软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(12):5674−5694 [doi: 10.13328/j.cnki.jos.007403] [CSTR: 32375.14.jos.007403]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于嵌入模型的知识图谱准确性评估*

张明韬[1,2], 杨国利[1], 白晓颖[1]

[1](北京大数据先进技术研究院, 北京 100195)

[2](北京大学 计算机学院, 北京 100871)

通信作者: 杨国利, E-mail: yanggl@aibd.ac.cn; 白晓颖, E-mail: baixy@aibd.ac.cn

摘　要: 知识图谱构造常面临三元组错误或缺失等质量问题, 准确性评估是选择和优化知识图谱的基础, 对提升下游应用的可信性具有重要意义. 引入嵌入模型, 降低对人工标注数据的依赖性, 提升大规模数据处理效率. 将三元组正误判定转化为无标注的自动化阈值选择问题, 提出了 3 种阈值选择策略, 增强评估的鲁棒性. 提出结合三元组重要性的评估方法, 从网络结构和关系语义两方面定义重要性指标, 对关键结构、频繁访问的三元组赋予更高关注度. 从嵌入模型表征能力、知识图谱稠密度、三元组重要性计算方式等多个角度, 分析比较了对评估方法性能的影响. 实验表明, 相比现有知识图谱准确性的自动化评估方法, 在零样本条件下, 所提出的方法可有效降低评估误差, 平均降低接近 30%, 在错误率较高、稠密图谱的数据集上效果尤为显著.

关键词: 知识图谱; 准确性评估; 嵌入模型

中图法分类号: TP182

中文引用格式: 张明韬, 杨国利, 白晓颖. 基于嵌入模型的知识图谱准确性评估. 软件学报, 2025, 36(12): 5674–5694. http://www.jos.org.cn/1000-9825/7403.htm

英文引用格式: Zhang MT, Yang GL, Bai XY. Knowledge Graph Accuracy Evaluation Using Embedding Model. Ruan Jian Xue Bao/Journal of Software, 2025, 36(12): 5674–5694 (in Chinese). http://www.jos.org.cn/1000-9825/7403.htm

Knowledge Graph Accuracy Evaluation Using Embedding Model

ZHANG Ming-Tao[1,2], YANG Guo-Li[1], BAI Xiao-Ying[1]

[1](Advanced Institute of Big Data, Beijing, Beijing 100195, China)

[2](School of Computer Science, Peking University, Beijing 100871, China)

Abstract: Quality issues, such as errors or deficiencies in triplets, become increasingly prominent in knowledge graphs, severely affecting the credibility of downstream applications. Accuracy evaluation is crucial for building confidence in the use and optimization of knowledge graphs. An embedding-model-based method is proposed to reduce reliance on manually labeled data and to achieve scalable automatic evaluation. Triplet verification is formulated as an automated threshold selection problem, with three threshold selection strategies proposed to enhance the robustness of the evaluation. In addition, triplet importance indicators are incorporated to place greater emphasis on critical triplets, with importance scores defined based on network structure and relationship semantics. Experiments are conducted to analyze and compare the impact on performance from various perspectives, such as embedding model capacity, knowledge graph sparsity, and triplet importance definition. The results demonstrate that, compared to existing automated evaluation methods, the proposed method can significantly reduce evaluation errors by nearly 30% in zero-shot conditions, particularly on datasets of dense graphs with high error rates.

Key words: knowledge graph; accuracy evaluation; embedding model

知识图谱广泛应用于智能搜索、机器学习、自动问答、知识推理等领域, 例如 IBM Waston 进行智能问答时利用了 YAGO[1]知识图谱与 DBpedia[2]知识库, CMU 设计 NELL[3]系统收集网络信息进行知识整合与推理, Google

* 基金项目: 国家自然科学基金 (72201275)
收稿时间: 2024-07-13; 修改时间: 2024-09-26; 采用时间: 2025-02-10; jos 在线出版时间: 2025-06-18
CNKI 网络首发时间: 2025-06-19

搜索引擎基于 Freebase[4]数据增强搜索. 随着对海量、时变信息高效分析处理需求的日益凸显, 知识图谱自动化构建技术成为一个重要的研究方向. 相比于人工构建, 自动化构建技术可大幅提升知识图谱的规模、效率和时效性, 但也面临着严重的质量问题. 原始网络文本抽取产生的三元组集合可信程度往往较低, 例如 NELL 数据集在迭代抽取网络信息进行扩充的过程中, 整体正确率一度降至 71%, 其中 25% 的类型实体相关的抽取正确率低至 60% 以下 (见 NELL 官网, http://rtw.ml.cmu.edu/rtw/overview). 表 1 显示了几种典型的知识图谱错误. 其中, 下划线部分是对错误三元组修正的部分.

表 1 知识图谱质量问题示意

序号	错误三元组	来源数据集	正确三元组
1	(Gene Stupnitsky, Awards on Ceremony, Writers Guild of America Awards 2006)	FB15K-237 (Freebase)	(Gene Stupnitsky, Awards on Ceremony, Writers Guild of America Awards 2007), 或 (Gene Stupnitsky, Nomination on Ceremony, Writers Guild of America Awards 2006)
2	(Dimitri Marick, has Child, Erica Kane)	YAGO	(Dimitri Marick, is Married To, Erica Kane)
3	(Washington, is Taller than, two million people)	NELL-995 (NELL)	(two million people, Enrolled in, Health Insurance Plans), 及 (Washington, Report in News, Information about Insurance Plans)

(1) 三元组结构错误: 实体混淆了相似的实体或关系导致错误链接, 例如表 1 中的错误 1, 根据 IMDb (https://www.imdb.com/name/nm1969144/awards/), Gene 赢得 2007 年美国作家协会奖, 在 2006 年典礼上仅获得提名, 可能的错误原因是混淆了 2006 年典礼与 2007 年典礼, 或混淆了“获奖”与“提名”两类关系. 错误 2 在 KGEval[5]提供的 YAGO 样本数据 (见文献 [5] 附件 https://aclanthology.org/D17-1183/) 中被人工标注为错误三元组, 头尾实体人物事实上为婚姻关系, 可能的错误原因是文本描述中以整个家庭为单位, 信息整合时混淆了“结婚”与“子女”两类关系.

(2) 三元组语义错误: 由于语句解析和理解错误导致提取的三元组语义错误, 例如表 1 中错误 3, 根据相关文献 [6] 的追溯, 此三元组来源于新闻稿: “WASHINGTON, Dec 31 (Reuters)—Over two million people have enrolled in health insurance plans”, 模板将“Over”识别为关系词, 错误理解语句导致出现完全无含义的信息

知识图谱质量严重影响下游任务的可信性, 错误信息可能导致存在歧义的推理或搜索结果. 准确性评估可为知识图谱选择、取信、优化提供定量和定性的依据, 是一项具有重要意义和挑战性的任务. 评估过程通常包括样本三元组选取、样本正确率评估和知识图谱准确性评估这 3 个环节, 采样阶段抽取一定数量三元组样本, 以其正确率代表知识图谱整体的准确性, 样本评估阶段将样本与人工标注判断、其他可信知识库等信息进行匹配, 将匹配结果按照一定标准划分为正确/错误三元组, 统计样本正确率, 最终以样本正确率度量知识图谱整体准确性. 整体流程如图 1 所示, 其中特殊标记的深色区域表示可能需要人工标注或其他可信知识的部分, 可见需要此类信息的环节集中在样本评估阶段, 成为准确性评估问题的瓶颈.

知识图谱准确性评估现有工作主要存在以下两方面的问题. 一是现有的评估方法严重依赖外部信息, 如人工标注数据、专家知识、规则等. 例如 DeFacto[7]或 TAA[8]验证三元组依赖于互联网文本或其他知识图谱, 基于嵌入模型判断错误的方法中, KGTtm[9]需要人工标注数据训练分类器, ACTC[10]等阈值标定方法也需要人工动态标注确定正误三元组的划分阈值. 二是现有准确性评估方法常以正确三元组比例作为衡量知识图谱准确性的唯一量化指标, 忽略了三元组在不同领域、关注程度、访问频率上的差异. NELL 系统抽取的初始知识中, 正确率在不同领域之间存在严重失衡现象, 接近 3/4 实体类型的信息有 90% 以上能够被正确抽取, 而剩余的实体类型该比例仅在 25%–60% (见 NELL 官网: http://rtw.ml.cmu.edu/rtw/overview). 不同领域之间三元组正确比例差别显著, 而在查询者角度, 三元组的分布与查询同样存在长尾分布的倾向[11]. 以正确率平均值难以有效反映知识图谱整体的准确率.

针对上述问题, 本文提出采用基于嵌入模型的评估方法, 定义多种阈值选择策略, 并结合了三元组重要性度量. 嵌入模型在知识图谱补全、错误发现等领域受到广泛应用[12–18], 其将三元组实体、谓词嵌入隐式空间, 通过调整嵌入向量使之与知识图谱对齐. 在验证阶段, 通过嵌入模型对三元组的评分评估三元组与知识图谱的匹配程度, 作为对其合理性表示[10,17]. 在此基础上构建分类器, 通过阈值设定判断三元组正误. 在阈值设计上, 结合不同嵌入模型的损失函数, 提出固定阈值、动态阈值、修正阈值这 3 种选择策略, 提升判定的合理性和鲁棒性. 进而依据每

个子领域重要程度、每个三元组被访问的概率等信息, 定义三元组的重要性, 采用加权均值度量其准确性, 细化评估的粒度, 此时评估结果更符合应用场景、贴近用户体验. 本文从嵌入模型表征能力、知识图谱稠密度、三元组重要性计算方式等多个角度, 分析比较了对评估方法性能的影响. 实验表明, 相比现有知识图谱准确性的自动化评估方法, 在零样本条件下, 本文所提出的方法可有效降低评估误差, 平均降低接近 30%, 在错误率较高、稠密图谱的数据集上效果尤为显著.

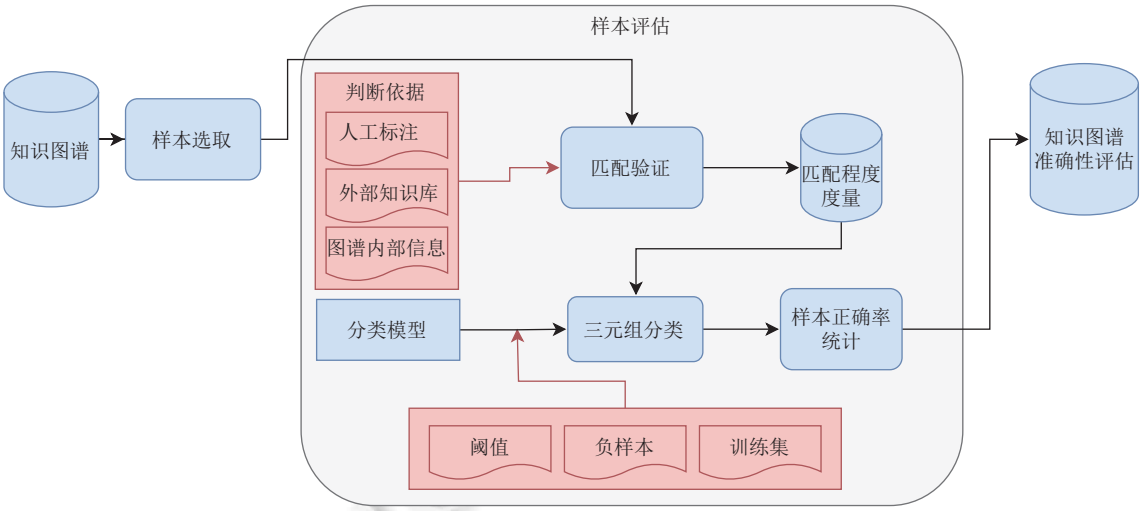

图 1 知识图谱正确性评估主要过程

本文第 1 节为问题定义. 第 2 节提出评估方法的框架. 第 3 节对当前方法进行实验分析. 第 4 节为准确性评估相关研究的介绍. 在第 5 节中进行总结.

1 问题定义

定义知识图谱 $G=\{t|t=(s,p,o)\}$, 记 $\mathcal{E}$ 为 G 当中实体构成的集合, $\mathcal{R}$ 为 G 中关系构成的集合, 因此 s, $o\in\mathcal{E}$, $p\in\mathcal{R}$, $G\subseteq\mathcal{E}\times\mathcal{R}\times\mathcal{E}$. 对于知识图谱当中某个三元组 $t\in G$, 以映射 f 表示单个三元组是否正确, 定义为 $f:\mathcal{E}\times\mathcal{R}\times\mathcal{E}\to\{0,1\}$, 0/1 分别表示三元组为错误/正确. 定义三元组重要性 $w:G\to\mathbb{R}^+$, $w(t)$ 的值越大表明三元组 t 在应用中对于知识图谱整体有着更大的影响. 由于不同用户查询的子领域、三元组流行程度的定义可能存在不同, w 的赋值并不唯一, 在不考虑用户查询情况、只强调三元组在知识图谱中作用时, 本文采用了均匀赋值、基于 PageRank 数值赋值以及结合实体出现频率这 3 种赋值方式作为样例. 以 $\mu_w(G)$ 表示知识图谱 G 结合三元组重要性 w 的正确率指标, 则:

$$\mu_w(G)=\left(\sum\nolimits_{t\in G}w(t)f(t)\right)\Big/\sum\nolimits_{t\in G}w(t) \tag{1}$$

本文以 $\mu_{\text{uni}}(G)$ 表示 w 满足 $\forall t\in G$, $w(t)=1$ 情况下的正确率, 称此时的 w 为均匀赋值下的重要性, 此时 w 记作 w_{uni}, 可见 $\mu_{\text{uni}}(G)$ 即为传统的知识图谱正确率定义. 而对于一般 μ_w, 本文将其解释为某一种查询方式下知识图谱正确反馈信息的比例. 具体而言, 针对用户提供的信息给出相关的三元组, 令 $q(u)$ 为用户 u 查询的三元组, 记 $p_u(t)=\Pr(q(u)=t)$, 在 w 取值为 $p_u(t)$ 时:

$$\mu_w(G)=\sum\nolimits_{t\in G}f(t)w(t)=\sum\nolimits_{t\in G}f(t)p_u(t)=\sum\nolimits_{t\in G}\Pr(f(t)=1|q(u)=t)\Pr(q(u)=t)=\Pr(f(q(u))=1) \tag{2}$$

因此使用 $\mu_{\text{uni}}(G)$ 即为计算 $\Pr(f(t)=1,t\sim U(G))$, 即从 G 集合中均匀采样, 相比之下可见 $\mu_w(G)$ 有着更好的物理含义. 即使在真实 $p_u(t)$ 值不可用时, 利用随机游走产生的 PageRank 衡量每个实体被访问的概率, 也与 YAGO 等知识库中基于图形化链接进行搜索的形式相一致. 然而, $w(t)$ 的不确定性也导致 $\mu_w(G)$ 的计算更加困难, 基于抽

样的方法需要成倍提升其样本所需规模以保证评估结果在不同权重下仍然形成无偏近似, 进一步降低了评估效率.

对于准确性评估问题, 记方法判断三元组 t 的结果为 $\hat{f}(t)\in\{0,1\}$, 则知识图谱正确率估计值如下, 可见此时评估误差主要受到 $\hat{f}$ 的影响.

$$\hat{\mu}_w(G)=\left(\sum_{t\in G}w(t)\hat{f}(t)\right)\Big/\sum_{t\in G}w(t) \tag{3}$$

本文目标为通过 $\hat{\mu}_w(G)$ 近似 $\mu_w(G)$, 即最小化 $ERR:=|\hat{\mu}_w(G)-\mu_w(G)|_{\text{abs}}$, 其中, $|\cdot|_{\text{abs}}$ 表示取绝对值.

图 2 给出了知识图谱准确性评估示例, 共 5 个三元组中存在一个虚线边, 代表该信息存在错误, 整体正确率为 $4/5=80\%$. 若用户搜索 $e2$ 相关的信息时, $(e3,r3,e4)$ 的正误并不重要, 此时用户访问的正确率结果为 3/4=75%; 若用户在该图上进行随机游走, 从某一实体访问其相邻实体, 则过程中访问 $e2$ 的可能性高于其余实体, 包含 $e2$ 的三元组其重要性同样高于其他三元组, 该示例说明结合重要性进行准确性评估的重要性.

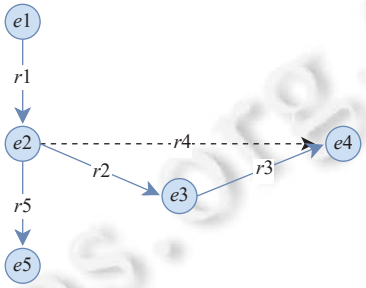

图 2 知识图谱示例

2 基于嵌入模型的评估方法

本文应用嵌入模型进行自动化的知识图谱准确评估. 嵌入模型整体可表示为函数 $Func_\theta:\mathcal{E}\times\mathcal{R}\times\mathcal{E}\to\mathbb{K}$, $Func_\theta(s,p,o)=\widehat{Func}(\boldsymbol{e}_s,\boldsymbol{e}_p,\boldsymbol{e}_o)$.

实体嵌入层: $EntEmbed_\theta:\mathcal{E}\to\mathbb{K}^{d_1}$, 关系嵌入层: $RelEmbed_\theta:\mathcal{R}\to\mathbb{K}^{d_2}$. 嵌入层将实体 s, o 或关系 p 分别映射成为向量 $\boldsymbol{e}_s$, $\boldsymbol{e}_o\in\mathbb{K}^{d_1}$ 或 $\boldsymbol{e}_p\in\mathbb{K}^{d_2}$.

嵌入函数: $\widehat{Func}:\mathbb{K}^{d_1}\times\mathbb{K}^{d_2}\times\mathbb{K}^{d_1}\to\mathbb{R}$, 作用于嵌入层输出向量, 其结果衡量该三元组是否可能成立, 记为该三元组评分. 该函数既可以基于建模方式定义为简单的向量运算结果 (例如 TransE 模型[12]、ComplEx 模型[13]), 也可以基于参数的反向传播习得 (例如 ConvE 模型[14]).

嵌入模型根据其建模方式可分为基于平移的嵌入模型 (例如 TransE 模型[12], 其嵌入函数形如向量之间的距离), 基于乘法的嵌入模型 (例如 ComplEx 模型[13], 其嵌入函数形如双线性函数), 以及其他类型模型 (例如定义在旋转意义下的 RotatE[15], 增加了其他特征的 CKRL[16]等).

2.1 基于嵌入模型的准确性评估框架

基于嵌入模型的准确性评估流程如图 3 所示, 整体分为 3 个阶段.

(1) 选择训练集与测试集 $\mathcal{S}_1(G)$ 和 $\mathcal{S}_2(G)$, 其中测试集即为进行准确性评估所用样本.

(2) 选择训练集后, 为减少错误三元组影响, 仅对嵌入模型 $Func_\theta$ 进行少量轮次训练, 并对测试集中的三元组进行评估, 嵌入模型分别对三元组返回模型函数值 $Func_\theta(t)$, 称其为三元组评分, 作为在无可对照真实数据情况下三元组合理性表示.

(3) 根据三元组评分选择合理阈值 λ, 进而利用标签映射函数 $Label_\lambda$ 将三元组评分转化为正误标签 (即 $\hat{f}(t)=Label_\lambda(Func_\theta(t))$), 根据定义计算加权平均值.

其中, 第 1 阶段需满足测试集 $\mathcal{S}_2(G)$ 的评估结果与整体知识图谱一致, 因此 $\mathcal{S}_2(G)$ 选择中首先根据 $w(t)$ 选择三元组中的主要部分, $w(t)$ 即为前文所述的任意某种三元组重要性, 随后采用抽样方式选取样本测试集, 即 $\mathcal{S}_2(G)=$

$Sample(Filter_{\xi}(G,w(t)))$, $Filter_{\xi}$ 表示满足三元组重要性大于 ξ 的三元组集合, *Sample* 为预定义的随机抽样方法, 本文采用简单随机抽样方法. 附录 A 中证明 $\mathcal{S}_2(G)$ 以高概率近似 G 的评估结果, 其评估误差与 *Sample* 样本的数量相关. $\mathcal{S}_1(G)$ 为满足 $\mathcal{S}_1(G)\subseteq G-\mathcal{S}_2(G)$ 的三元组集合, 用于进行嵌入模型的训练. 由于 $|\mathcal{S}_1(G)|$ 过大会导致更多计算复杂度, 过小则导致训练信息不足, 本文令 $\mathcal{S}_1(G)=G-\mathcal{S}_2(G)$ 使得训练集最优.

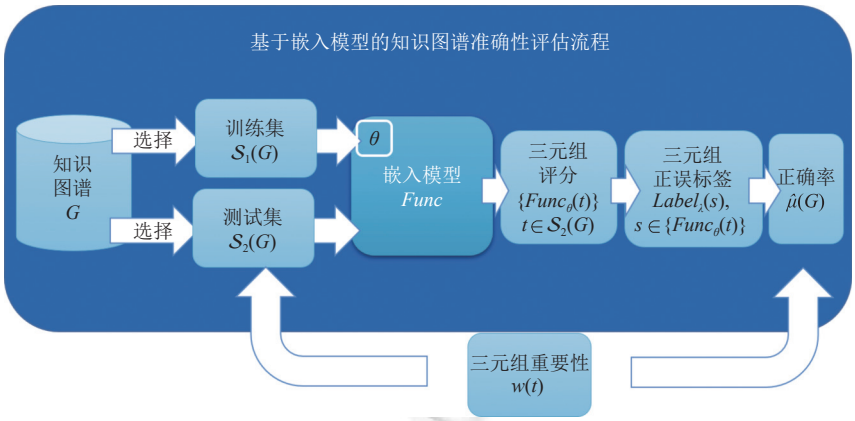

图 3 基于嵌入模型的知识图谱准确性评估框架

第 2 阶段进行嵌入模型的训练. 由于正确三元组联系紧密, 更容易被模型表示, 为减少训练集中错误三元组影响, 进行少量轮次训练以保证正确三元组表示的同时, 减少对错误三元组的过拟合, 便于后续在测试集中分离正确三元组与错误三元组. 考虑到更改训练集样本的损失权重可能忽略某些实际正确的三元组, 本文未采用通过调整某些样本的损失权重[16]等缓解三元组错误影响的方法.

第 3 阶段, $Label_\lambda$ 将输入评分转化为三元组正误标签, 其形式为在阈值 λ 发生突变的跳跃函数, 本文假定正确三元组评分大于错误三元组 (由于此处只涉及三元组评分的相对大小关系, 可通过取相反数使得三元组评分满足统一形式, 此处借鉴相关文献 [18] 的形式化), 则:

$$Label_\lambda(s)=\begin{cases}0, & s<\lambda\\ 1, & s\geqslant\lambda\end{cases} \tag{4}$$

2.2 三元组正误判断阈值优化

阈值 λ 的选择是自动化评估中最具挑战之处, 现有方法往往需人工标注数据构建分类器, 本文借助嵌入模型的定义以及负采样策略下的优化策略, 提出 3 种不同阈值选择策略以实现自动化评估.

(1) 固定阈值: 对于基于乘法的嵌入模型, 定义固定阈值, 记为 λ_0, $\lambda_0:=0$. 对于基于乘法的嵌入模型, 知识图谱中的关系被表示为双线性函数 (将该线性函数表示为矩阵 M_r, 则嵌入函数即为 $h^{\mathrm{T}}M_rt$), 模型评分 $h^{\mathrm{T}}M_rt$ 可解释为头实体与关系表示与尾实体向量之间的内积 $(M_r^{\mathrm{T}}h,t)=\left|M_r^{\mathrm{T}}h\right||t|\cos\langle M_r^{\mathrm{T}}h,t\rangle$, 该解释中评分的正负反映了向量的同向/反向, 因而存在可解释的固定阈值 $\lambda_0:=0$. 对于其他类型的嵌入模型该阈值并不适用.

(2) 动态阈值: 正确、错误三元组数量未知时, 由于利用已有信息无法直接表示 *ERR*, 能使 *ERR* 最小的 λ (记为 $\mathrm{argmin}_\lambda ERR$) 难以直接作为最优化目标. 由附录 B 中证明, *ERR* 的上界为 *ACC* (定义为 f 与 $\hat{f}$ 判断相同的比例 $\left|\left\{t\in\mathcal{S}_2(G):\hat{f}(t)=f(t)\right\}\right|_w\Big/|\mathcal{S}_2(G)|_w$, 其中 $|\cdot|_w$ 表示按权重 w 计算加权和) 控制, 本文通过近似 *ACC* 的最优阈值, 记为 $\lambda_{\mathrm{opt}}=\mathrm{argmax}_\lambda ACC$ 实现阈值选择. 形式上, 由于知识图谱 G 中存在一定比例错误, 因此有 $G=G_T\cup G_F$, G_T, G_F 代表正确/错误的三元组集合, 令 $Score_T=\{Func_\theta(t),t\in G_T\}$, $Score_F=\{Func_\theta(t),t\in G_F\}$, 则 $Score_G=\{Func_\theta(t),t\in G\}=Score_T\cup Score_F$.

若采用均匀加权 w_u (其他加权同理), 则选择的最优阈值 λ_{opt} (实验中可通过 GlobalOpt[17]等方式进行标定, GlobalOpt 即为根据所有三元组的标注结果, 在一系列阈值中根据某指标 (原文选择 *ACC*) 选择统一最优阈值, 此

处沿用了相关文献 [10] 中的命名) 满足:

$$\lambda_{\text{opt}} = \text{argmax}_{\lambda}\left(|Score_T \geqslant \lambda|_{\text{ineq}} + |Score_F < \lambda|_{\text{ineq}}\right) \tag{5}$$

$|S\ OPT\ \lambda|_{\text{ineq}}$ 表示 S 集合中满足不等式的元素个数, 即 $\{s \in S | s\ OPT\ \lambda\}$ 的元素个数. 在未给定具体标签的情况下, 阈值选择问题相对具有挑战性, 本文采用负采样策略生成负例集 S_{Neg}, 满足 $Score_{\text{Neg}} = \{Func_{\theta}(t), t \in S_{\text{Neg}}\}$ 与 $Score_F$ 数值分布上相似, 则 λ_{opt} 也可根据新生成的错误三元组评分集合进行选择:

$$\begin{aligned}\lambda_{\text{opt}} &\approx \text{argmax}_{\lambda}\left(|Score_T \geqslant \lambda|_{\text{ineq}} + |Score_F < \lambda|_{\text{ineq}} + |Score_{\text{Neg}} < \lambda|_{\text{ineq}}\right)\\ &= \text{argmax}_{\lambda}\left(|Score_T \geqslant \lambda|_{\text{ineq}} + |(Score_F \cup Score_{\text{Neg}}) < \lambda|_{\text{ineq}}\right)\end{aligned} \tag{6}$$

由于 $Score_G = Score_T \cup Score_F$, 则近似 λ_{opt} 有:

$$\begin{aligned}\lambda_{\text{opt}} &\approx \text{argmax}_{\lambda}\left(|Score_G \geqslant \lambda|_{\text{ineq}} - |Score_F \geqslant \lambda|_{\text{ineq}} + |Score_F < \lambda|_{\text{ineq}} + |Score_{\text{Neg}} < \lambda|_{\text{ineq}}\right)\\ &= \text{argmax}_{\lambda}\left(|Score_G \geqslant \lambda|_{\text{ineq}} + |Score_{\text{Neg}} < \lambda|_{\text{ineq}} - 2 \times |Score_F \geqslant \lambda|_{\text{ineq}}\right)\end{aligned} \tag{7}$$

由于 $|Score_F \geqslant \lambda|_{\text{ineq}} \leqslant |G_F|$, 在 $|G_F|$ 占较小比例时, 模型评分的相对大小关系反映了三元组的合理性, 进一步降低了 $|Score_F \geqslant \lambda|_{\text{ineq}}$ 的数量, 因此忽略该部分以近似 λ_{opt}, 记作动态阈值 λ_1:

$$\lambda_1 := \text{argmax}_{\lambda}\left(|Score_G \geqslant \lambda|_{\text{ineq}} + |Score_{\text{Neg}} < \lambda|_{\text{ineq}}\right) \approx \lambda_{\text{opt}} \tag{8}$$

可见, $\left|S_{\text{Neg}}\right|$ 过大会导致 $|Score_G \geqslant \lambda|_{\text{ineq}} + |Score_{\text{Neg}} < \lambda|_{\text{ineq}}$ 中 λ_1 更容易受到 $Score_{\text{Neg}}$ 集合的支配, 反之同理, 实验设定 $|S_{\text{Neg}}| = |G|$, 降低集合大小对阈值选择的影响.

(3) 修正阈值: 根据定义可见 λ_{opt} 与 λ_1 之间大小关系: 定义 $h(\lambda) = |Score_G \geqslant \lambda|_{\text{ineq}} + |Score_{\text{Neg}} < \lambda|_{\text{ineq}}$, 根据公式 (7) 和公式 (8), 有 $|Score_G \geqslant \lambda|_{\text{ineq}} + |Score_{\text{Neg}} < \lambda|_{\text{ineq}} - 2 \times |Score_F \geqslant \lambda|_{\text{ineq}} = h(\lambda) - 2 \times |Score_F \geqslant \lambda|_{\text{ineq}}$, 由于已知 $|Score_F \geqslant \lambda|_{\text{ineq}}$ 随 λ 增加而递减, 因此当公式 (8) 取得极值时, 公式 (7) 中函数仍在上升, 故 $\lambda_1 \leqslant \lambda_{\text{opt}}$, 证明当前近似方法相对真实的最优阈值存在一定偏差, 忽略了 $|Score_F \geqslant \lambda|_{\text{ineq}}$ 的影响.

值得注意的, $\lambda_1 \leqslant \lambda_{\text{opt}}$ 仅为理论分析的结果, 在特殊情况下, 稀疏的知识图谱 (关系或实体的数量较多, 整体三元组数量较少的知识图谱) 在嵌入模型训练时易过拟合错误三元组, 导致计算动态阈值中同分布假设有误, 该大小关系自然不成立. 而当符合假设要求时, 可利用 λ_1 与 λ_{opt} 的关系对 $|Score_F \geqslant \lambda_1|_{\text{ineq}}$ 进一步近似, 从而对 λ_1 进行优化.

具体而言, 由于 $\lambda_1 \leqslant \lambda_{\text{opt}}$, 本文假设 $Score_{F1} = \{s \in Score_G,\ s < \lambda_1\} \subset Score_F$, $Score_F$ 近似于正态分布. 利用最大似然法算法最大化 $Score_F$ 的似然概率, 可近似补全 $Score_{F1}$ 中大于等于 λ_1 的部分. 但采用最大似然法估算高斯分布的参数时有 $\mu_{\text{gauss}} = Mean(Score_F)$, $\sigma_{\text{gauss}} = Std(Score_F)$, μ_{gauss}, σ_{gauss} 为分布的参数, $Mean$, Std 代表集合的平均值与标准差. 由于 $Score_{F1}$ 为 $Score_F$ 中相对更小的一部分, 其平均值近似 μ_{gauss} 存在偏差, 因此本文将分数平均划分为若干区间, 将 μ_{gauss} 设置为含有最多集合元素的区间的中点. 通过 $Score_F$ 的预测高斯分布可计算修正后的错误三元组数量, 进而计算修正后的三元组正误判断阈值. 记 λ_1 调整后阈值为 λ_1', 即修正阈值, 相对于 λ_1 运用了更强的假设进行修正.

本文将在后续知识图谱准确性评估实验中分析对比 λ_0 (即上述固定阈值), λ_1 (即上述动态阈值), λ_1' (即上述修正阈值) 三者性能.

2.3 三元组重要性定义

由于目前缺少三元组重要性相关工作, 本文主要借鉴了实体重要性衡量的模式与定义. 实体重要性可通过实体被访问概率直观反映, 由于开源知识图谱中一般缺少访问的历史记录信息, GENI 方法[19]将三元组中的实体/关系映射到互联网网页, 利用网页被访问频率构建真实重要性度量. PageRank 分数基于拓扑结构进行随机游走, 衡量节点的重要性, PPR (personalized PageRank)[20]允许用户自定义节点游走权重, 增强灵活度. HAR[21]在 PR 和 PPR 基础上扩展, 利用图拓扑结构以及自定义输入的同时, 区分知识图谱中的不同关系. GENI[19]同时考虑图的拓扑结构以及对于每个实体而言不同谓词的作用, 利用深度学习对标准集进行拟合.

借鉴实体重要性的度量要素可从以下两个方面定义三元组重要性: 一是网络结构信息, 由于相邻实体彼此交互, 并且它们倾向于共享共同特征, 因此受关注的实体, 其相邻的实体、关系也倾向被关注; 二是关系语义信息, 知

识图谱由多种类型的关系组成, 三元组中的谓词在确定三元组重要性方面起着重要作用. 本文希望通过这两种简单直接的三元组重要性特征, 直观衡量三元组不同重要性分布对准确性评估框架的影响, 并证明当前框架在不同重要性分布下的有效性.

本文采用均匀分布、基于网络结构、基于关系语义这 3 种加权方式作为三元组重要性度量.

(1) 基于均匀分布 (Uniform) 的加权 $w_{\text{uni}}: w_{\text{uni}}(t)=1$, 即为将所有三元组等同看待, 赋予相同权重. 此为传统正确率定义中采用的加权, 评估结果即为 $\hat{\mu}_{\text{uni}}(G)$.

(2) 基于网络结构 (Structure) 的加权 w_s, 即为考虑三元组结构信息情况下的加权计算. w_s 的赋权往往利用网络结构衡量节点的重要性程度. 本文计算上基于 PageRank 算法[20], 将每个三元组的重要性表示为三元组被图上随机游走中被访问的概率. 具体计算如算法 1.

(3) 基于关系语义 (Relation) 的加权 w_r, 关注不同节点之间的不同关系带来的影响. 例如, 一个著名的导演其拍摄的电影往往具有更高的被访问概率, 而在某一个著名的城市取景的电影, 其受欢迎程度不一定与城市自身知名度相关, 计算上通过设计指标衡量三元组提供的实体信息从而衡量三元组的重要性.

基于 PageRank 算法实现基于结构的重要性加权算法. 由于图形化知识图谱查询方式与互联网超链接形式类似, PR 分数反映了在图上随机游走后到达某个节点的概率, 可用于度量节点的关注程度. 本文在知识图谱上构建反边进行随机游走, 访问三元组的概率即为其两个实体的 PR 分数平均. 伪代码如算法 1 所示.

算法 1. 基于结构的三元组重要性计算.

输入: 知识图谱 G; 测试集/样本 S;
输出: 样本 S 中每个三元组对应的重要性.

1. $PR = PageRank(G)$;
2. **for** $triple=(s,p,o)\in S$;
3. $\quad triple.importance = \dfrac{PR(s)}{OutDegree(s)} + \dfrac{PR(o)}{InDegree(o)}$;
4. **return** $\{triple.importance | triple \in S\}$.

对于基于关系语义的三元组重要性加权方法, 本文认为三元组的语义体现为其为三元组实体提供的信息. 本文此处直接估算三元组限制的集合大小, 以度量每个三元组的增加对实体不确定性的降低.

给定知识图谱三元组 (s,p,o), 令 $F_p(s)$ 表示集合 $\{o|(s,p,o)\in G\}$, $F_p^{-1}(o)$ 表示集合 $\{s|(s,p,o)\in G\}$, $Im(F_p)$ 表示集合 $\{o|\exists s_0,(s_0,p,o)\in G\}$, $Def(F_p)$ 表示集合 $\{s|\exists o_0,(s,p,o_0)\in G\}$. 由于该三元组表明 $s\in F_p^{-1}(o)$, $o\in F_p(s)$, 但考虑到三元组中通过 s 访问 o 的情况多于以 o 搜索 s 的情况, 这种有向性使得两者从重要性上并不完全对称, 同时图谱中的错误可能使得该集合限制并不准确, 因此定义该三元组的语义信息为 $\dfrac{1}{|F_p^{-1}(o)|}\times\dfrac{1}{|Im(F_p)|}$, 以降低 o 重要性的影响 ($F_p(s)\subset Im(F_p)$, $|F_p(s)|\leqslant|Im(F_p)|$). 伪代码如算法 2.

算法 2. 基于关系语义的三元组重要性计算.

输入: 知识图谱 G; 测试集/样本 S;
输出: 样本 S 中每个三元组对应的重要性.

1. 计算每个关系 p 相关的信息;
2. **for** $triple=(s,p,o)\in S$;
3. $\quad triple.importance = \dfrac{1}{|F_p^{-1}(o)|}\times\dfrac{1}{|Im(F_p)|}$;
4. **return** $\{triple.importance | triple \in S\}$.

本文所提出的框架并不限制特定形式的重要性度量, 此处给出了两种简单基础的三元组重要性度量, 并以此证明本文框架对不同三元组重要性的可行性与泛用性. 两种基础定义之间存在着差异与共性: 基于网络结构计算三元组重要性时, 假定重要的实体之间联系紧密, 访问往往从一个待查实体出发, 通过三元组访问到另一个实体, 该过程中重要的实体有更高的概率被频繁地访问到, 因此可以被基于随机游走的 PageRank 算法模拟; 基于关系语义的三元组重要性则是从关系的角度, 衡量三元组中关系是否为实体的类型提供了重要依据 (例如 (*A*, lives_in, *B*) 是一个常见的三元组, 提供的信息仅为为 *A* 为人名、*B* 为地名、*A* 的居住地, 其中前两个信息几乎无用, 而 (*A*, is_the_leader_of, *B*) 隐含着 *A* 可能是政客且归属于 *B*、*B* 为某种政党或组织, 提供了更多额外的信息). 基于网络结构与关系语义的三元组重要性出发点与侧重点存在不同, 但均能够一定程度反映三元组的重要程度, 体现了知识图谱中三元组是否处在关键结构/易被频繁访问的倾向.

3 实验分析

本节对方法框架 (主要为阈值选择策略)、嵌入模型与数据集选择、三元组重要性等进行实验分析.

3.1 实验数据集

在验证本文流程的可行性时, 本文选择具有常见知识图谱, 并在其中加入一定比例人工构造的错误三元组生成模拟数据集 (为保证最终结果的有效性, 此处构造的错误三元组与近似最优阈值时采用的负采样策略分布上并不相同), 主要包含以下 3 种.

(1) FB15K-237: FB15K 与 FB15K-237 数据集基于 Freebase[4]建立, FB15K 为对 Freebase 原数据抽取其中 15 000 个实体对应三元组构建而成, FB15K-237 对 FB15K 中数据筛选产生, 除去了其中的反向关系.

(2) NELL-995: 该数据集抽取自 NELL (never ending language learning) 数据集[3]中数据, 其通过对网络相关文本的实体识别、关系抽取产生, 其代表自动化生成的知识图谱.

(3) YAGO3-10: 该数据集抽取自 YAGO 知识图谱[1], 它产生于对部分百科词条的收集归纳, 代表基于已有结构化、半结构化数据而产生的知识图谱.

关于知识图谱大小、知识图谱稠密性 (可被平均每个实体/每类关系对应三元组反映) 等相关因素见表 2.

表 2 数据集相关信息

因素	FB15K-237	NELL-995	YAGO3-10
实体数	14 541	75 492	123 182
关系数	237	200	37
三元组数 (train)	272 115	149 678	1 079 040
三元组数 (valid)	17 535	543	5 000
三元组数 (test)	20 466	3 992	5 000
平均每个实体对应三元组	21.33	2.04	8.84
平均每类关系对应三元组	1 308.51	771.07	29 433.51

经过人工的抽样检测, FB15K-237 的正确率大约在 93%, NELL-995 的正确率大约在 90%, YAGO3-10 的正确率大约在 95% (错误样例可见表 1). 本文假定其原有三元组整体正确, 构造模拟数据. 沿用 KGTtm 的错误数据构造方法[9], 对数据集添加构造错误, 构造策略如下: 等概率随机选择一定比例的三元组, 替换其中的头实体或尾实体为相近实体 (相近实体指与当前三元组有着相同关系 *p* 的其他三元组中, 当前待替换实体对应位置的实体, 具体样例可见图 4), 或随机替换其关系 *p*; 构建出的新三元组不在知识图谱中时, 假定新产生的三元组为错误三元组.

由此得到 FB15K-237-SYN10, FB15K-237-SYN20, FB15K-237-SYN40 等, FB15K-237-SYN“*X*”表示在 FB15K-237 数据中加入 *X*% 比例错误产生的知识图谱, 此处“*X*”取值主要借鉴了 NELL 网页初始抽取结果的正确率等级. 对于 NELL-995、YAGO3-10 等有着类似的一系列测试数据集.

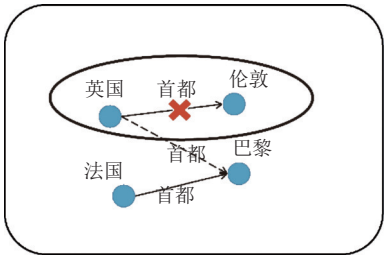

图 4 数据集错误样例构造示意图

3.2 实验环境设置

由于嵌入模型的内存占用一般随知识图谱实体、关系数量线性增长, 本文提供了模型运行的条件要求: 由于训练所用时间或内存至少随着知识图谱大小线性增长, 因此本文设定 GPU 内存限制大小为 4 GB, 时间限制为 5 h.

在嵌入模型的实现以及超参数设定上, 本文对 OpenKE[22]中的实现进行一定修改, 利用原有的超参数设定 (按第 2.1 节框架中讨论, 训练轮次降低为原 1/10), 对于 KGTtm[9], CKRL[16], CAGED[23], KGClean[24]等未在 OpenKE 库中实现的方法, 本文根据文章描述进行了方法整合以实现对比. 其余模型的定义可参见相关综述[18]以及已有实验[25].

根据问题定义, 知识图谱准确性评估存在如下指标 (*ERR* 和 *ACC* 的定义与前文相同, 为主要参考指标).

● *ERR* 代表最终评估的误差情况, 反映评估结果的优劣, 计算公式为:

$$ERR = |\mu_w(G) - \hat{\mu}_w(G)|_{\text{abs}} \tag{9}$$

其中, $|\cdot|_{\text{abs}}$ 表示计算数值的绝对值.

● *ACC* 代表标签正确率, 即在输出三元组标签中, 与事实一致的标签的占比 (本文使用 $\mu_w(G)$, $\hat{\mu}_w(G)$ 表示知识图谱真实的正确率和评估所得正确率, 此处 *ACC* 表示输出正误标签中, 与事实一致的标签比例 (即被正确判断的三元组比例), 请注意三者的含义以及符号表示), 反映 $\hat{f}(t) := Label_\lambda \circ Func_\theta(t)$ 的优劣, 计算公式如下:

$$ACC = \left|\left\{t \in \mathcal{S}_2(G) : \hat{f}(t) = f(t)\right\}\right|_w \Big/ |\mathcal{S}_2(G)|_w \tag{10}$$

其中, $|\cdot|_w$ 代表结合 w 权重计算集合大小, 即 $|S|_w = \sum_{t\in S} w(t)$.

● *ROC-AUC* 代表评分对于正确和错误三元组评分的区分程度, 直接反映 $Func_\theta$ 的优劣. 它的定义为 ROC 曲线 (receiver operating characteristic curve, 受试者工作特征曲线) 下的面积, 曲线表示了评分模型分类结果随阈值选择的变化, 指标可被解释为:

$$ROC\text{-}AUC = \Pr(Func_\theta(t_1) > Func_\theta(t_0) | t_1 \in G_T, t_0 \in G_F) + \frac{1}{2}\Pr(Func_\theta(t_1) = Func_\theta(t_0) | t_1 \in G_T, t_0 \in G_F) \tag{11}$$

其中, $Func_\theta$ 代表嵌入模型, G_T, G_F 分别代表正确的三元组集和错误的三元组集.

● *COST* 代表评估所用的成本, 包括时间成本、人工投入、计算资源等, 记为 *COST*. 其中时间成本用评估所用时间表示, 计算资源用评估所需要的内存等表示, 该指标反映了评估方法在多个角度的评估成本, 进而限制了评估方法对于一定规模的知识图谱是否可用.

整体上, 本文实验既包括对于评估框架可行性的探讨, 又包括对不同条件下嵌入模型、数据集之间的对比, 同时包括结合重要性的准确性评估, 与其他补全任务/错误识别任务的实验相比, 有以下特点.

(1) 实验针对知识图谱准确性评估问题, 评估误差是评估方法的重要指标, 而其他任务中一般以排序中前 K 位的命中比例 (Hits@K) 为指标.

(2) 嵌入学习过程中在知识图谱训练集增加错误, 并针对常见知识图谱情况设计了相对极端的错误比例.

(3) 知识图谱准确性评估任务要求针对整体知识图谱进行评估, 考虑到知识图谱规模以及人工评估所需时间成本, 本文在训练中增加了额外的评估时间限制.

(4) 本文将多个嵌入模型应用于三阶段框架中, 对嵌入模型各阶段的评估结果、针对准确性评估问题的适配

性、阈值的选择等均进行了对比.

3.3 阈值选择策略对比

本文首先对 $\mu_{\text{uni}}(G)$ 评估情况进行实验, 在该条件下所有三元组的重要性权重无差别设置. 在训练集正确率分别为 60%, 80% 和 90% 条件下, 表 3 反映了嵌入模型三元组评分的 *ROC-AUC* 值. 由表 3 可见其大多处在较高水平, 说明大部分模型能够区分测试集中的正确或错误三元组, 其中 TransD, RotatE, CKRL 在该环境设置下效果普遍较好. 整体上在训练集含有错误的情况下, 嵌入模型依然具有较好鲁棒性.

表 3 嵌入模型在不同数据集中 *ROC-AUC* 指标 (排序前 3 的模型以粗体标明)

模型类型	模型	FB15K-237			NELL-995			YAGO3-10		
		90%	80%	60%	90%	80%	60%	90%	80%	60%
基于平移	TransE	0.842	0.822	0.805	0.700	**0.716**	0.705	**0.921**	**0.885**	**0.832**
	TransH	0.841	0.824	0.804	0.713	0.715	**0.707**	**0.911**	**0.888**	**0.824**
	TransD	**0.847**	**0.832**	**0.807**	0.711	0.702	**0.706**	**0.906**	**0.889**	**0.823**
基于乘法	SimplE	0.804	0.766	0.716	0.683	0.632	0.602	0.888	0.860	0.797
	ComplEx	0.804	0.754	0.684	**0.778**	**0.776**	0.618	—	—	—
	DistMult	0.800	0.768	0.726	0.655	0.625	0.583	0.881	0.863	0.797
其他方法	RotatE	**0.863**	**0.844**	**0.806**	0.654	0.625	0.593	0.904	0.874	0.797
	CKRL	**0.845**	**0.834**	**0.806**	**0.730**	**0.729**	**0.706**	—	—	—
	CAGED	0.740	0.735	0.731	**0.716**	0.697	0.696	0.614	0.584	0.625

其次, 对于第 2.2 节中提出的 3 种阈值选择策略: λ_0 为选择固定阈值进行划分, 动态阈值 λ_1 为借助负采样策略近似最优阈值, 修正阈值 λ_1' 基于更强的高斯分布假设, 在 λ_1 基础上进一步调整其与最优阈值之间的偏差, 进行对比. 本文将基于上述 λ_0, λ_1, λ_1' 的 $Label_\lambda$ 与基于线性层进行分类的 KGTtm 方法[9]、基于 GlobalOpt[17]标定阈值的方法、离群值检测方法 Z-score (超参数 K 选择 1.5) 这 3 类基线方法进行对比 (KGTtm 要求采用标注标签进行训练, 本文以现有三元组为正例, 加入构造三元组为负例进行训练; GlobalOpt 与 ACTC 选择阈值策略一致, 其区别主要在不同人工标注顺序影响了收敛速度, 此处以 GlobalOpt 作为代表; Z-score 假定数据为高斯分布, 将偏离期望值过多的值视作离群值, 为常用一维数据的离群值检测方法), 最终在各数据集上结果如表 4, 表中每一类阈值下 *ACC* 的最优值以及小于 10% 的 *ERR* 值突出显示. 为保证一致性, 令 $\mathcal{S}_1(G) := G - \mathcal{S}_2(G)$. 需注意表 4 中, 基于乘法的方法同时考虑其 λ_0, λ_1 的评估结果, CAGED 对于数据的要求使得此时 λ_1 计算困难, 且其输出分数范围已知 (经过 Sigmoid 后在 (0, 1) 间), 因此采用固定阈值 $\lambda_0 = 0.99$, 在表 4 中记作 CAGED*, 其他模型均仅采用 λ_1 阈值进行评估; 表 3 可见 KGClean 不适合在数据含有错误时进行评估, 故这里没有验证该方法.

通过实验发现, 当知识图谱正确率较高时大部分评估效果较好, 随着知识图谱正确率的下降, 评估误差增大, NELL-995 上 *ACC* 指标相对其他数据集模型效果明显下降, 说明数据集本身特性能够显著影响自动化评估方法的效果. GlobalOpt 标定结果有着普遍最好的性能, 说明 λ_{opt} 作为近似目标的合理性.

本文提出的基于嵌入模型和阈值优化的评估方法与基线方法 KGTtm 与 Z-score 相比, λ_0 对应的 *ACC* 指标下降, 但平均 *ERR* 指标在较为稠密的 FB15K-237 与 YAGO3-10 上降低 37.7%, λ_1 对应的 *ACC* 指标与基线方法效果接近, 但在 3 个数据集上平均 *ERR* 降低达到 29.0%, 同时在正确率较低的数据上有着相对更小的误差, λ_1' 的修正针对错误三元组较多的情况, 与 λ_1 相比 *ERR* 进一步下降, 说明 λ_0, λ_1, λ_1' 阈值选择策略的有效性.

比较 λ_1 与 λ_0 可见, 采用阈值 λ_1 时有着较好的 *ACC* 指标, 但其评估误差 *ERR* 在知识图谱正确率下降时评估误差大幅上升; 相反, 采用固定阈值 λ_0 的嵌入模型尽管 *ERR* 更小, 然而面对特殊数据 (例如较为稀疏的 NELL 数据集), 基于乘法的嵌入模型可能面临更严重的训练不足的情况, 其评估误差常处于较高水平, 说明了两类阈值选择策略的不同适用性.

对比 λ_1', λ_1 可见, 尽管修正阈值 λ_1' 调整所基于的假设较强, 但实验证明修正 λ_1' 相对于 λ_1 在 *ERR* 上有了较大改善, 在 $\mu_{\text{uni}}(G)$ 较低时仍能保持较小 *ERR*. 在 3 个数据集上平均 *ERR* 指标降低 8.9% 的同时, 在 FB15K-237, YAGO3-10 上评估误差降低 37.2%, $\mu_{\text{uni}}(G)$ 较低的数据中误差下降更为显著; 同时 λ_1' 的 *ACC* 指标相较于 λ_1 也普

遍提高, 部分数据中 λ_1' 的 *ACC* 相对下降是由于 $\mu_{\text{uni}}(G)$ 较高引起的标签不平衡, 但鉴于此时 λ_1' 也拥有与 λ_1 近似的效果, 而 $\mu_{\text{uni}}(G)$ 下降后对 $|Score_F < \lambda|_{\text{ineq}}$ 的估计与修正显著降低了评估误差, 因此本文认为该优化是简单而有效的.

表 4 基于阈值 λ_0, λ_1, λ_1' 以及基线方法评估结果 *ACC* 和 *ERR* 指标

指标	阈值	模型	FB15K-237			NELL-995			YAGO3-10		
			90%	80%	60%	90%	80%	60%	90%	80%	60%
ACC	λ_0	ComplEx	0.730	0.674	0.618	0.727	0.686	0.551	—	—	—
		SimplE	0.822	0.756	0.652	0.525	0.517	0.549	0.830	**0.798**	**0.732**
		DistMult	0.821	0.764	0.663	0.532	0.503	0.526	0.832	0.801	0.731
		CAGED*	**0.904**	**0.789**	**0.669**	**0.864**	**0.784**	**0.621**	**0.892**	0.794	0.654
	λ_1	TransE	0.910	0.856	0.721	0.752	0.711	0.664	**0.904**	**0.840**	**0.671**
		TransH	0.910	0.855	0.714	0.694	0.693	0.667	0.902	0.839	0.662
		TransD	**0.926**	0.856	0.721	**0.761**	**0.753**	**0.669**	0.896	0.833	0.657
		ComplEx	0.877	0.797	0.644	0.739	0.601	0.569	—	—	—
		SimplE	0.876	0.800	0.643	0.718	0.688	0.561	0.896	0.819	0.660
		DistMult	0.877	0.803	0.645	0.719	0.670	0.560	0.881	0.820	0.655
		RotatE	0.917	0.853	0.698	0.635	0.603	0.569	0.896	0.815	0.619
		CKRL	0.906	**0.861**	**0.728**	0.738	0.738	0.655	—	—	—
	λ_1'	TransE	**0.883**	0.840	**0.746**	**0.686**	**0.687**	**0.676**	**0.892**	0.845	0.708
		TransH	0.882	**0.844**	0.737	0.660	0.587	0.646	0.884	0.846	0.699
		TransD	0.864	0.837	0.741	0.617	0.671	0.671	0.881	**0.848**	**0.740**
	基线方法	KGTtm	0.905	0.813	0.617	0.876	0.743	**0.697**	—	—	—
		Z-score (TransE)	0.915	**0.859**	0.680	0.862	0.787	0.614	0.873	0.828	0.622
		GlobalOpt (TransE)	**0.926**	**0.859**	**0.749**	**0.899**	**0.790**	0.666	**0.905**	**0.851**	**0.754**
ERR	λ_0	ComplEx	0.221	0.209	0.111	0.173	0.314	0.297	—	—	—
		SimplE	0.101	**0.075**	**0.053**	0.434	0.389	0.254	0.124	**0.094**	**0.001**
		DistMult	0.102	**0.056**	**0.079**	0.422	0.401	0.277	0.117	**0.095**	**0.007**
		CAGED*	**0.005**	**0.029**	0.264	**0.005**	0.113	0.332	**0.009**	0.196	0.340
	λ_1	TransE	**0.006**	**0.077**	0.240	0.149	**0.100**	0.120	**0.003**	**0.078**	0.228
		TransH	**0.006**	**0.080**	0.240	0.224	0.122	**0.091**	**0.002**	**0.075**	0.282
		TransD	**0.096**	**0.056**	0.228	0.155	**0.027**	0.153	**0.003**	**0.085**	0.288
		ComplEx	**0.003**	**0.071**	0.244	0.141	0.192	**0.003**	—	—	—
		SimplE	**0.006**	**0.083**	0.246	0.174	**0.042**	**0.016**	**0.006**	**0.085**	0.251
		DistMult	**0.001**	**0.085**	0.251	0.166	**0.074**	**0.017**	**0.010**	**0.085**	0.272
		RotatE	**0.017**	**0.087**	0.236	0.271	0.190	**0.039**	**0.005**	**0.086**	0.290
		CKRL	**0.003**	**0.070**	0.215	0.172	**0.050**	**0.090**	—	—	—
	λ_1'	TransE	**0.036**	**0.013**	**0.038**	0.241	0.159	0.015	**0.034**	**0.033**	0.161
		TransH	**0.036**	**0.030**	0.178	0.272	0.230	**0.080**	**0.052**	**0.042**	0.210
		TransD	**0.060**	**0.003**	0.139	0.218	0.180	**0.028**	**0.057**	**0.044**	**0.121**
	基线方法	KGTtm	**0.094**	0.172	0.383	**0.089**	0.241	**0.038**	—	—	—
		Z-score (TransE)	**0.014**	**0.100**	0.307	**0.007**	0.122	0.341	**0.027**	0.124	0.353
		GlobalOpt (TransE)	**0.059**	**0.100**	**0.079**	**0.100**	0.195	0.131	**0.089**	**0.020**	**0.041**

3.4 不同嵌入模型与数据集对比

本节针对实验中的不同模型和数据进行进一步对比分析.

(1) 嵌入模型的对比与选择

首先, 嵌入模型之间进行对比时, 针对某一正确率分别取其对应的 3 个数据集上指标的平均值 (对于无数据的情况, 以其他模型除去最高、最低值的平均值代替, 由于 λ_0 与 λ_1' 均只针对一部分嵌入模型设置, 因此对比嵌入模型时以 λ_1 的性能为代表) 得到模型平均的 *ROC-AUC* 指标 (图 4(a))、平均基于 GlobalOpt 标定阈值的 *ACC* 指标

(记作 ACC_{opt}, 图 4(b)), 平均基于动态阈值 λ_1 的 ACC_1, ERR_1 指标 (图 4(c), (d)).

整体而言, 如图 5 所示, 各个指标相对于知识图谱真实正确率呈现明显的分层现象. 通过图 4 数据整体分析可见, 基于平移的模型与 CKRL 等相对于基于乘法的方法存在优势, 例如其 *ROC-AUC* 指标相对其余模型提升 8%. 在相关实验[22,25]中, 最为简单的 TransE 相对其他模型性能也具有竞争力, 本文中, TransE 性能与 TransH, TransD 接近, 这与现有其他工作的实验结果可能存在不同. 整体上模型建模能力与其实验性能的关系分析如下.

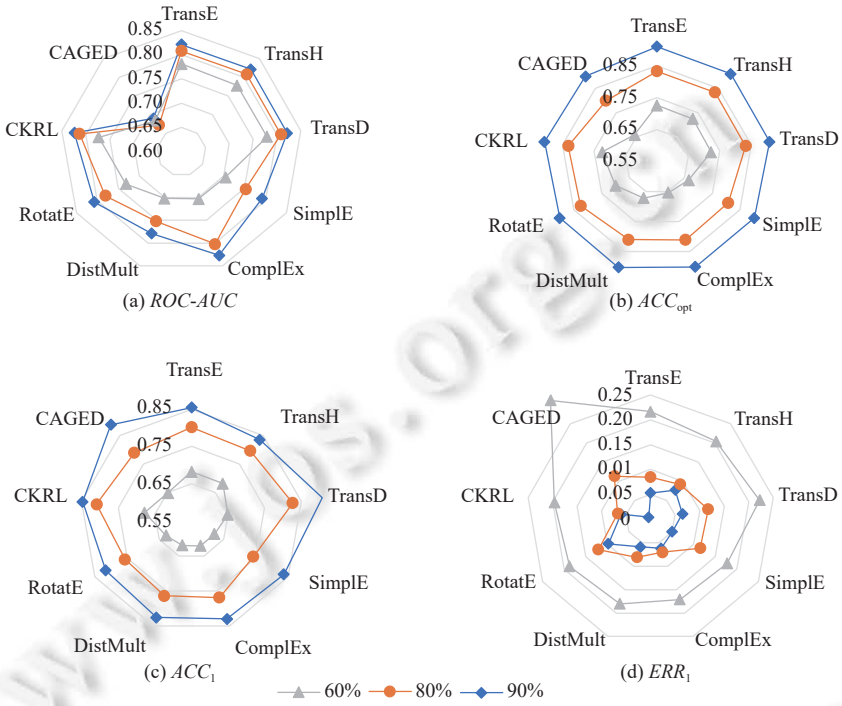

图 5 不同模型评估性能对比

1) 模型能否建模 1-N、N-N 关系不会显著影响模型性能, 与知识图谱补全问题、错误发现问题不同 (例如补全中常采用的指标 *Hits@K* 等, 侧重于排序后分数排名更靠前的部分三元组是否正确判断), 准确性评估任务不关注三元组评分的相对大小关系, 在三元组正误判断中允许 1-N 关系对应的多个三元组都满足设定阈值条件, 一定程度上降低了对建模此类关系的需求.

2) 模型能否建模关系的传递性可能显著地影响了模型的判断是否合理, 直观上看, 利用嵌入模型进行三元组正误判断即为衡量头尾实体之间路径所代表的多种关系的复合是否与当下询问的三元组中关系相匹配, 基于平移思想的模型大多满足该条件, 而基于乘法的方法中的 ComplEx 等模型不能建模关系传递性, 可能导致判断时出现更多的错误. 本文认为这导致了基于乘法的嵌入模型评估 *ACC* 指标较低.

3) 为了防止对称关系的干扰, 本文采用的 FB15K-237 数据集等有意识地去除了此类三元组, 可能导致难以建模对称关系的嵌入模型 (如 TransE) 评估性能高于其他知识图谱中的表现, 但由于实际情况中对称关系容易通过规则识别, 而增加对称关系则可能导致模型指标不合理的提升[25], 因此本文认为该简化是合理的.

在模型评估误差能力之外, 不同嵌入模型涉及运算不同在评估过程中所用时间存在区别, 图 6 展示了在 FB15K-237 上不同模型的运行时间以及人工评估 200 个三元组大致所用时间[26]. 可见, 此规模下大部分模型所用时间低于人工评估所用时间, 考虑不同 w, 人工方法所需样本规模会进一步扩大. 但 CKRL, KGTtm 等应用图结构特征的模型往往因为预处理计算量大、采用 CPU 等原因, 耗时多于人工评测用时. 尽管目前其他大部分嵌入模型自动化处理方法评估效率优于人工用时, 但知识图谱规模增大导致批处理的数据数量至少线性增大, 对于大规模

知识图谱上嵌入模型训练时间不可控. 人工方法只需评估单个三元组是否正确, 评估时间只与样本数量有关, 相对稳定. 因此嵌入模型训练集的抽样策略可作为一种重要的未来研究方向.

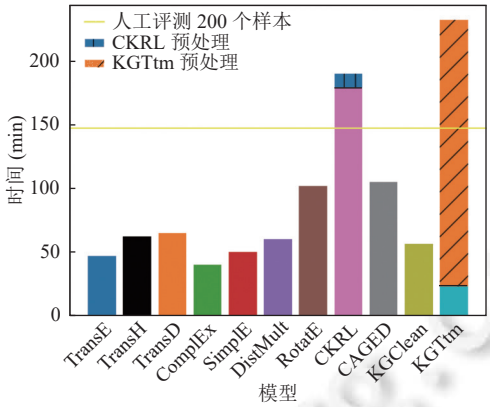

图 6 各模型评估时间对比

(2) 知识图谱数据集对比

为进一步探索不同数据集的作用, 本文取嵌入模型中除去最高、最低值后的均值作为缺省值, 分析数据集差异可能造成的影响 (如表 5 所示). FB15K-237 与 YAGO3-10 有着类似的指标范围与变化趋势, 而 NELL-995 在 *ACC* 层面显著低于剩余数据集, 平均 *ACC* 指标相对其他数据集下降 13%, 在 *ERR* 的变化趋势上也有着较高的随机性. 这说明了 NELL-995 数据集一定程度上并不适合此类自动化评估方法.

表 5 各数据集在不同模型下的指标平均值对比

数据集	正确率 (%)	ACC_1	ERR_1
FB15K-237	90	**0.901**	**0.024**
	80	0.828	**0.081**
	60	0.670	0.269
NELL-995	90	**0.764**	0.149
	80	0.702	**0.033**
	60	0.623	**0.075**
YAGO3-10	90	**0.896**	**0.005**
	80	0.825	**0.094**
	60	0.692	0.286

该现象可能源于 NELL-995 的稀疏性, 相关文献 [27] 中通过定义与实验分析, 说明了平均每个实体、每类关系对应的三元组数量可以一定程度反映知识图谱的相对稠密性, 其他条件不变的情况下, 对三元组进行增减能够影响嵌入模型的性能, 这使得稀疏性成为指导嵌入模型是否可以使用的指标, NELL-995 数据集相对其他两个知识图谱, 平均每个实体、每个关系对应的三元组数量较少, 一方面相关文献 [27] 中的实验结果说明, 嵌入模型本身的训练质量依赖于密集的训练数据, 稀疏性可能导致第二阶段 (嵌入模型训练步骤) 存在缺陷; 另一方面, 由于此时数据集中存在错误三元组, 数据较少时对错误三元组训练易过拟合, 导致第 3 阶段分类时更加难以分离正确与错误三元组, 导致真实正确率下降时评估误差情况与其他数据集表现不同. 可见将嵌入模型应用于知识图谱时, 待评估知识图谱的正确率 ($\mu_{\text{uni}}(G)$) 以及知识图谱自身的稠密程度 (例如实体、关系平均所对应的三元组数) 都影响着嵌入模型应用的性能. 直观上看, 嵌入模型利用知识图谱内部信息验证样本三元组是否正确, 其他三元组错误过多或提供的可用信息不足均会影响嵌入模型的性能. 因此图谱的先验正确率区间与图谱的稀疏性这两个特征可指导评估人员衡量是否适宜利用嵌入模型进行直接评估.

3.5 结合三元组重要性的准确性评估

在增加权重后, 嵌入模型在不同重要性定义下的 ACC_1 指标如表 6 所示 (λ_0, λ_1' 阈值只适用于部分模型, 因此

这里只展示基于 λ_1 阈值下的 *ACC* 指标; CAGED 在 NELL-995 上效果远低于其他模型, 不具有参考价值, 因此略去), 结果证明在知识图谱正确率较高的情况下, 利用嵌入模型结合三元组重要性进行评估具有一定可行性.

表 6 模型在网络结构加权 w_s/关系语义加权 w_r 重要性定义下的 ACC_1 指标

模型类型	模型	w选择	FB15K-237			NELL-995			YAGO3-10		
			90%	80%	60%	90%	80%	60%	90%	80%	60%
基于平移	TransE	w_{uni}	0.910	0.856	0.721	0.752	0.711	0.664	**0.904**	**0.840**	**0.671**
		w_s	0.910	0.823	0.751	0.746	0.700	0.651	**0.921**	0.841	0.631
		w_r	0.901	0.762	0.653	0.657	0.610	0.512	**0.890**	0.881	0.635
	TransH	w_{uni}	0.910	0.855	0.714	0.694	0.693	0.667	0.902	0.839	0.662
		w_s	0.903	0.854	0.724	0.695	0.682	**0.661**	0.918	**0.856**	0.664
		w_r	0.879	0.798	0.598	0.648	0.616	0.543	0.876	0.867	0.651
	TransD	w_{uni}	**0.926**	0.856	0.721	**0.761**	**0.754**	**0.669**	0.896	0.833	0.657
		w_s	**0.929**	0.854	**0.755**	**0.747**	**0.735**	0.653	0.917	0.840	**0.652**
		w_r	0.902	0.774	**0.710**	0.648	0.618	0.520	0.888	**0.888**	**0.654**
基于乘法	ComplEx	w_{uni}	0.877	0.797	0.644	0.739	0.601	0.569	—	—	—
		w_s	0.873	0.790	0.653	0.738	0.605	0.584	—	—	—
		w_r	0.877	0.738	0.526	0.678	0.618	0.507	—	—	—
	SimplE	w_{uni}	0.876	0.800	0.643	0.718	0.688	0.561	0.896	0.819	0.660
		w_s	0.879	0.793	0.646	0.722	0.574	0.566	0.872	0.821	0.637
		w_r	0.866	0.728	0.516	0.673	0.594	0.488	0.868	0.847	0.616
	DistMult	w_{uni}	0.877	0.803	0.645	0.719	0.670	0.560	0.881	0.820	0.655
		w_s	0.881	0.802	0.652	0.682	0.667	0.565	0.900	0.810	0.668
		w_r	0.879	0.759	0.515	0.677	0.575	0.494	0.884	0.846	0.619
其他方法	RotatE	w_{uni}	0.917	0.853	0.698	0.635	0.603	0.569	0.896	0.815	0.619
		w_s	0.922	0.857	0.706	0.653	0.603	0.572	0.901	0.840	0.610
		w_r	**0.913**	0.807	0.573	0.669	0.583	0.528	0.874	0.888	0.576
	CKRL	w_{uni}	0.906	**0.861**	**0.728**	0.738	0.738	0.655	—	—	—
		w_s	0.907	**0.868**	0.740	0.736	0.722	0.646	—	—	—
		w_r	0.909	**0.844**	0.694	**0.685**	**0.626**	**0.533**	—	—	—
	CAGED*	w_{uni}	0.904	0.789	0.669	—	—	—	0.892	0.794	0.654
		w_s	0.906	0.690	0.706	—	—	—	0.711	0.623	0.618
		w_r	0.903	0.703	0.709	—	—	—	0.667	0.584	0.568

注: 粗体为基于结构加权或关系语义加权下各数据集对应的最高指标

除 NELL-995 外, 数据集表现整体较好, 与表 4 对比可见基于网络结构重要性 w_s 的模型接近基于均匀分布 w_{uni} 模型的性能, 而基于关系语义重要性 w_r 模型表现略有下降. 说明在只考虑图上的连接关系权重时, 无论是对于 PR 分数较高, 易被访问的三元组, 还是对 PR 分数较低, 重要性较低的三元组, 嵌入模型性能整体结果保持稳定. 而与基于关系语义的加权 w_r 对比可见, 其 ACC_1 指标往往低于 w_s和w_{uni} 的对应指标, 平均 ACC_1 整体下降 5.3%, 这证明了能够为实体提供独特信息的三元组往往难以被嵌入模型根据统计信息正确判断, w_r 较高的三元组在嵌入模型的判断过程中可能更容易出现错误. 尽管基于 w_r 加权时嵌入模型自动化评估存在性能下降的风险, 但反之利用 w_r 的分布, 也可能筛选出易判断错误的三元组集合, 侧面提升模型的性能.

比较 3 个加权方式下不同模型的性能, 可发现: 相对于均匀加权模型性能排名基本保持不变, TransD 和 CKRL 均呈现较好评估结果. 在 w_{uni} 与 w_s 权重赋值下, TransD 在最多数据集上表现突出, CKRL 则在 w_r 上各数据集中表现突出, 在不考虑 NELL-995 数据集时, TransE 与 TransD 取得各权重赋值下突出的结果, 说明 TransE、TransD、CKRL 等更多关注到 w_s和w_r 赋值下较为重要的三元组, 相对其他方法在此类重要性赋值下表现更优.

4 相关工作

准确性是重要的知识图谱质量维度, 正确率是评估准确性的重要指标[28–31]. 知识图谱准确性相关方法分为 5 类, 包括随机抽样检测、规则辅助抽样、外部数据验证、图依赖关系验证、嵌入模型衍生方法. 本节围绕图 1 中知识图谱准确性评估流程各环节的实现, 对相关方法的代表性工作进行了调研分析.

4.1 随机抽样检测

随机抽样检测法主要包括简单随机抽样方法、多阶段加权抽样方法等.

简单随机抽样检测方法将知识图谱视作三元组集合, 基于人工核验判断三元组正误. 根据抽样检测的理论分析, 采用有放回随机抽样时, 最终评估的置信区间只与样本大小以及样本的标注结果相关, 不受知识图谱整体三元组数量的影响, 证明了抽样检测方法对于大规模知识图谱的可行性.

多阶段加权抽样方法[26]针对知识图谱准确性评估问题提出将三元组按照头实体分组后、按照分组三元组数量作为权重进行抽样检测, 避免了多次重复识别头部实体, 进而优化了以人工核验标注作为判断依据时样本评估的效率. 该工作首次从人工标注三元组用时的角度定义了高效准确性评估任务的优化目标, 并比较了大量抽样方法在统计学意义上理论要求的最低样本数量, 说明了 TWCS 理论上的高效. 除 TWCS 方法及其误差分析外, 该工作同时优化了知识图谱内容动态增加情况下的增量抽样检测方法.

4.2 规则辅助抽样

规则辅助下的抽样法主要包括 KGEval 方法、人工-机器协同工作方法等.

KGEval 方法: Ojha 等人[5]提出 KGEval 方法, 以人工核验、外部规则 (该方法中采用实体类型约束、形如 $(h_1,r_1,t_1)\wedge(t_1,r_2,t_2)\rightarrow(h_1,r_3,t_2)$ 的霍尔语句) 以及已有的标注结果为判断依据, 利用三元组之间可能存在的依赖关系, 减少人工标注的数量. KGEval 针对整体知识图谱进行三元组正误判断, 在单个三元组判断过程中, 若已标注三元组结果满足外部规则中的推理前提, 则该规则的推理结果即可证明另一三元组的正误, 从而减少额外的人工标注数量, 因此验证阶段三元组选择的顺序为方法整体效率的重要影响因素. KGEval 利用启发式规则进行待标注三元组的选取, 降低了人工标注的数量.

人工-机器协同工作方法: Qi 等人[32]综合 TWCS、KGEval 两种方法的优势, 提出利用基于最优化问题的人工-机器协同工作方法. 方法以人工核验及外部规则作为判断语句, 以存在推理关系的三元组分为一组, 以分组中三元组数量进行加权采样; 在匹配验证阶段, 该方法利用蒙特卡洛算法进行搜索, 选择下一步人工标注的三元组, 从而尽可能先标注外部规则的推理前提中的三元组, 推理出更多的未标注三元组, 减少人工标注的数量.

4.3 外部数据验证

外部数据验证下的准确性评估法主要包括 TAA 方法、ProVe 方法等.

TAA 方法: Liu 等人[8]提出利用其他的知识图谱与当前待评测知识图谱之间实体的 SameAs 链接, 完成其他知识图谱到当前知识图谱的验证. 该方法以其他若干知识图谱作为判断依据, 在匹配验证阶段将当前三元组中的头尾实体通过 SameAs 链接找到其他知识图谱当中的对应实体, 并利用字符串匹配程度衡量其他知识图谱中三元组与当前待验证三元组的一致程度; 分类阶段, 通过人工预设阈值, 将上阶段中多个外部知识图谱的匹配程度进行综合, 在匹配数量满足阈值要求时, 判断三元组为正确三元组.

ProVe 方法[33]: 该方法以网络信息作为判断依据进行事实验证, 在匹配验证阶段, 以三元组为关键词搜索获取对应网页, 并基于字符串匹配策略衡量该网页是否支持待验证三元组; 在分类阶段, 通过预设阈值, 筛选得到足够网页数量支持的三元组为正确三元组.

4.4 图依赖关系验证

图依赖关系验证法, 即利用知识图谱规则等条件进行验证. Fan 等人[34,35]通过对一般的属性依赖增加其所需要满足的子图结构, 将一般知识库中各列属性的依赖关系迁移至知识图谱结构上, 从而进行规则发现、矛盾发现. 该方法以三元组实体关系、属性值等方面的依赖关系作为判断依据 (在实验中该依赖关系源于对知识图谱内部三

元组信息的挖掘), 通过属性与实体的依赖关系判断当前三元组是否违背了依赖条件、是否存在推理规则支持该三元组. 尽管图依赖关系能够发现知识图谱中不符合约束条件的情况, 然而在分类阶段缺少对应的实现方式, 一种较为简单的分类即为将所有符合依赖关系约束条件的三元组均判断为正确三元组.

4.5 嵌入模型衍生方法

由于嵌入模型自身难以完成分类阶段, 嵌入模型衍生方法主要实现按照阈值或特征的分类, 常见如 KGTtm 方法、ACTC 方法等.

KGTtm 方法: Jia 等人[9]的工作在嵌入模型方法基础上增加更多知识图谱结构、路径特征辅助判断, 并设计分类器实现分类阶段. 具体的, 在验证阶段对三元组进行评分时利用了知识图谱 3 方面的特征: 图谱结构特征、嵌入模型对三元组的评分、头尾实体节点最短路径上的节点信息, 并利用神经网络融合 3 方面特征实现分类器, 进而给出输入三元组是否可信的判断, 在分类阶段利用已标注数据训练分类器, 因此需要人工标注.

ACTC 方法: Sedova 等人[10]在匹配验证阶段采用嵌入模型, 利用图谱内部结构信息进行验证, 在分类阶段提出阈值标定任务, 采用人工核验三元组是否正确的方式, 获得部分三元组标注数据, 确定近似的最优分类阈值. ACTC 方法为目前阈值标定任务中的最先进方法, 在一般的简单随机抽样进行阈值标定的基础上, 利用线性回归等方式辅助判断未标注数据, 并根据已判断三元组选择下一个待标注三元组, 避免样本中三元组评分分布存在倾斜, 从而更快收敛至最优阈值.

4.6 评估方法多维度对比

本节对比知识图谱准确性评估方法的可信程度、评估效率、灵活性等多个维度, 说明当前方法存在的优势与不足之处.

比较各类方法对三元组判断的能力如表 7 所示. 其中针对整体知识图谱, 比较其评估结果是否无偏; 针对每种错误类型, 比较是否能够被方法完整判断 (即不存在难以识别该类型错误、无法给出判断结果的情况)、判断结果是否可信 (即识别结果合理程度) 两个维度; 针对正确三元组, 按照判断可信程度进行比较. 针对其中判断结果可信程度这一维度, 表格利用“+”的数量衡量其判断结尾为基本可信 (“+++”)、较为可信 (“++”) 与可信程度一般 (“+”), 若判断结果与是否有该类型错误无关, 则划去该表格.

表 7 方法对三元组判断可信程度对比

方法类型	评估是否无偏	判断完整性			判断结果可信程度			
		结构错误	语义错误	实体/属性层面错误	结构错误	语义错误	实体/属性层面错误	正确三元组
I. 随机抽样检测	√	√	√	√	+++	+++	+++	+++
II. 规则辅助抽样	√	√	√	√	+++	+++	+++	+++
III. 外部数据验证	—	—	—	—	+	—	+	+++
IV. 图依赖关系验证	—	—	—	—	+++	+++	+++	+
V. 嵌入模型衍生方法	√	√	√	—	++	++	—	++

由表 7 可见, 在判断完整性上, 随机抽样检测以及规则辅助抽样两类方法针对正确三元组以及各种类型错误都能进行完整识别判断; 基于外部数据验证以及图依赖关系验证的方法针对多种错误存在无法完全进行识别的情况, 这是由于其判断依据与待判断三元组之间的匹配仅为全部三元组的一部分, 导致其判断往往局限于知识图谱中的一个子集, 对于剩余三元组既无法证真或判伪; 嵌入模型衍生方法通过知识图谱内部结构验证判断三元组层面错误, 但无法判断实体/属性层面错误.

在判断结果可信程度的对比上, 随机抽样检测以及规则辅助抽样两类方法具有最优的错误识别能力, 基于外部数据验证的方法能够较为可信地识别正确三元组, 但针对错误三元组往往缺少判断能力, 基于图依赖关系验证的方法则与之相反, 在本文目标类型错误上嵌入模型衍生方法有着较为均衡的判断能力.

比较各类方法评估效率如表 8 所示. 针对方法人工标注的时间成本、计算机计算的时间成本两方面, 针对三

元组正误判断中匹配验证、分类两阶段进行对比. 此处“+”数量仅反映了某一成本在各方法之间定性估计的相对大小, 因此不代表实际比例, 同时标注时间成本与计算成本之间存在数量级的差距, 无法直接对比.

表 8　不同方法评估效率对比

方法类型	匹配验证阶段		分类阶段	
	标注时间成本	计算时间成本	标注时间成本	计算时间成本
I. 随机抽样检测	+++	+	—	+
II. 规则辅助抽样	++	++	—	+
III. 外部数据验证	—	++	+	+
IV. 图依赖关系验证	—	+++	+	+
V. 嵌入模型衍生方法	—	+++	++	+

根据表 8, 规则辅助抽样的方法相对于随机抽样检测利用额外的计算机计算以降低人工标注时间成本; 外部数据验证方法在分类阶段需要人工设定正确三元组需满足的匹配比例, 因此存在较少的标注时间成本与计算时间成本; 基于图依赖关系验证的方法与嵌入模型衍生方法均需要对知识图谱整体三元组进行信息挖掘, 因此拥有更高的计算时间成本; 现有嵌入模型衍生的方法在分类阶段需要更多人工标注, 因此其标注时间成本更高. 一般情况下, 人工评估时间成本相对于计算时间成本影响更大, 嵌入模型衍生方法需借助少量人工标注完成分类阶段, 使得标注时间成本成为嵌入模型衍生方法评估效率上的瓶颈.

最后比较各类方法灵活性如表 9 所示. 整体上, 方法的灵活性取决于其是否能够简单地迁移到另一个知识图谱、另一个领域, 不需要额外进行调整, 针对方法在验证阶段、分类阶段是否需要额外人工核验、是否需要特定领域信息、是否需要人工调整参数进行衡量, 以对比方法灵活性.

表 9　方法灵活性对比

方法类型	匹配验证阶段			分类阶段		
	需要人工核验/标注	需要特定额外领域信息	需要额外人工设定参数	需要人工核验/标注	需要特定额外领域信息	需要额外人工设定参数
I. 随机抽样检测	√	√	—	—	—	—
II. 规则辅助抽样	√	√	—	—	—	—
III. 外部数据验证	—	√	√	—	—	√
IV. 图依赖关系验证	—	—	√	—	—	√
V. 嵌入模型衍生方法	—	—	—	√	—	—

随机抽样检测、规则辅助抽样的方法在验证阶段需要人工标注以及部分规则, 因此对于新的知识图谱或领域需要重新获取领域信息与人工标注, 灵活性较差; 外部数据验证方法在验证阶段需要用于匹配的外部领域信息, 并在匹配验证、分类阶段根据外部信息的数量修改匹配策略, 其灵活性也有一定损失; 图依赖关系验证方法与之类似, 在匹配验证阶段调整图依赖关系挖掘中的参数, 在分类阶段需要根据依赖关系数量调整分类阈值; 嵌入模型衍生方法基于知识图谱内部信息进行建模, 验证阶段不需要人工标注或特定领域信息, 但在分类阶段, 为将三元组按照嵌入模型评分进行分类, 需要进行额外人工标注以确定分类阈值, 存在一定灵活性损失.

整体上看, 基于嵌入模型的方法满足自动化的要求, 平衡了各个维度的性能, 本文将其作为匹配验证的主体, 然而现有具体方法在分类阶段均要求人工标注信息, 成为其评估过程中的一大瓶颈, 对相关工作的分析说明了本文阈值选择策略在嵌入模型衍生方法中分类阶段的重要性.

5 总　结

目前的知识图谱正确率定义往往仅考虑了正确三元组的比例, 忽略了不同三元组的重要程度对知识图谱整体使用的影响, 本文结合三元组重要性, 提供更为丰富的知识图谱质量信息评估策略, 但任务要求的提高需要高效而自动化的评估策略. 考虑到当前评估方法大多基于人工, 结合三元组重要性时评估低效的问题进一步凸显, 本文提

出基于嵌入模型的准确性评估方法, 并通过错误三元组形式假设构建负样本, 自动化选择阈值, 实验证明了本文选取的阈值选择策略相对于无监督分类器或离群值检测方法的优越性. 此外在实验中, 本文衡量了一般情况下以及结合三元组重要性时的准确性评估情况, 给出不同情况下的方法评估误差与运行成本进行分析, 阐明当前嵌入模型衡量的优势、劣势以及在不同需求下、不同数据中对于嵌入模型选择的建议.

未来进一步研究, 将在评估误差、评估代价等方面开展研究, 以提高评估的可信性, 减少影响评估的偶然因素. 具体研究内容可包括: 进一步优化嵌入模型, 提高其进行推理、验证的能力; 利用图谱过程中的信息来源等, 提高 $Label_\lambda$ 判断的能力. 针对知识图谱稀疏时评估困难的问题, 本文认为数据方面通过数据增强的方式, 通过引入外来知识图谱、进行知识图谱补全, 模型方面, 使用更为鲁棒的嵌入模型等方式, 可以降低稀疏性带来的影响. 此外, 三元组重要性的衡量是一个仍在探索的领域, 尽管目前实体重要性衡量仍主要采用网络结构与语义关系作为描述重要性的重要特征, 构造含义更加丰富的三元组重要性定义并探究其对本文框架的影响, 仍然是一个未来待探究的方向. 本文在未来进一步研究中希望通过进一步引入其他重要性特征、融合多种重要性特征等方式以实现更丰富的度量.

References:

[1] Suchanek FM, Kasneci G, Weikum G. YAGO: A core of semantic knowledge. In: Proc. of the 16th Int'l Conf. on World Wide Web. Banff: ACM, 2007. 697–706. [doi: 10.1145/1242572.1242667]

[2] Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes PN, Hellmann S, Morsey M, van Kleef P, Auer S, Bizer C. DBpedia—A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web, 2015, 6(2): 167–195. [doi: 10.3233/SW-140134]

[3] Carlson A, Betteridge J, Kisiel B, Settles B, Hruschka E, Mitchell T. Toward an architecture for never-ending language learning. In: Proc. of the 24th AAAI Conf. on Artificial Intelligence. Atlanta: AAAI, 2010. 1306–1313. [doi: 10.1609/aaai.v24i1.7519]

[4] Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J. Freebase: A collaboratively created graph database for structuring human knowledge. In: Proc. of the 2008 ACM SIGMOD Int'l Conf. on Management of Data. Vancouver: ACM, 2008. 1247–1250. [doi: 10.1145/1376616.1376746]

[5] Ojha P, Talukdar P. KGEval: Accuracy estimation of automatically constructed knowledge graphs. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017. 1741–1750. [doi: 10.18653/v1/D17-1183]

[6] Wang YQ, Ma FL, Gao J. Efficient knowledge graph validation via cross-graph representation learning. In: Proc. of the 29th ACM Int'l Conf. on Information & Knowledge Management. ACM, 2020. 1595–1604. [doi: 10.1145/3340531.3411902]

[7] Gerber D, Esteves D, Lehmann J, Bühmann R, Usbeck R, Ngomo ACN, Speck R. DeFacto-temporal and multilingual deep fact validation. Journal of Web Semantics, 2015, 35: 85–101. [doi: 10.1016/j.websem.2015.08.001]

[8] Liu SY, d'Aquin M, Motta E. Measuring accuracy of triples in knowledge graphs. In: Proc. of the 1st Int'l Conf. on Language, Data, and Knowledge. Galway: Springer, 2017. 343–357. [doi: 10.1007/978-3-319-59888-8_29]

[9] Jia SB, Xiang Y, Chen XJ, Wang K, Shi J. Triple trustworthiness measurement for knowledge graph. In: Proc. of the 2019 World Wide Web Conf. San Francisco: ACM, 2019. 2865–2871. [doi: 10.1145/3308558.3313586]

[10] Sedova A, Roth B. ACTC: Active threshold calibration for cold-start knowledge graph completion. In: Proc. of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto: ACL, 2023. 1853–1863. [doi: 10.18653/v1/2023.acl-short.158]

[11] Li Q, Li YL, Gao J, Su L, Zhao B, Demirbas M, Fan W, Han JW. A confidence-aware approach for truth discovery on long-tail data. Proc. of the VLDB Endowment, 2014, 8(4): 425–436. [doi: 10.14778/2735496.2735505]

[12] Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: ACM, 2013. 2787–2795.

[13] Trouillon T, Welbl J, Riedel S, Gaussier E, Bouchard G. Complex embeddings for simple link prediction. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York: PMLR, 2016. 2071–2080.

[14] Schlichtkrull M, Kipf TN, Bloem P, van den Berg R, Titov I, Welling M. Modeling relational data with graph convolutional networks. In: Proc. of the 15th Int'l Conf. on the Semantic Web. Heraklion: Springer, 2018. 593–607. [doi: 10.1007/978-3-319-93417-4_38]

[15] Sun ZQ, Deng ZH, Nie JY, Tang J. RotatE: Knowledge graph embedding by relational rotation in complex space. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019.

[16] Xie RB, Liu ZY, Lin F, Lin LY. Does William Shakespeare REALLY write Hamlet? Knowledge representation learning with confidence.

In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2017. 4954–4961. [doi: 10.1609/aaai.v32i1.11924]

[17] Speranskaya M, Schmitt M, Roth B. Ranking vs. Classifying: Measuring knowledge base completion quality. In: Proc. of the 2020 Conf. on Automated Knowledge Base Construction. AKBC, 2020. [doi: 10.24432/C57G65]

[18] Wang Q, Mao ZD, Wang B, Guo L. Knowledge graph embedding: A survey of approaches and applications. IEEE Trans. on Knowledge and Data Engineering, 2017, 29(12): 2724–2743. [doi: 10.1109/TKDE.2017.2754499]

[19] Park N, Kan A, Dong XL, Zhao T, Faloutsos C. Estimating node importance in knowledge graphs using graph neural networks. In: Proc. of the 25th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. Anchorage: ACM, 2019: 596–606. [doi: 10.1145/3292500.3330855]

[20] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: Bringing order to the Web. 1999. http://ilpubs.stanford.edu:8090/422/

[21] Li XT, Ng MK, Ye YM. HAR: Hub, authority and relevance scores in multi-relational data for query search. In: Proc. of the 12th SIAM Int'l Conf. on Data Mining. Anaheim: SIAM, 2012. 141–152. [doi: 10.1137/1.9781611972825]

[22] Han X, Cao SL, Lv X, Lin YK, Liu ZY, Sun MS, Li JZ. OpenKE: An open toolkit for knowledge embedding. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing: System Demonstrations. Brussels: ACL, 2018. 139–144. [doi: 10.18653/v1/D18-2024]

[23] Zhang QG, Dong JN, Duan KY, Huang X, Liu YZ, Xu LC. Contrastive knowledge graph error detection. In: Proc. of the 31st ACM Int'l Conf. on Information & Knowledge Management. Atlanta: ACM, 2022. 2590–2599. [doi: 10.1145/3511808.3557264]

[24] Ge CC, Gao YJ, Weng HH, Zhang C, Miao XY, Zheng BH. KGClean: An embedding powered knowledge graph cleaning framework. arXiv:2004.14478, 2020.

[25] Akrami F, Saeef MS, Zhang QH, Hu W, Li CK. Realistic re-evaluation of knowledge graph completion methods: An experimental study. In: Proc. of the 2020 ACM SIGMOD Int'l Conf. on Management of Data. Portland: ACM, 2020. 1995–2010. [doi: 10.1145/3318464.3380599]

[26] Gao JY, Li X, Xu YE, Sisman B, Dong XL, Yang J. Efficient knowledge graph accuracy evaluation. Proc. of the VLDB Endowment, 2019, 12(11): 1679–1691. [doi: 10.14778/3342263.3342642]

[27] Pujara J, Augustine E, Getoor L. Sparsity and noise: Where knowledge graph embeddings fall short. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017. 1751–1756. [doi: 10.18653/v1/D17-1184]

[28] Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 1996, 12(4): 5–33. [doi: 10.1080/07421222.1996.11518099]

[29] Zaveri A, Rula A, Maurino A, Pietrobon R, Lehmann J, Auer S. Quality assessment for linked data: A survey: A systematic literature review and conceptual framework. Semantic Web, 2016, 7(1): 63–93. [doi: 10.3233/SW-150175]

[30] Xue BC, Zou L. Knowledge graph quality management: A comprehensive survey. IEEE Trans. on Knowledge and Data Engineering, 2023, 35(5): 4969–4988. [doi: 10.1109/TKDE.2022.3150080]

[31] Acosta M, Zaveri A, Simperl E, Kontokostas D, Flöck F, Lehmann J. Detecting linked data quality issues via crowdsourcing: A DBpedia study. Semantic Web, 2016, 9(3): 303–335. [doi: 10.3233/SW-160239]

[32] Qi YF, Zheng WG, Hong L, Zou L. Evaluating knowledge graph accuracy powered by optimized human-machine collaboration. In: Proc. of the 28th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. Washington: ACM, 2022. 1368–1378. [doi: 10.1145/3534678.3539233]

[33] Amaral G, Rodrigues O, Simperl E. ProVe: A pipeline for automated provenance verification of knowledge graphs against textual sources. Semantic Web, 2024, 15(6): 2159–2192. [doi: 10.3233/SW-233467]

[34] Fan WF. Dependencies for graphs: Challenges and opportunities. Journal of Data and Information Quality, 2019, 11(2): 1–12. [doi: 10.1145/3310230]

[35] Fan WF, Hu CM, Liu XL, Lu P. Discovering graph functional dependencies. ACM Trans. on Database Systems (TODS), 2020, 45(3): 15. [doi: 10.1145/3397198]

附录 A. 评估误差范围分析

对于简单随机抽样方法, 若对于知识图谱 G 有 $|G|=N$, $\mu_1(G)=p$, $G=G_T+G_F$, G_T 代表知识图谱中的正确三元组集合, G_F 代表错误三元组集合, 则 $|G_T|=N\times p$, 如果令 $\mathcal{S}_2(G)$ 表示经过简单随机抽样方法抽取出的 n 个三元

组样本构成的测试集, 则 $\Pr(|\mathcal{S}_2(G)\cap G_T|=i)=C_n^i\times p^i\times(1-p)^{(n-i)}$, 根据大数定律, 伯努利分布在 n 较大的情况下趋近于高斯分布 (对于 $n>30$, $n\times p>5$), 即:

$$X=\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}\sim Gaussian(0,1) \tag{A1}$$

其中, $\hat{p}:=\mu_{\text{uni}}(\mathcal{S}_2(G))=i/n$, 于是有:

$$\Pr\left(-z_{1-\frac{\alpha}{2}}<X<z_{1-\frac{\alpha}{2}}\right)=\alpha,\ \forall\alpha\in[0,1]$$

$$\Pr\left(-z^*<\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}<z^*\right)=\alpha,\ \forall\alpha\in[0,1],\ z^*:=z_{1-\frac{\alpha}{2}} \tag{A2}$$

由于抽样所产生的误差 $err_1=|\mu_{\text{uni}}(\mathcal{S}_2(G))-\mu_{\text{uni}}(G)|_{\text{abs}}=|\hat{p}-p|_{\text{abs}}$, $\hat{p}(1-\hat{p})\leqslant\frac{1}{4}$, 因此对给定置信度 α 可以确定 z^*, 进而使得 $\Pr\left(|err_1|_{\text{abs}}<\frac{z^*}{2\sqrt{n}}\right)=\alpha$, 若要求 $|err_1|_{\text{abs}}<\varepsilon$, 可使其上界 $\frac{z^*}{2\sqrt{n}}\leqslant\varepsilon$, 计算对应的 n. 因此在置信度 α 的概率下, 通过确定 n 的范围使误差 err_1 被 $\frac{z^*}{2\sqrt{n}}$ 限制, 进而被 ε 限制, 此时 $n\geqslant\frac{z_{1-\frac{\alpha}{2}}^{\ 2}}{4\varepsilon^2}$.

例如给定 $\alpha=95\%$, 则根据 $\Pr(-z^*<X<z^*)=\alpha$, $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}=X\sim N(0,1)$, 有 $z^*=z_{1-\frac{\alpha}{2}}\approx1.96$, 若要求误差范围在 5%, 应当有 $\frac{z^*}{2\sqrt{n}}\leqslant0.05$, $n\geqslant384.16$. 即对于符合 $n\geqslant384.16$ 条件的 n 组随机抽取结果, 能够在 95% 的概率下认定利用该样本的评估误差在 5% 以内. 对于不同重要性定义 $w(t)$, 有类似的方式证明: 随着样本数量增加, 随机抽样样本与实际三元组集合之间评估误差会以大概率被限制在给定范围.

由于:

$$\begin{cases}ERR:=|\mu_w(G)-\hat{\mu}_w(G)|_{\text{abs}}\leqslant|\mu_w(G)-\mu_w(\mathcal{S}_2(G))|_{\text{abs}}+|\mu_w(\mathcal{S}_2(G))-\hat{\mu}_w(G)|_{\text{abs}}\\ \text{for a given }\epsilon>0,\ C>0,\ \exists\mathcal{S}_2(G),\ \Pr(|\mu_w(G)-\mu_w(\mathcal{S}_2(G))|_{\text{abs}}\geqslant\epsilon)\geqslant C\\ \hat{\mu}_w(G)=\dfrac{\sum_{t\in\mathcal{S}_2(G)}w(t)\hat{f}(t)}{|\mathcal{S}_2(G)|_w},\ \hat{f}(t)=Label_\lambda(Func_\theta(t))\\ |\mu_w(\mathcal{S}_2(G))-\hat{\mu}_w(G)|_{\text{abs}}=\left|\dfrac{\sum_{t\in\mathcal{S}_2(G)}w(t)\left(f(t)-\hat{f}(t)\right)}{|\mathcal{S}_2(G)|_w}\right|_{\text{abs}}\end{cases} \tag{A3}$$

因此, 对于给定的 ε, 随机选择符合条件的 n 个三元组构成评估样本, 则:

$$ERR\leqslant\varepsilon+\left|\frac{\sum_{t\in\mathcal{S}_2(G)}w(t)\left(f(t)-\hat{f}(t)\right)}{|\mathcal{S}_2(G)|_w}\right|_{\text{abs}} \tag{A4}$$

此时, 在整体误差 ERR 中的主要误差即为在 $\mathcal{S}_2(G)$ 上 $Label_\lambda\circ Func_\theta$ 与三元组真实正误 f 的区别.

附录 B. 模型与评估误差的关系

首先证明实验指标中标签正确率 ACC 与整体评估误差 ERR 的关系. 由于已知 $ERR=|\hat{\mu}_w(G)-\mu_w(G)|$, $ACC=\left|\left\{t\in\mathcal{S}_2(G):\hat{f}(t)=f(t)\right\}\right|_w/|\mathcal{S}_2(G)|_w$, 根据评估结果的计算方式与公式 (A3) 中的证明, 可见有:

$$ERR=|\hat{\mu}_w(G)-\mu_w(G)|_{\text{abs}}\leqslant\left|\frac{\sum_{t\in\mathcal{S}_2(G)}w(t)(Label_\lambda(Func_\theta(t))-f(t))}{|\mathcal{S}_2(G)|_w}\right|_{\text{abs}}+|\mu_w(\mathcal{S}_2(G))-\mu_w(G)|_{\text{abs}} \tag{B1}$$

省略后部分随机抽样导致的误差后 (该误差可通过样本规模被限制), 有 $ERR\approx\left|\sum_{t\in\mathcal{S}_2(G)}w(t)(Label_\lambda\circ Func_\theta(t)-f(t))\right|_{\text{abs}}/|\mathcal{S}_2(G)|_w$, 定义 $Label_\lambda\circ Func_\theta$ 对 $\mathcal{S}_2(G)$ 分类的情况 TP, TN, FP, FN 为:

$$\begin{cases} TP = |\{t \in \mathcal{S}_2(G) : Label_\lambda \circ Func_\theta(t) = f(t) = 1\}|_w \\ TN = |\{t \in \mathcal{S}_2(G) : Label_\lambda \circ Func_\theta(t) = f(t) = 0\}|_w \\ FP = |\{t \in \mathcal{S}_2(G) : Label_\lambda \circ Func_\theta(t) = 1 \text{ and } f(t) = 0\}|_w \\ FN = |\{t \in \mathcal{S}_2(G) : Label_\lambda \circ Func_\theta(t) = 0 \text{ and } f(t) = 1\}|_w \\ sum = |\mathcal{S}_2(G)|_w = TP + TN + FP + FN \end{cases} \tag{B2}$$

通过定义有:

$$\begin{cases} ERR \approx \left| \dfrac{TP+FP}{sum} - \dfrac{TP+FN}{sum} \right|_{\text{abs}} = \left| \dfrac{FP-FN}{sum} \right|_{\text{abs}} \\ ACC = \dfrac{TP+TN}{sum} = \dfrac{sum-FP-FN}{sum} \end{cases} \tag{B3}$$

所以对于 λ 的不同, $Label_\lambda$ 改变后可能出现 ERR 为 0 的情况, 此时对应着 $FP = FN$. 同时由于 $FP + FN = (1 - ACC) \times sum$, 则:

$$\begin{cases} |FP - FN|_{\text{abs}} \leqslant |FP|_{\text{abs}} + |FN|_{\text{abs}} = (1 - ACC) \times sum \\ ERR \approx |FP - FN|_{\text{abs}} / sum \leqslant 1 - ACC \end{cases} \tag{B4}$$

因此不考虑 $\mathcal{S}_2$ 选择所导致的误差时, 模型评估误差 ERR 不会超过 $1 - ACC$. 这是一个相当宽松的上界: 由于 $FP, FN \geqslant 0$, $|FP - FN|_{\text{abs}} \leqslant \max(FP, FN)$, 因此仅在 FP 或 FN 为 0 时取得等号.

不等式 (B4) 也证明, 若假设同一模型在不同数据集上, 评估三元组正误的正确判断 ACC 接近定值时, 该模型评估误差存在上界 $1 - ACC$, 且模型判断三元组正误的正确率 ACC 越高时, 评估误差上界越小, 评估越可靠.

张明韬(2002－), 男, 博士生, 主要研究领域为图数据分析, 时空数据分析.

白晓颖(1973－), 女, 博士, 研究员, 博士生导师, 主要研究领域为软件工程, 软件测试, 服务计算.

杨国利(1987－), 男, 博士, 助理研究员, CCF 专业会员, 主要研究领域为复杂网络结构, 图数据分析.