfoNCE损失$\mathcal{L}_{\text{InfoNCE}}$，并据此更新模型参数，最终输出训练好的CGRL模型。

3.6.2 规则生成流程

在CGRL的规则生成阶段，首先对候选规则计算置信度，并在此基础上依次采用以下筛选策略，构建高质量且多样化的规则库。

1）置信度排序。CGRL根据模型输出置信度分数对候选规则从高到低排序，形成优先级队列。

2）关系均衡化。CGRL按规则头关系类型分组，确保每种关系至少保留k条规则。

3）冗余规则消除。CGRL计算规则对(Q_i, Q_j)的相似度：

$$J(Q_i, Q_j) = \frac{|\mathcal{E}_{Q_i} \cap \mathcal{E}_{Q_j}| + |\mathcal{R}_{Q_i} \cap \mathcal{R}_{Q_j}|}{|\mathcal{E}_{Q_i} \cup \mathcal{E}_{Q_j}| + |\mathcal{R}_{Q_i} \cup \mathcal{R}_{Q_j}|}\text{，}$$

其中$\mathcal{E}_{Q_i}$表示规则Q_i中涉及的节点集合，$\mathcal{R}_{Q_i}$表示其边集合。当$J(Q_i, Q_j) > 0.7$时，视为冗余，仅保留置信度更高者；否则同时保留。

通过以上策略，CGRL输出高置信度且具备跨关系与结构适用性的规则，进而应用到下游任务。

3.6.3 时间复杂度分析

CGRL的主要计算消耗来自子图采样阶段和对比学习阶段。在子图采样阶段，对于节点数为$|\mathcal{E}|$的知识图谱，Node2Vec的时间复杂度为$O(|\mathcal{E}| \cdot l_1 \cdot w)$，其中$l_1$是随机游走长度，$w$是每个节点的游走次数。自适应聚类的时间复杂度为$O(|\mathcal{E}| \cdot k \cdot iter \cdot d_1)$，其中$k$是类别数，$iter$是迭代次数，$d_1$为节点嵌入维度。节点级子图采样的时间复杂度为$O(|\mathcal{E}| \cdot s^{len})$，其中$len$是采样深度，每一跳采样$s$个邻居，此外，若采用PageRank计算节点重要性，设PageRank的迭代次数为$iter_1$，则其

时间复杂度为$O(iter_1(|\mathcal{E}|+|\mathcal{R}|))$。

在对比学习阶段,对于批量大小为B的样本,图编码器的时间复杂度为$O(B\cdot L_1\cdot(|\mathcal{R}_{\text{sub}}|^2\cdot d+|\mathcal{R}_{\text{sub}}|\cdot d^2))$,其中$|\mathcal{R}_{\text{sub}}|$是子图中的关系数,$d$是嵌入维度,$L_1$是Transformer编码层数。投影头的复杂度$O(B\cdot d\cdot d')$,其中$d'$是投影维度。InfoNCE损失计算的复杂度为$O(B^2\cdot d')$。

因此,CGRL整体的时间复杂度为$O(|\mathcal{E}|\cdot(l_1\cdot w+k\cdot iter+s^{len})+iter_1(|\mathcal{E}|+|\mathcal{R}|)+T\cdot(B\cdot L_1\cdot(|\mathcal{R}_{\text{sub}}|^2\cdot d+|\mathcal{R}_{\text{sub}}|\cdot d^2)+B\cdot d\cdot d'+B^2\cdot d'))$,其中$T$是对比学习的迭代次数。在实际训练中,若将随机游走次数、聚类迭代次数、PageRank迭代次数、采样深度以及Transformer层数视为固定超参数,则上述复杂度可进一步近似为$O(|\mathcal{E}|\cdot s^{len}+T\cdot(B\cdot(|\mathcal{R}_{\text{sub}}|^2\cdot d+|\mathcal{R}_{\text{sub}}|\cdot d^2)+B^2\cdot d'))$。

4 实 验

4.1 实验设置

本节通过在多种公开知识图谱基准上的实验,评估CGRL在规则学习与知识补全任务中的有效性。具体的实验设置如下。

4.1.1 数据集

本文选取6个开源数据集对CGRL的性能进行全面评估,包括FB15k-237[39], WN18RR[11], Family[40], Kinship[41], UMLS[41], YAGO3-10[42],对于所有数据集,本文均采用开源数据集提供的标准训练集、验证集和测试集,不重新切分数据。各数据集的统计信息及其在CGRL中的类别数和*CV*值如表2所示。

Table 2 Dataset Statistics and Label Classification

表2 数据集统计信息与标签分类情况

数据集	实体数量	关系类型	三元组数量	类别数	*CV*值
FB15k-237	14 541	237	310 116	8	0.42
WN18RR	40 943	11	93 003	12	0.38
Family	3 007	12	28 356	6	0.33
Kinship	104	25	9 587	4	0.31
UMLS	135	46	5 960	5	0.35
YAGO3-10	123 182	37	1 089 040	15	0.45

4.1.2 对比模型和评价指标

本文选取了代表性的知识图谱嵌入方法和基于规则学习方法作为对比基线。基于嵌入的方法包括DistMult[8], RotatE[10], ConvE[11], ComplEx[9], GoldE[1]。这些方法通过将实体和关系映射到低维向量空间,利用不同的数学操作建模三元组之间的语义关系,从而实现对缺失链接的预测。

基于规则学习的方法包括Neural-LP[12], DRUM[28], RNNLogic[13], RLogic[14], NCRL[15], MPRM[31]。这些方法侧重于从知识图谱中自动学习规则,并通过端到端学习或微分优化等方式提升规则推理的效果,尤其适用于可解释性要求较高的任务。同时本节还比较了新颖的使用大模型实现知识图谱补全方法ChatRule[32]。对于上述基线方法,本文在实验过程中采用原始论文中报告的默认参数设置;其中,ChatRule采用ChatGPT作为其大模型基底。

为全面评估CGRL的性能,本文评估知识图谱补全任务,并采用平均倒数排序(MRR)与命中率(Hits@1和Hits@10)为评价指标。

4.1.3 参数设置

本文实验参数设置如下:CGRL中节点嵌入采用Node2Vec方法,随机游走长度为80,每个节点的游走次数为10,返回参数p=1.0,邻居参数q=1.0,节点嵌入维度为512,层数为3,数据采样比例为0.5。模型参数方面,隐藏层维度为512,投影头维度为128,批大小设置为32,训练轮次为20,学习率是0.001,随机丢弃率为0.1。在训练与优化过程中,模型采用分阶段优化策略。首先进行对比学习训练以获得规则表示,随后在冻结对比编码器参数的条件下进一步训练规则预测器。优化器方面,对比学习模型与规则预测模型均采用AdamW进行参数更新,并使用ExponentialLR指数学习率调度器对学习率进行衰减。

本文方法基于PyTorch实现。本文实验均在1台运行Ubuntu 20.04 LTS操作系统的高性能服务器上进行。该服务器配备了AMD Ryzen Threadripper PRO 5995WX 64核处理器、4张NVIDIA Geforce RTX 4090 24 G显卡和256 GB内存。所有实验均重复运行5次,并求平均结果。

4.2 实验结果分析

本节对CGRL在6个基准数据集上的整体表现进行系统评估,并结合各数据集的结构特性展开深入分析。实验结果如表3所示,CGRL在主流大规模基准数据集FB15k-237与WN18RR上均取得了当前最优和次优的结果。其中,CGRL在FB15k-237上的MRR与Hits@1均大幅领先,而在WN18RR上3个指标均达到最优,表明CGRL能有效捕捉复杂语义路径并具备较强的长距离依赖建模能力。在Family数据集上,CGRL的Hits@10达到99.9%,几乎接近满分,说明即使在高度结构化的家族关系推理场景下,模型依然能够有效捕捉高置信度规则。在UMLS上,

Table 3 Comparative Results Between CGRL and Other Baselines on all Datasets

表 3 CGRL 与其他基线方法在所有数据集上的对比结果

模型	FB15k-237			WN18RR			UMLS		
	MRR	Hits@1/%	Hits@10/%	MRR	Hits@1/%	Hits@10/%	MRR	Hits@1/%	Hits@10/%
ConvE	0.32	23.7	50.1	0.43	40.2	52.3	0.91	86.2	98.1
RotatE	0.34	24.1	53.3	0.48	42.8	57.1	**0.95**	91.4	**99.4**
DistMult	0.24	15.5	41.9	0.43	39.0	49.0	0.75	65.1	93.0
ComplEx	0.25	15.8	42.8	0.44	41.0	51.0	0.83	74.8	96.1
GoldE	0.37	27.7	**55.8**	0.59	52.5	61.5	0.93	**93.2**	99.2
Neural-LP	0.24	16	39.9	0.40	31.3	56.4	0.71	57.9	91.5
DRUM	0.25	18.7	37.4	0.42	39.5	56.5	0.81	67.2	97.5
RNNLogic	0.35	25.8	53.3	0.51	47.1	59.7	0.77	69.1	89.9
NCRL	0.3	20.9	47.3	0.75	67.3	88.6	0.75	67.3	94.1
RLogic	0.31	20.3	50.1	0.47	44.3	53.7	0.71	56.6	93.2
MPRM	0.33	23.4	51.8	0.54	47.8	63.2	0.37	19.4	81.2
ChatRule	0.3	20.7	50.2	0.35	30.1	40	0.78	68.5	94.8
CGRL（本文）	**0.41**	**34.2**	55.1	**0.80**	**73.4**	**91.2**	0.87	79.5	98.3

模型	Kinship			Family			YAGO3-10		
	MRR	Hits@1/%	Hits@10/%	MRR	Hits@1/%	Hits@10/%	MRR	Hits@1/%	Hits@10/%
ConvE	**0.83**	**77**	**96.7**	0.63	48.7	91.9	0.44	35.4	61.6
RotatE	0.65	50.4	93.2	0.86	78.7	93.3	0.44	40.2	67
DistMult	0.35	18.9	75.5	0.54	36.0	88.5	0.34	24.3	53.3
ComplEx	0.42	24.2	81.2	0.81	72.7	94.6	0.34	24.8	54.9
GoldE	0.80	76.3	95.2	0.85	79.0	95.2	0.59	**51.0**	**72.3**
Neural-LP	0.62	47.2	91.1	0.88	80.1	98.5	OOM	OOM	OOM
DRUM	0.58	42.3	90.1	**0.93**	**87.6**	99.7	OOM	OOM	OOM
RNNLogic	0.60	42.9	94.6	0.91	85.4	99.1	OOM	OOM	OOM
NCRL	0.60	46.3	90.5	0.91	85.2	99.3	0.38	27.4	53.6
RLogic	0.58	43.4	87.2	0.88	81.3	97.2	0.36	25.2	50.4
MPRM	0.36	17.4	85.5	0.86	76.1	99.0	**0.63**	45	**77.3**
ChatRule	0.32	17.9	55.9	0.79	70.4	89.9	0.45	40.1	58.8
CGRL（本文）	0.61	46.5	90.6	0.88	85.7	**99.9**	0.52	46.5	59.9

注：最佳结果用黑体数值标出，次优结果用下划线数值标出，OOM 表示方法内存溢出。

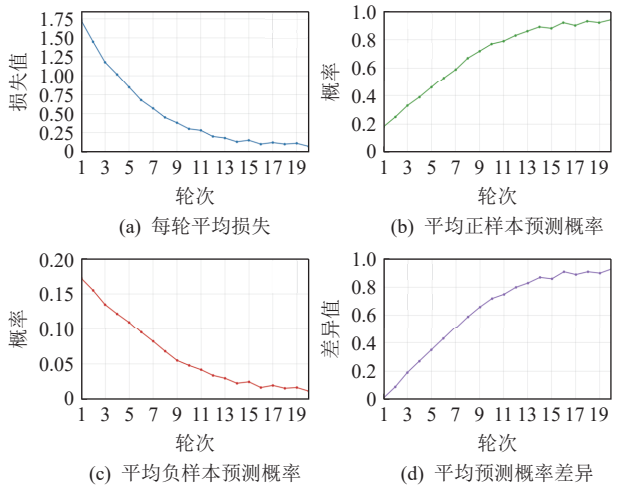
CGRL 的 MRR 为 0.87，略低于 RotatE 的 0.95，但其表现仍优于大多数神经网络方法，体现出较强的普适性。相比之下，在 Kinship 数据集上，CGRL 的 MRR 为 0.61，略逊于 ConvE（0.83）和 RotatE（0.65），表明在结构单一或关系紧凑的场景中，链式规则方法即可覆盖大部分推理需求，因此 CGRL 的优势未能完全发挥。对于 YAGO3-10 这样超大规模数据集，CGRL 的 MRR 为 0.52，略低于 GoldE（0.59）和 MPRM（0.63），但在 Hits@1 上达到 46.5%，比 RotatE（40.2%）等绝大部分方法更优，说明其在大规模场景下仍具备一定的稳定性。

同时，基于大模型的规则学习方法 ChatRule 展现了潜力，但在 FB15k-237，WN18RR 等核心基准数据集上，其 MRR 分别只有 0.30 和 0.35，明显落后于 CGRL。此外，在与 MPRM 的对比中，CGRL 在 FB15k-237，WN18RR，Family 等复杂数据集上均表现更佳，展现了更强的跨数据集适应性。

综上，CGRL 在复杂大规模知识图谱上展现出推理能力和语义建模优势，这一结果表明，CGRL 更适用于结构复杂、语义关联丰富的知识推理任务。

同时，为进一步分析 CGRL 的训练效能与区分能力，图 4 展示了在 Family 数据集上的训练曲线，包

括每轮训练的平均损失、正负样本的平均预测概率及其差值。



(a) 每轮平均损失 (b) 平均正样本预测概率

(c) 平均负样本预测概率 (d) 平均预测概率差异

Fig. 4 Training convergence curve of CGRL on Family dataset

图 4 CGRL 在 Family 数据集上的训练收敛曲线

如图 4 所示, 模型大约在第 13 轮迭代后开始收敛, 此时正样本预测概率超过 90%, 负样本概率降至 2% 左右。同时, 正负样本概率差值持续增大, 表明 CGRL 在训练过程中逐步增强了对有效规则与错误规则的判别能力, 整体趋势表明验证了对比学习机制能够有效优化规则体与规则头的语义一致性。

4.3 消融实验

为验证 CGRL 模型各组件的有效性, 本节设计了对照实验验证下述方法的效果: 节点分类方法、采样策略、结构感知位置编码以及对比学习, 并在下文详细分析其实验设计与结果。

4.3.1 采样模块

为验证不同采样策略在 CGRL 中的性能, 本文比较了随机采样、CGRL 串行采样以及 CGRL 并行采样(采用 4 个处理单元实现, 未进行额外并行优化)在 YAGO3-10 数据集上采样 20 000 个子图的运行时间, 实验结果如表 4 所示。

根据表 4 的结果可知, 引入节点重要性评分的 CGRL 采样方法虽然使单线程运行时间比随机采样增加约 10%, 但通过采用并行化策略, 其运行时间较串行版本缩短约 48%, 有效缓解了节点重要性评分带来的额外计算开销。

Table 4 Sampling Time Comparison of Different Sampling Methods on YAGO3-10

表 4 不同采样方法在 YAGO3-10 上的采样时间对比 s

采样方法	平均采样时间
随机采样	142
CGRL 串行采样	156
CGRL 并行采样	81

4.3.2 节点分类方法

本节将 CGRL 自适应聚类分类方法与无分类方法和基于邻域半径的 GraIL 双半径分类方法进行比较, 以检验自适应聚类分类在规则学习中的有效性。实验结果如图 5 所示。

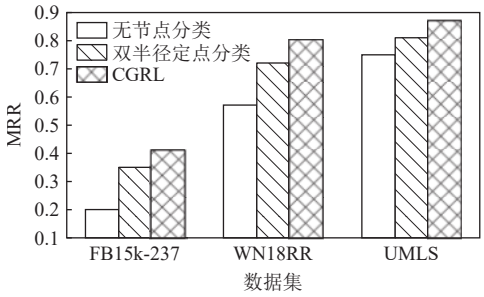


Fig. 5 Ablation study for CGRL node classification

图 5 CGRL 节点分类消融实验

如图 5 所示, 节点分类对模型性能影响显著: 当所有节点统一为同一类别时, 3 个数据集性能均下降, 尤其是关系复杂的 FB15k-237, 说明节点分类有助于捕捉图结构特征。与不带节点分类的方法相比, 引入软标签后推理性能平均提升约 25 个百分点, 从下游推理角度验证了节点分类方法的有效性。相比双半径方法, CGRL 自适应聚类在 WN18RR 和 UMLS 上 MRR 分别提升约 10 个百分点, 验证了基于表示学习的分类方法更能适应不同知识图谱结构, 从而增强规则学习效果。

4.3.3 结构感知的位置编码策略

为了验证 CGRL 设计的结构感知位置编码的效果, 本节将其与传统的正余弦位置编码进行对比。实验结果如图 6 所示, 使用正余弦编码会导致 FB15k-237、WN18RR 和 UMLS 这 3 个数据集的 MRR 指标下降, 其中 FB15k-237 上的下降幅度最为明显, 说明传统位置编码难以适应知识图谱的多样化结构特征。相比之下, 结构感知位置编码通过引入图结构信息, 提高了模型对不同位置节点的区分能力, 从而显著提升了规则学习效果。

4.3.4 对比学习模块

为验证对比学习在 CGRL 中的作用, 本节将其替换为多层感知器并采用均方误差损失优化规则头与规则体的关系。实验结果如图 7 所示, 替换对比学习导致 3 个数据集 MRR 指标显著下降: FB15k-237 相对下降约 50 个百分点, WN18RR 相对下降约 38 个百分点, UMLS 相对下降约 20 个百分点, 说明

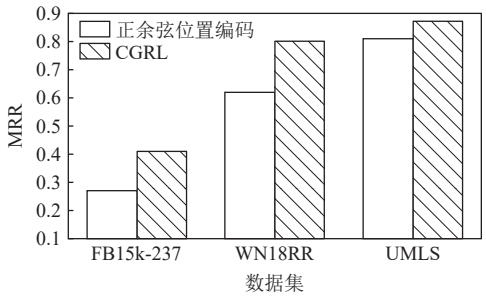


Fig. 6 Ablation study for CGRL structure-aware positional encoding strategies

图 6 CGRL 结构感知的位置编码策略消融实验

对比学习能通过正负样本特征对比有效捕捉规则的深层语义关联。在关系类型复杂的知识图谱中，对比学习促进模型区分语义相近的规则模式，而多层感知器仅建立规则头与规则体的直接映射，无法学习样本间的相对语义关系。

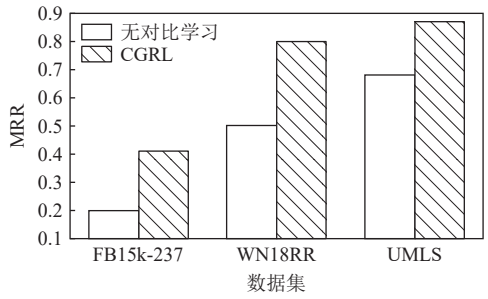


Fig. 7 Ablation study for CGRL contrastive learning module

图 7 CGRL 对比学习模块消融实验

4.4 软标签分析

本节进一步分析模型中节点分类机制，包括通过定量和案例分析说明自适应聚类生成的软标签，以及分析软标签对规则泛化能力的提升。

4.4.1 软标签分析

本节详细分析 CGRL 提出的自适应聚类的有效性。在真实知识图谱中，合理的关系与实体类型在统计上应呈现“高内聚、低耦合”的规律，即特定关系（如毕业于）应集中且一致地连接特定类型（如人物→学校），形成明显的头部效应；同时，它们的组合应具有稀疏性，即逻辑互斥的组合（如人物→发动机排量）在数据中应表现为结构化的零值或极低频，而非均匀分布。例如，如果一个聚类出的“人物”标签下，实体的关系分布既包含“出生日期”又包含“占地面积”且无明显主次，这就违反了集中性与稀疏模式，说明该软标签不合理。

基于这一结构特征，本节可以对软标签进行间接验证：具体可通过统计实体在特定标签簇内的关系分布一致性来侧面验证，即计算各簇内实体所参与关系的集中度，若某一标签下的实体均高度集中在若干特定关系上，且该分布模式与其他簇存在显著差异，则表明该软标签符合图谱的内在结构规律，具有合理性。

本节使用簇纯度指标验证衡量同一软标签下的实体在关系实例层面是否主要集中于特定关系类型，定义为：$Purity(C)=\max_r \frac{n_r}{|I(C)|}$，其中 C 是软标签聚类后的簇，$|I(C)|$为关系实例中头实体或尾实体属于簇 C 的实例大小，n_r表示在 $I(C)$中关系为 r 的实例数量。

实验结果如图 8 所示，大多数类别中得到的关系实例并未均匀分布于多种关系类型，而是明显集中于少数特定关系，这一现象显著强于随机划分情形。

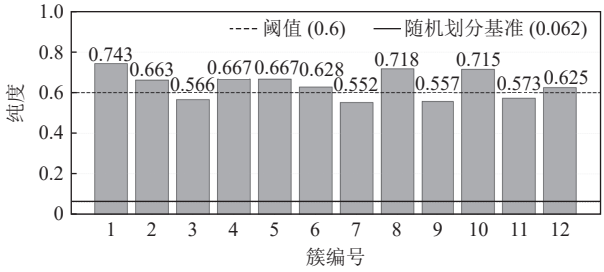


Fig. 8 Purity distribution of semantic clusters in CGRL (WN18RR dataset)

图 8 CGRL 的语义簇纯度分布（WN18RR 数据集）

同时，本节用热力图来分析不同软标签下实体在关系类型上的分布差异，关系占比越高，其颜色越深，如图 9 所示，不同类别中高频出现的关系类型明显不同：某些关系在特定类别中频繁出现，而在其他类别中则很少出现甚至几乎不存在。

本节进一步以簇为单位，统计各簇中实体在知识图谱中所涉及的主导头关系，以及它们的主导头关系占比，如表 5 所示，在每一个簇内对应的高频关系类型也存在明显差异。该结果表明，软标签划分后，不同节点类别在规则实例层面呈现出可区分的关系参与模式，而非随机混合。

4.4.2 软标签对规则泛化能力的验证

本节进一步分析自适应软标签机制在不同数据分布偏移场景下对规则泛化能力的影响，以系统评估其鲁棒性与跨分布适应能力。具体而言，本节在同一知识图谱上构造多种具有显著统计与结构差异的训练分布，包括：1）长尾分布，通过提升低频实体与稀疏结构路径的采样权重以模拟知识图谱中的稀疏语义区域；2）关系均衡分布，对不同关系类型进行

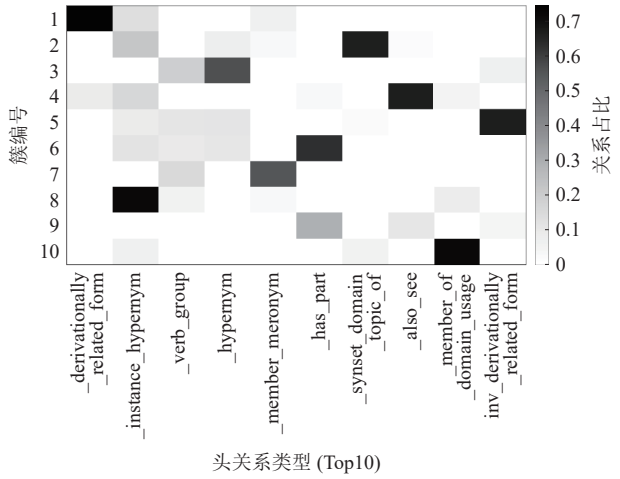


Fig. 9 Heat map of semantic clusters after CGRL clustering (WN18RR dataset)

图 9 CGRL 聚类后的语义簇热力图（WN18RR 数据集）

Table 5 Dominant Head Relation Types of Sematic Clusters (WN18RR Dataset)

表 5 语义簇的主导头关系类型（WN18RR 数据集）

簇编号	主导头关系类型（英文原名）	主导头关系占比/%
1	派生相关形式（_derivationally_related_form）	74
2	同义词领域主题（_synset_domain_topic_of）	66
3	上位词（_hypernym）	57
4	另见（_also_see）	66
7	成员整体关系（_member_meronym）	55

均匀采样以消除关系频率偏置；3)实体幂律分布，基于原始知识图谱中实体的节点度设定采样权重(系数为 2.5)，并据此对包含该实体的三元组进行加权采样，使训练集在实体出现频次上呈现更强的幂律特征。在上述分布偏移设置下，对比分析引入与不引入软标签机制时 CGRL 的性能变化。实验结果如图 10 所示，在 3 种不同的数据分布场景下，引入软标签均持续带来稳定且显著的性能提升。具体而言，在长尾分布与关系均衡分布下，MRR 分别提升约 40.0 个百分点与 39.3 个百分点，在实体幂律分布下仍保持约 30.3 个百分点的相对增益。不同分布条件下的性能提升幅度整体保持一致，表明自适应聚类所学习得到的语义分组先验并未过拟合于特定训练分布，而是在面对显著数据分布变化时仍能够稳定约束规则语义结构，从而有效提升模型的跨分布泛化能力。

4.5 模型稳健性分析

本节从规则大小、多路径推理复杂度及噪声扰动 3 个维度，对 CGRL 的稳健性进行系统分析。除特别说明外，所有实验均采用相同的参数配置。

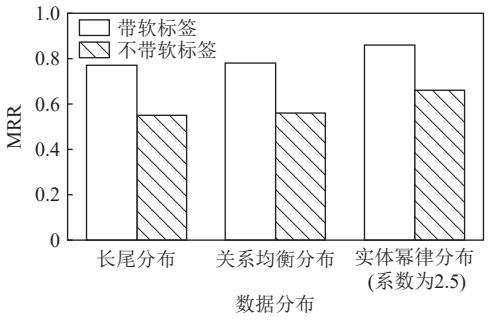


Fig. 10 Robustness evaluation of CGRL under different data distribution disturbance (WN18RR dataset)

图 10 不同数据分布扰动下 CGRL 的鲁棒性评估（WN18RR 数据集）

4.5.1 不同规则大小下 CGRL 稳健性分析

本节分析 CGRL 不同规则长度下的稳健性。具体而言，本节以规则体中包含的关系数作为规则长度度量指标，通过逐步扩大规则结构所覆盖的关系范围和节点连接复杂度，构造不同规模的图结构规则，用以评估 CGRL 在推理结构持续复杂化条件下的性能稳定性。

实验结果如图 11 所示，在所有数据集上，随着规则规模逐步扩大，CGRL 均未出现显著性能崩塌，表明模型在推理结构复杂性提升条件下保持良好稳定性。

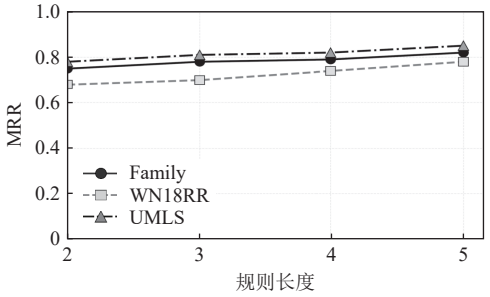


Fig. 11 Performance evaluation of CGRL on different datasets under varying rule size

图 11 不同数据集下 CGRL 面向规则长度变化的性能评估

4.5.2 多路径复杂度下 CGRL 稳健性分析

在 CGRL 中，规则由采样得到的子图生成，而子图的结构复杂性主要来源于其构建过程中所涉及的关系路径数量。不同数量的路径对应着规则体中可用推理路径的规模变化，从而反映了规则体结构在多路径条件下的复杂程度。为系统分析模型在多路径复杂度变化条件下的稳健性，本节通过在子图采样阶段控制所采样的路径数量，构造由低到高多路径复杂度的子图结构。随着采样路径数量的增加，

生成的子图包含更多并行路径，推理结构逐渐复杂。在此基础上，本节评估模型在多路径复杂度逐步增加条件下的性能变化与稳定性表现。

具体地，本节固定规则体规模为 5，并将子图构建阶段的采样路径数量设为 10, 20, 30, 40。每个结构模式最多可获得 50 条候选路径。实验结果如图 12 所示。

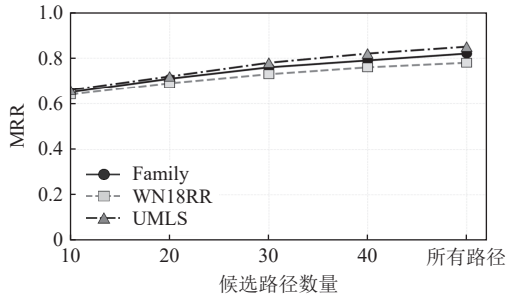


Fig. 12 Performance variation of CGRL under multi-path complexities on different datasets

图 12 不同数据集下 CGRL 多路径复杂度下性能的变化

如图 12 所示，在规则体规模保持不变的条件下，当候选结构分支数量逐步增加时，CGRL 在 3 个数据集上的推理性能整体呈现出持续但逐渐放缓的提升趋势。尽管路径复杂度不断增加，模型性能未出现明显波动或退化，表明 CGRL 对多路径结构复杂度变化具有较好的稳健性。

4.5.3 噪声干扰条件下 CGRL 稳健性分析

本节在训练阶段通过随机扰动部分图结构元素逐步引入不同比例的结构噪声，以模拟现实知识图谱中的结构不确定性与观测误差。具体而言，设噪声注入比例为$\eta \in [0,1]$，则对于每个三元组，以概率η对其进行结构扰动操作，包括随机替换关系类型或随机替换头实体与尾实体中的一项，从而构造噪声来增强训练样本。在此基础上，本节分析 CGRL 在多个数据集上的性能变化趋势。

实验结果如图 13 所示，随着噪声注入比例逐步增加，模型性能整体呈现出平滑、可控的下降趋势，而未出现性能突变或失稳现象，表明模型对噪声具有较好的鲁棒性。即使在 0.2 的噪声比例下，模型在 3 个数据集上仍能保持相对稳定的推理性能。

4.6 参数敏感性分析

为全面评估 CGRL 在关键参数变化下的性能，本节重点考察 3 项核心参数：节点采样比例、嵌入维度和投影头维度。其中，节点采样比例体现训练数据规模与计算效率之间的权衡，嵌入维度影响关系

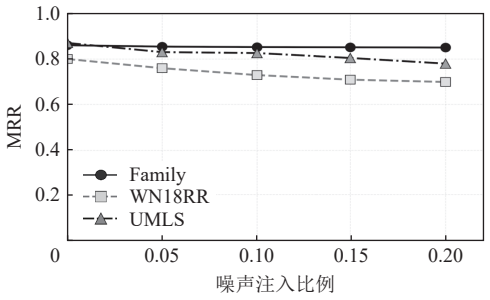


Fig. 13 Performance analysis of CGRL under noise perturbations on different datasets

图 13 不同数据集下 CGRL 面向噪声扰动的性能分析

表示能力，投影头维度则影响对比学习中的特征解耦效果。实验采用控制变量设计，在知识图谱补全任务上以 MRR 为指标，系统分析不同参数设置对模型性能及跨图谱适应性的影响。

4.6.1 采样比例的影响

采样比例决定了从知识图谱中抽取规则样本的比例，直接影响训练时间和模型性能。本文在不同采样比例设置下(0.1, 0.3, 0.5, 0.7, 0.9)开展实验，评估其对模型表现的影响。

实验结果如图 14 所示，当采样比例过小时，所有数据集上的 MRR 指标均较低，主要原因在于训练样本不足，限制了模型的学习能力。当采样比例提升到 0.5 时，模型在各数据集上均达到最佳性能，说明适中的样本量有助于捕捉图谱中的结构特征与语义模式。然而，当采样比例继续增大(超过 0.5)时，FB15k-237 和 UMLS 等数据集的性能略有下降，表明模型可能出现过拟合趋势。此外，较高的采样比例还会显著增加训练时间，但带来的性能提升却相对有限。综上，当采样比例设置为 0.5 时，模型能够在训练效率与推理性能之间取得较好平衡，适用于不同规模和复杂度的数据集，既避免了过拟合，又保障了模型的学习能力。

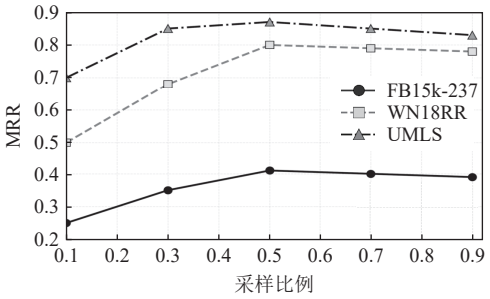


Fig. 14 Parameter sensitivity analysis of CGRL varying with sampling ratios on different datasets

图 14 不同数据集下 CGRL 对采样比例变化的参数敏感性分析

4.6.2 嵌入维度的影响

嵌入维度直接影响规则在低维空间中的表达能力。本文在不同嵌入维度设置下(64, 128, 256, 512, 1 024)开展实验,评估其对模型性能的影响。

实验结果如图 15 所示,在 FB15k-237 数据集上,较小的嵌入维度不足以建模复杂的关系结构,导致模型性能较低。随着嵌入维度的增加,模型的表达能力增强,性能随之提升,表明更高维度有助于捕捉复杂语义。然而,当维度超过一定阈值后,性能提升趋于饱和,同时带来额外的计算开销,性价比降低。在 WN18RR 和 UMLS 数据集上,由于图谱结构相对简单,较小的嵌入维度已能充分捕捉其中的语义模式,继续增加维度对性能提升有限,显示出在不同图谱特性下嵌入维度的敏感性差异。

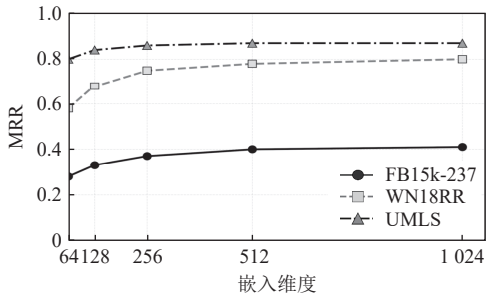


Fig. 15 Parameter sensitivity analysis of CGRL for the embedding dimension variation on different datasets

图 15 不同数据集下 CGRL 对嵌入维度变化的参数敏感性分析

4.6.3 投影头维度的影响

投影头维度是对比学习模块中的关键参数,直接影响特征在低维语义空间中的分布质量和区分能力。本文在多个评估投影维度设置下(32, 64, 128, 256, 512)对模型性能的影响。如图 16 所示,在 FB15k-237 数据集上,当投影维度为 128 时,模型的 MRR 表现达到峰值。维度过低会导致特征表示压缩严重,信息损失较大;而维度过高则可能引发过拟合,干扰模型对有用特征的提取,说明在复杂关系建模中,适中的投影维度有助于在信息保留与噪声过滤之间取得平衡。相比之下,在 WN18RR 和 UMLS 数据集中,当投影维度为 64 时已能获得较高的 MRR,当投影维度超过 128 时增加维度反而导致性能下降。这表明在结构较简单或数据量较小的图谱中,较低维度的投影空间更利于模型泛化。实验结果表明,最佳投影头维度依赖于图谱的规模与结构复杂度,实际应用中应根据数据特性灵活调整该参数。

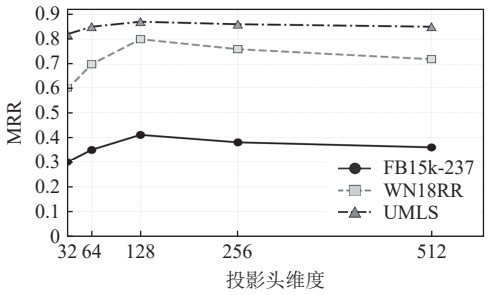
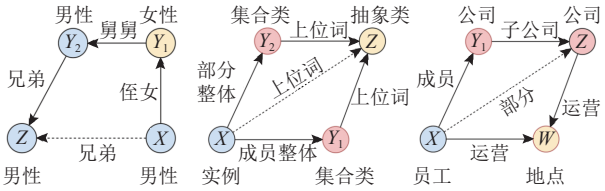


Fig. 16 Parameter sensitivity analysis of CGRL on variation of projection head dimension on different datasets

图 16 不同数据集下 CGRL 对投影头维度变化的参数敏感性分析

4.7 实证分析

图 17 展示了 CGRL 在不同数据集上挖掘出的 3 个高置信度规则实例,它们在结构上呈现出由简至繁的特点,反映出 CGRL 捕捉多样化语义模式的能力:图 17(a)表示若某人X有 1 名侄女 Y_1(即 X与Y_1存在侄女/舅舅关系),且该侄女Y_1有 1 名舅舅Y_2,同时该舅舅Y_2有 1 名兄弟Z,则可推断:X为Z的兄弟。图 17(b)体现出更丰富的语义层级与关系组合:若实体X同时作为集合Y_1的成员和作为某整体Y_2的部分,并且Y_1, Y_2均属于Z的下位类,则X可被推断为Z的下位类对象(即Z是对X的一个更抽象类别)。此规则不是简单的链式传递,而是多个条件的协同约束。图 17(c)展现出高度复杂的多跳推理与异构约束能力:如果X是Y_1的成员,且Y_1是Z的子公司,并且X与Z在同一地点运营,则可以推断X是Z的一部分。



(a) 亲属关系规则实例 (b) 概念层级规则实例 (c) 企业组织规则实例

注:大写字母表示节点编号,节点周边文字表示节点分类所得到的标签,关系周边文字表达关系类型,虚线表示推理出来的关系。

Fig. 17 Examples of rules learned by CGRL

图 17 CGRL 学习到的规则样例

这些图结构规则能够表达链式规则难以建模的多路径依赖、复合语义约束和跨关系推理模式,体现了图结构规则在知识推理中的潜在价值。

4.8 实验总结

CGRL 在公开数据集的实验效果说明了其优秀的性能。同时,消融实验说明了 CGRL 组件的重要性。参数敏感性分析表明,适度的采样比例、合理的嵌入维度与投影维度对于模型取得最优性能至关重

要。此外，实证分析表明，CGRL 能够挖掘出多类型、高置信度规则。需要指出的是，尽管模型在知识图谱中学习到的部分规则可能包含统计偏差，导致错误的推理结果，但 CGRL 整体能提取在推理任务中具有实用价值的潜在语义规律，为下游任务提供可靠支持。

5 结　论

本文针对知识图谱推理中规则学习方法局限于链式规则而难以有效挖掘图结构规则的问题，提出了一种基于对比学习的图结构规则学习方法 CGRL。该方法通过 3 个核心模块解决上述挑战：1)采用节点自适应聚类策略生成软标签，为规则学习引入类别语义约束；2)设计基于多跳采样的语义子图构建与规则样本生成机制，以支持复杂规则提取；3)引入结构感知的 Transformer 编码器与对比学习框架，共同优化规则的结构与语义表示，实现规则体内与规则头间的高效对齐。实验结果表明，CGRL 在多个公开数据集上的知识图谱补全任务中取得了优异性能，验证了图结构规则在推理任务中的有效性。未来的研究将探索图结构规则与传统链式规则的融合机制，并进一步拓展 CGRL 至时序知识图谱推理等复杂应用场景。

作者贡献声明：董博文负责模型的设计和实现以及论文的撰写和修改；刘雪莉提出研究方向，把握论文的创新性，并指导论文修改；冯潇洋负责模型的设计和实现；麦珍珍负责论文修改；王文俊提供项目支持。

参　考　文　献

[1] Li Rui, Li Chaozhuo, Shen Yanming, et al. Generalizing knowledge graph embedding with universal orthogonal parameterization[C]// Proc of the 41st Int Conf on Machine Learning. New York: PMLR, 2024: 28040−28059

[2] Yang Yuhao, Huang Chao, Xia Lianghao, et al. Knowledge graph contrastive learning for recommendation[C]//Proc of the 45th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2022: 1434−1443

[3] Wang Meng, Wang Haofen, Li Bohan, et al. Survey on key technologies of new generation knowledge graph[J]. Journal of Computer Research and Development, 2022, 59(9): 1947−1965 (in Chinese)
(王萌, 王昊奋, 李博涵, 等. 新一代知识图谱关键技术综述[J]. 计算机研究与发展, 2022, 59(9): 1947−1965)

[4] Liu Jingping, Cen Zhedong, Zhang Weiyan, et al. Open domain relation reasoning method based on generative model[J]. Journal of Computer Research and Development, 2026, 63(2): 378−392 (in Chinese)
(刘井平, 岑哲栋, 张维彦, 等. 基于生成模型的开放域关系推理方法[J]. 计算机研究与发展, 2026, 63(2): 378−392)

[5] Yang Haotong, Lin Zhouchen, Zhang Muhan. Rethinking knowledge graph evaluation under the open-world assumption[C]//Proc of the 36th Conf on Neural Information Processing Systems. New York: Curran Associates, Inc, 2022: 8374−8385

[6] Cui Yuanning, Sun Zequn, Hu Wei. A pre-trained universal knowledge graph reasoning model based on rule prompts[J]. Journal of Computer Research and Development, 2024, 61(8): 2030−2044 (in Chinese)
(崔员宁, 孙泽群, 胡伟. 基于规则提示的知识图谱通用推理预训练模型[J]. 计算机研究与发展, 2024, 61(8): 2030−2044)

[7] Bordes A, Usunier N, García-Durán A, et al. Translating embeddings for modeling multi-relational data[C]//Proc of the 27th Int Conf on Neural Information Processing Systems. New York: Curran Associates, Inc, 2013: 2787−2795

[8] Yang Bishan, Yih W T, He Xiaodong, et al. Embedding entities and relations for learning and inference in knowledge bases[J]. arXiv preprint, arXiv: 1412.6575, 2014

[9] Trouillon T, Welbl J, Riedel S, et al. Complex embeddings for simple link prediction[C]//Proc of the 33rd Int Conf on Machine Learning. New York: PMLR, 2016: 2071−2080

[10] Sun Zhiqing, Deng Zhihong, Nie Jianyun, et al. RotatE: Knowledge graph embedding by relational rotation in complex space[C/OL]// Proc of the 7th Int Conf on Learning Representations. 2019[2025-10-10]. https://openreview.net/forum?id=HkgEQnRqYQ

[11] Dettmers T, Minervini P, Stenetorp P, et al. Convolutional 2D knowledge graph embeddings[C]//Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto CA: AAAI, 2018: 1811−1818

[12] Yang Fan, Yang Zhilin, Cohen W W. Differentiable learning of logical rules for knowledge base reasoning[C]//Proc of the 31st Int Conf on Neural Information Processing Systems. New York: Curran Associates, Inc, 2017: 2316−2325

[13] Qu Meng, Chen Jian, Xhonneux L P, et al. RNNLogic: Learning logic rules for reasoning on knowledge graphs[C/OL]//Proc of the 9th Int Conf on Learning Representations. 2021[2025-10-10]. https://openreview.net/forum?id=tGZu6DlbreV

[14] Cheng Kewei, Liu Jiahao, Wang Wei, et al. RLogic: Recursive logical rule learning from knowledge graphs[C]//Proc of the 28th ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2022: 179−189

[15] Cheng Kewei, Ahmed N K, Sun Yizhou. Neural compositional rule learning for knowledge graph reasoning[C/OL]//Proc of the 11th Int Conf on Learning Representations. 2023[2025-10-10]. https://openreview.net/forum?id=F8VKQyDgRVj

[16] Xu Zezhong, Ye Peng, Chen Hui, et al. Ruleformer: Context-aware

rule mining over knowledge graph[J]. arXiv.2209.05815

[17] Chen Ye, Hong Xu, Zhang Hua, et al. Grier: Graph repairing based on iterative embedding and rules[J]. Knowledge and Information Systems, 2023, 65(8): 3273−3294

[18] Yang Yuan, Song Le. Learn to explain efficiently via neural logic inductive learning[C/OL]//Proc of the 8th Int Conf on Learning Representations. 2020[2025-10-10]. https://openreview.net/forum?id=SJlh8CEYDB

[19] Galárraga L A, Teflioudi C, Hose K, et al. AMIE: Association rule mining under incomplete evidence in ontological knowledge bases[C]//Proc of the 22nd Int Conf on World Wide Web. New York: ACM, 2013: 413−422

[20] Galárraga L A, Teflioudi C, Hose K, et al. Fast rule mining in ontological knowledge bases with AMIE+[J]. The VLDB Journal, 2015, 24(5): 707−730

[21] Du Xueying, Liu Mingwei, Shen Liwei, et al. A survey of knowledge graph embedding methods for link prediction[J]. Journal of Software, 2024, 35(1): 87−117 (in Chinese)
(杜雪盈, 刘名威, 沈立炜, 等. 面向链接预测的知识图谱表示学习方法综述[J]. 软件学报, 2024, 35(1): 87−117)

[22] Liang Ke, Meng Lingyuan, Liu Meng, et al. A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 9456−9478

[23] Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases[C]//Proc of the 20th Int Conf on Very Large Data Bases. San Francisco, CA: Morgan Kaufmann, 1994: 487−499

[24] Cordella L P, Foggia P, Sansone C, et al. A (sub)graph isomorphism algorithm for matching large graphs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(10): 1367−1372

[25] Cheng Kewei, Yang Ziqing, Zhang Ming, et al. UniKER: A unified framework for combining embedding and definite Horn rule reasoning for knowledge graph inference[C]//Proc of the 2021 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2021: 9753−9771

[26] Meilicke C, Chekol M W, Betz P, et al. Anytime bottom-up rule learning for large-scale knowledge graph completion[J]. The VLDB Journal, 2024, 33(1): 131−161

[27] Fan Wenfei, Jin Ruochun, Liu Muyang, et al. Capturing associations in graphs[J]. Proceedings of the VLDB Endowment, 2020, 13(12): 1863−1876

[28] Sadeghian A, Armandpour M, Ding P, et al. DRUM: End-to-end differentiable rule mining on knowledge graphs[C/OL]//Proc of the 33rd Conf Int Conf on Neural Information Processing Systems. New York: Curran Associates, Inc, 2019: 15321−15331

[29] Zhang Wen, Paudel B, Wang Liang, et al. Iteratively learning embeddings and rules for knowledge graph reasoning[C]//Proc of the 28th Int World Wide Web Conf. New York: ACM, 2019: 2366−2377

[30] Chen Shengyuan, Fang Huang, Cai Yunfeng, et al. Differentiable neuro-symbolic reasoning on large-scale knowledge graphs[C]//Proc of the 37th Int Conf on Neural Information Processing Systems. New York: Curran Associates Inc, 2023: 28139−28154

[31] Li Mingyang, Wang Song, Cai Ning. MPRM: A Markov path-based rule miner for efficient and interpretable knowledge graph reasoning[J]. arXiv preprint, arXiv: 2505.12329, 2025

[32] Luo Linhao, Ju Jiaxin, Xiong Bo, et al. ChatRule: Mining logical rules with large language models for knowledge graph reasoning[C]//Proc of the 29th Pacific-Asia Conf on Knowledge Discovery and Data Mining. Berlin: Springer, 2025: 314−325

[33] Wang Liang, Zhao Wei, Wei Zhuoyu, et al. SimKGC: Simple contrastive knowledge graph completion with pre-trained language models[C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics: ACL 2022. Dublin: Association for Computational Linguistics, 2022: 4281−4294

[34] Grover A, Leskovec J. Node2Vec: Scalable feature learning for networks[C]//Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 855−864

[35] Permadi V A, Tahalea S P, Agusdin R P. K-means and elbow method for cluster analysis of elementary school data[J]. Progres Pendidikan, 2023, 4(1): 50−57

[36] Hamilton W L, Ying R, Leskovec J. Inductive representation learning on large graphs[C]//Proc of the 31st Int Conf on Neural Information Processing Systems. New York: Curran Associates, Inc, 2017: 1025−1035

[37] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks and ISDN Systems, 1998, 30(1): 107−117

[38] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//Proc of the 2016 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770−778

[39] Toutanova K, Chen Danqi. Observed versus latent features for knowledge base and text inference[C]//Proc of the 3rd Workshop on Continuous Vector Space Models and Their Compositionality. Stroudsburg, PA: ACL, 2015: 57−66

[40] Hinton G E. Learning distributed representations of concepts[C]//Proc of the 8th Annual Conf of the Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum Associates, 1986: 1−12

[41] Kok S, Domingos P. Statistical predicate invention[C]//Proc of the 24th Int Conf on Machine Learning. New York: ACM, 2007: 433−440

[42] Suchanek F M, Kasneci G, Weikum G. YAGO: A core of semantic knowledge[C]//Proc of the 16th Int Conf on World Wide Web. New York: ACM, 2007: 697−706



Dong Bowen, born in 1995. PhD candidate. Student member of CCF. His main research interests include big data computing, data mining, and complex network analysis.

董博文, 1995 年生。博士研究生。CCF 学生会员。主要研究方向为大数据计算、数据挖掘、复杂网络分析。



Liu Xueli, born in 1987. PhD, associate professor. Senior member of CCF. Her main research interests include big data computing, data mining, and data security & privacy computing.

刘雪莉, 1987 年生。博士，副教授。CCF 高级会员。主要研究方向为大数据计算、数据挖掘、数据安全与隐私计算。



Feng Xiaoyang, born in 1999. Master. His main research interest includes rule mining.

冯潇洋, 1999 年生。硕士。主要研究方向为规则挖掘。



Mai Zhenzhen, born in 1990. PhD candidate. Her main research interests include big data computing, data mining, and rule learning and its applications in big data.

麦珍珍, 1990 年生。博士研究生。主要研究方向为大数据计算、数据挖掘、规则学习和其在大数据中的应用。



Wang Wenjun, born in 1970. PhD, professor, PhD supervisor. Member of CCF. His main research interests include big data mining, complex network analysis, and computational social science.

王文俊, 1970 年生。博士，教授，博士生导师。CCF 会员。主要研究方向为大数据挖掘、复杂网络分析、计算社会科学。