计算机科学
COMPUTER SCIENCE

基于生成式对抗网络和正类无标签学习的知识图谱补全算法

胡斌皓，张建朋，陈鸿昶

引用本文

胡斌皓，张建朋，陈鸿昶. 基于生成式对抗网络和正类无标签学习的知识图谱补全算法[J]. 计算机科学, 2024, 51(1): 310-315.

HU Binhao, ZHANG Jianpeng, CHEN Hongchang. Knowledge Graph Completion Algorithm Based on Generative Adversarial Network and Positiveand Unlabeled Learning [J]. Computer Science, 2024, 51(1): 310-315.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300006

基于生成式对抗网络和正类无标签学习的知识图谱补全算法

胡斌皓[1,2]　张建朋[2]　陈鸿昶[2]
1 郑州大学网络空间安全学院　郑州 450002
2 信息工程大学信息技术研究所国家数字交换系统工程技术研究中心　郑州 450002
(hu15181620732@163.com)

摘　要　随着知识图谱的应用越来越广泛,绝大多数真实世界的知识图谱通常具有不完备性,限制了知识图谱的实际应用效果。因此,知识图谱补全成为了知识图谱领域的热点。然而,现有方法大多集中在评分函数的设计上,少部分研究关注了负样本抽样策略。在改善负样本抽样的知识图谱补全算法的研究中,基于生成式对抗网络的方法取得了不错的进展。然而,现有研究并没有关注到负样本存在假阴性标签的问题,即生成的负样本中可能包含真实的事实。为了缓解假阴性标签问题,提出了一种基于生成式对抗网络和正类无标签学习的知识图谱补全算法。该方法利用生成式对抗网络生成无标签样本,并使用正类无标签学习缓解假阴性标签问题。在基准数据集上进行的大量实验证明了所提算法的有效性与准确性。

关键词:知识图谱补全;生成式对抗网络;正类无标签学习;负样本抽样
中图分类号　TP391.1

Knowledge Graph Completion Algorithm Based on Generative Adversarial Network and Positive and Unlabeled Learning

HU Binhao[1,2],ZHANG Jianpeng[2] and CHEN Hongchang[2]
1 School of Cyber Science and Engineering,Zhengzhou University,Zhengzhou 450002,China
2 National Digital Switching System Engineering & Technological R&D Center(NDSC),Institute of Information Technology,University of Information Engineering,Zhengzhou 450002,China

Abstract　With the widespread application of knowledge graphs,the majority of real-world knowledge graphs suffer from the problem of incompleteness,which hinders their practical applications. As a result,it makes knowledge graph completion become a hot topic in the field of knowledge graph. However,most existing methods focus on the design of scoring functions,with only a few studies paying attention to negative sampling strategies. In the research of knowledge graph completion algorithms which aims at improving negative sampling,the methods based on generative adversarial networks(GANs) have achieved significant progress. Nonetheless,existing studies have not addressed the false negative issue,meaning that generated negative samples may contain actual facts. To address this issue,this paper proposes a knowledge graph completion algorithm based on GAN and positive-unlabeled learning. In the proposed method,GANs are utilized to generate unlabeled samples,while positive unlabeled learning is employed to alleviate the false negative problem. Extensive experiments on benchmark datasets demonstrate the effectiveness and accuracy of the proposed algorithm.

Keywords　Knowledge graph completion,Generative adversarial network,Positive unlabeled learning,Negative sampling

1　引言

知识图谱以结构化的形式表示人类的知识,为人工智能系统提供其可处理的知识,使其具有像人类一样处理复杂问题的能力。目前,知识图谱在推荐系统、智能客服以及社会热点发现等领域有着广泛的应用。然而,现在的知识图谱都是依赖于人工构建或者半自动构建。虽然人工构建的知识图谱符合人类认知的直观逻辑,有着很好的解释性。但是人工构建成本高昂,不能满足实际需求,同时需要相关领域的专家协助。半自动方式构建的知识图谱虽然可以弥补以上不足,但是通常具有不完备性,即许多真实的事实被遗漏,难以保证所构建的知识图谱的质量。例如,Freebase 和 DBpedia 中超过 66%的人物条目缺少出生地[1]。这样的问题限制了知识图谱的实际应用,使得大量的学者投入到知识图谱补全任务的研究中。知识图谱补全任务旨在预测出知识图谱中缺失的、最合理的事实来补全图谱。知识图谱在形式上由众多实体以及实体之间的关系组成,通常将知识图谱表示为由头实体,关系,尾实体构成的三元组的集合。本文的研究旨在在给定

到稿日期:2023-03-01　返修日期:2023-06-13
基金项目:国家自然科学基金(62002384);嵩山实验室项目(221100210700-3)
This work was supported by the National Natural Science Foundation of China(62002384) and Song Shan Laboratory(221100210700-3).
通信作者:张建朋(j_zhang_edu@sina.com)

三元组中头(尾)实体和关系的情况下预测其他实体为缺失尾(头)实体的可能性。

近年来，对知识图谱补全任务的研究逐渐增多，使得该领域取得了巨大进步。目前对知识图谱补全算法的研究主要集中在设计各样的评分函数[2-7]或者通过与额外信息结合的方式提升嵌入向量的表征能力[8-10]。为了提高模型的鲁棒性和泛化能力，大多数模型需要进行对比训练。由于知识图谱中仅包含正确的三元组，即正样本三元组，但是在模型的训练过程中需要大量的负样本。负样本不仅可以提高模型的泛化性和鲁棒性，还可以加速模型收敛。然而，大多数现有研究并没有关注负样本抽样方式，而是选择将简单高效的均匀采样应用于各类知识图谱补全算法中。均匀采样指随机替换三元组中的头(尾)实体构造负样本，但它是一个固定方案，忽略了训练过程中负样本三元组的变化。均匀采样得到的样本集中的三元组大多数是容易被分类的。由于评分函数倾向于给正样本三元组较大的值，随着训练的进行，这些容易被分类的三元组将会迅速变得非常小，且对应的梯度会快速衰减至零[11]。在训练初期，当实体在语义空间中并没有分隔很开时，随机生成负样本的策略是可行的。但是，当训练到一定阶段之后，不同类型的实体在语义空间中彼此区分，随机生成的负样本可能很容易被模型辨别出来。如正三元组(中国，首都，北京)，经过随机生成的负样本为(中国，首都，火锅)，显然，在训练到一定阶段后，北京与火锅两个实体彼此之间分隔很开了，模型可以轻易地判断出这样的负样本，那么这样的负样本对这个阶段的训练没有任何帮助。此外，均匀采样还会导致得到假阴性标签的抽样的结果，即采样出的负样本三元组可能包含潜在的真实事实。

目前部分研究尝试使用生成式对抗网络来改进负样本的抽样策略。为了抽样出更高质量的负样本并缓解假阴性标签问题，本文受生成对抗模型和正类无标签学习的启发，提出了基于生成式对抗网络和正类无标签学习的知识图谱补全算法。首先，使用生成式对抗网络用于生成大量的三元组，这些三元组将作为候选三元组用于知识图谱的补全。其次，同大多数现有研究[3-6,12-13]一样，将知识图谱中已有的三元组视作正样本。不同的是，本文将新生成的三元组视作未知标记的样本，而非直接视为负样本。再次，本文利用正类无标签学习对正三元组和无标签三元组进行排序。最后在基准数据集上进行了详细的实验验证。实验结果表明，本文方法可以缓解假阴性标签问题对知识图谱补全任务的影响。本文的主要贡献如下：

(1)利用生成式对抗网络生成丰富的样本作为无标签样本，一定程度上缓解了知识图谱数据稀疏性的影响。

(2)为了缓解假阴性标签问题，提出了基于正类无标签学习的知识图谱补全算法，该算法通过针对成对排序优化的正类无标签学习获得候选三元组的最优排序来完成知识图谱补全任务。

(3)在真实的基准数据集上进行了广泛的实验评估和实验结果分析，验证了所提方法的有效性和准确性。

2 相关工作

2.1 知识图谱补全

知识图谱补全领域的研究主要集中在学习知识图谱中实体与关系的分布式表示[3-4,6]。从模型编码的角度上来看，分布式表示学习模型可以分为：(1)线性\双线性模型，其通过线性操作来编码实体和关系[4,12]。(2)张量分解的模型，其假设三元组的得分可以被分解成多个张量[2]。(3)几何模型，其通过头尾实体的转换建模关系[3,6]。TransE[3]在欧几里得空间中将关系定义为头尾实体之间的距离。RotatE[6]将实体嵌入到复数空间中，利用头尾实体的旋转定义关系。ATTH[7]认为在欧几里得空间嵌入知识图谱会导致数据失真，而双曲空间更适合知识图谱多层次结构的特性。因此，该模型将实体与关系映射到双曲空间，并取得了较好的结果。PJRE[9]通过结合规则挖掘、组合表示学习以及规则引导的优化，实现了知识图谱表示学习的高质量和有效推理。HAKE[10]通过考虑层次信息来提高链接预测的准确性，该模型将实体和关系嵌入分解为角度(Angle)和幅度(Modulus)两个部分。角度表示实体之间的语义关系，而幅度表示实体在层次结构中的位置。(4)深度神经网络模型，其利用深度神经网络嵌入知识图谱。MLP[14]将实体和关系共同编码进同一个全连接层，使用 Sigmoid 函数为三元组评分。ConvE[5]使用二维卷积嵌入和多层非线性特征，通过将头实体和关系重塑为二维矩阵来建模实体和关系之间的交互。RSN[15]设计了一种循环跳跃机制，通过区分关系和实体来增强语义表示。KG-BERT[8]借鉴了预训练语言模型的思想，利用 BERT 作为实体和关系的编码器。R-GCN[16]利用图神经网络对实体邻域和关系进行聚合。然而这些模型都没有关注负样本抽样，采用的都是均匀抽样，生成的负样本难以保证其质量。CompGCN[17]通过在图结构上执行卷积操作来捕捉实体和关系之间的上下文信息，同时利用复合函数来表示关系。

2.2 生成式对抗网络

生成式对抗网络最开始是为了在图像等连续空间中生成样本而提出的[18]。生成式对抗网络由生成器和判别器两个部分组成。生成器接受噪声输入并输出图像，判别器是一个二分类的分类器。在训练时，生成器和判别器进行一个最小化最大化博弈，其中生成器试图生成“真实”图像欺骗判别器，判别器试图将其与真实图像区别开来。SEQGAN[19]利用强化学习解决了生成式对抗网络不能生成离散样本的问题。KBGAN[20]将对抗学习框架引入知识图谱领域，判别器采用基于 Margin 的知识图谱嵌入模型(如 TransE[3])，而生成器使用的是基于概率和对数损失的嵌入模型(如 DisMult[12] 和 COMPLEX[4])。文献[21]的思想与 KBGAN 类似，但是其生成器使用的是一个两层的神经网络。然而，这些工作都未考虑假阴性标签问题。假阴性标签指一些真实的事实被错误地作为负样本，这限制了模型的性能。

2.3 正类无标签学习

正类无标签学习指模型只能访问正样本和无标签数据，而无标签数据包括正样本和负样本[22]。正类无标签学习主要包括两步求解法[23]和基于无偏风险估计的方法[24]。文献[25]利用代价敏感分类器框架，通过创立无偏风险评估来矫正分类器在正样本和无标签样本训练中产生的偏差。自文献[26]首次将无偏风险评估用于处理正类无标签学习问题以来，许多研究致力于改进此类方法。文献[27]提出的凸无偏风险评估可以降低计算成本。文献[23]利用非负风险估计量

可应对过度拟合问题。PUbN[28]构建了类似的无偏风险估计量，将小部分负样本用于训练分类器。文献[29]利用了少量有标注的数据进行关系学习，并通过无监督学习的方式自动生成大量的未标注数据。

知识图谱中仅包含了正三元组。为了满足正类无标签学习的需求，本文提出的方法利用生成式对抗网络模块产生大量无标签三元组。接下来，正类无标签学习模块对正三元组和无标签三元组进行排序，以便完成知识图谱的补全。

3 基于生成式对抗网络和正类无标签学习的知识图谱补全算法

本文提出了一种基于生成式对抗网络和正类无标签学习的知识图谱补全算法，该算法主要分为两部分：(1)生成式对抗网络模块，该模块通过添加随机噪声生成样本，这些样本中可能包含着正样本和负样本，因此不能直接视为负样本用于对比学习，本文将其视为无标签三元组；(2)正类无标签学习模块，该模块通过正类无标签学习获得三元组的最优排序以完成知识图谱补全任务。最后我们通过最小化最大化博弈训练所提模型。模型框架如图1所示。为了方便叙述，本文在后续章节中仅讨论替换尾实体的情况。

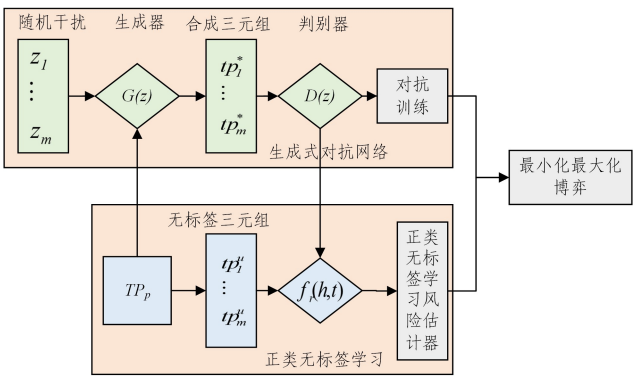


图1 GANPUL算法框架

Fig. 1 Framework of GANPUL

3.1 生成式对抗网络模块

知识图谱中仅包含正三元组，而知识图谱补全模型依赖于对比学习来训练模型，从而提高模型的准确率和泛化性。为了获取足够的负样本进行训练，目前大部分模型选择破坏知识图谱中已有的三元组，通过均匀抽样随机替换头(尾)实体获得负样本。然而，这样获取的负样本的质量得不到保证。例如，当模型初始化嵌入向量时，嵌入空间中各三元组的嵌入向量还没有彼此分离。在这样的情况下使用随机生成的负样本进行训练可以帮助模型区分不同三元组。然而随着训练的进行，嵌入空间中各个三元组已经彼此分离，而模型可以很轻易地识别这样的负样本，那么这些负样本对模型的训练已经没有任何帮助。因此本文需要获取高质量的负样本，以保证即使模型训练到一定程度，本文生成的负样本仍然可以对模型训练起到很好的指导作用。

受文献[21]的启发，本文可以通过生成式对抗网络来改善现有的负样本生成策略。在生成式对抗模型中需要生成器和判别器，生成器负责生成样本逼真的新数据实例，判别器负责判断给定的实例是真实的还是由生成器伪造的。本文通过对每个正三元组 $tp_i^p=(h_i^p,r_i^p,t_i^p)\in TP$ 添加一个随机扰动 z_{im}，再通过一个两层的全连接层来生成新的三元组 tp_{ij}^*。具体地，本文的生成器模型为：

$$G(z;\theta)=\mathrm{Tanh}(\boldsymbol{W}_2\cdot\mathrm{ReLU}(\boldsymbol{W}_1\cdot z+b_1)+b_2) \tag{1}$$

其中，$z\in\mathcal{N}(0,\sigma I)$ 为随机扰动；$\mathbf{I}\in R^{d\times d}$ 为单位矩阵，维度与嵌入向量维度 d 相同；σ 是输入噪声的偏差；θ 是两层全连接层的可学习参数 $\boldsymbol{W}_1$，$\boldsymbol{W}_2$，b_1，b_2 的集合。判别器模型则基于对数损失的DistMult。

$$D(tp;\theta)=f_r(h,t)=\boldsymbol{h}\times\boldsymbol{r}\times\boldsymbol{t} \tag{2}$$

因为假阴性问题的存在，我们没有直接将生成的三元组认定为负三元组，而是通过正类无标签学习来判别。本文通过生成式对抗网络生成了新的三元组。因为存在假阴性标签问题，所以本文区别于大多数直接将新生成的三元组作为负三元组用于模型的训练的方法[20-21]，将新生成的三元组视为无标签三元组。

3.2 正类无标签学习模块

知识图谱中仅包含真实的事实三元组，不存在显性的负样本。大多数现有模型采用均匀采样来获取负样本。然而，它们忽略了采样出来的样本可能是未发现的事实，即假阴性标签问题。因此，对于知识图谱而言，所有未见的三元组可能是正样本或负样本。故它们的真实标签应该是未知的。正无标签学习正是针对仅有正标签和未知标签的场景设计的。受到正类无标签学习的启发，本文利用训练正类无标签学习风险估计器来获得三元组的最优排序以完成知识图谱补全任务。

正样本的先验概率(即在所有可能样本中正样本所占的比例)为 $\pi_p=1-\pi_n$，其中 π_n 为负样本的先验概率(即所有样本中负样本所占比例)。为了优化排序所需的风险估计器，需要计算正样本先验概率的经验风险估计量。经验风险估计量表示模型关于训练集的平均损失，旨在通过最小化训练数据上的损失函数(风险函数)来寻找最优模型。其核心思想是根据观察到的有限样本数据来估计模型在整个数据分布上的期望风险，从而选择具有最低经验风险的模型作为最优解。经验风险估计量计算式为：

$$P_{p,n}(S)=\pi_pP_p^+(S)+\pi_nP_n^-(S) \tag{3}$$

其中，S 为决策函数；$P_p^+(S)$ 表示预测正样本为正的期望风险，类似地，$P_n^-(S)$ 表示预测负样本为负的期望风险。由于在正无标签学习中没有负样本，所以需要通过计算得到 $\pi_nP_n^-(S)=P_u^-(S)-\pi_pP_p^-(S)$，其中 $P_u^-(S)$ 为无标签样本预测为负的期望风险，$P_p^-(S)$ 为正样本被预测为负的期望风险。因此，正样本的先验概率的经验风险估计量可以表示为：

$$P_{p,n}(S)=\pi_pP_p^+(S)+P_u^-(S)-\pi_pP_p^-(S) \tag{4}$$

为了缓解决策函数过于复杂而导致的过拟合现象，本文使用一种非负风险估计量[23]。

$$P_{p,n}(S)=\pi_pP_p^+(S)+\max\{0,P_u^-(S)-\pi_pP_p^-(S)\} \tag{5}$$

本文选择 log-sigmoid 函数作为正无标签学习的损失函数[23]，以便可以通过现成的梯度方法使其最小化。正(负)样本被预测为正(负)的期望风险可以扩展为：

$$P_p^{\pm}(S)=-\frac{1}{|TP|}\sum_{i=1}^{|TP|}\ln\sigma(\pm S(tp_i^p)) \tag{6}$$

其中，$P_p^+(S)$ 表示正样本被预测为正类的期望风险概率，

$P_p^-(S)$表示负样本被预测为负类的期望风险概率，TP 是三元组集合，$|TP|$表示三元组集合总数。同理，无标签样本被预测为负样本的期望风险可以扩展为：

$$P_u^-(S)=-\frac{1}{|TP|}\frac{1}{N}\sum_{i=1}^{|TP|}\sum_{j=1}^{N}\ln\sigma(-S(tp_{i,j}^u)) \tag{7}$$

为了区分正负样本，需要使决策函数对正样本的评分高于未标记三元组。为了保证正三元组 tp_i^p 的排名始终高于无标签三元组 $tp_{i,j}^u$，进一步将无标签三元组被预测为负类的期望风险扩展为：

$$P_u^-(S)=\frac{1}{|TP|}\sum_{i=1}^{|TP|}-\frac{1}{N}\sum_{j=1}^{N}\ln\sigma(S(tp_i^p)-S(tp_{i,j}^u)) \tag{8}$$

其中，$-\frac{1}{N}\sum_{j=1}^{N}\ln\sigma(S(tp_i^p)-S(tp_{i,j}^u))$是获得非最优排序的风险。通过在原本的正类无标签学习的风险估计量中嵌入成对排序目标 $S(tp_i^p)-S(tp_{i,j}^u)$，将 $S(tp_i)$与 $S(tp_{i,j})$之间的差异最大化，以保证正三元组排名高于无标签三元组。

3.3 最小化最大化博弈

在知识图谱中，正样本仅占所有真实三元组的一部分。为了获取足够的候选样本，本文采用生成式对抗网络来生成新样本。为确保获得可靠的正样本和负样本，本文并未将新生成的样本直接视为正样本或负样本，而是将其视为无标签样本，并通过正类无标签学习方法对这些样本进行排序。生成式对抗网络的最大化最小化博弈可以表示为：

$$\min_D\max_G:-\frac{1}{|TP|}\frac{1}{|N|}\sum_{i=1}^{|TP|}\sum_{j=1}^{|N|}\ln\sigma(D(tp_i)-D(tp_{ij}^*)) \tag{9}$$

随后，将新生成的三元组作为无标签三元组，通过正类无标签学习对其成对进行排序。通过固定生成器 G 最小化公式，利用训练判别器 D 优化正三元组和新生成的三元组之间的排序，确保正三元组排序始终高于无标签三元组；再利用固定判别器 D 最大化公式，利用训练生成器 G 生成可以欺骗判别器 D 的三元组。

通过式(8)可以将最小化最大化博弈改写为：

$$P_*^-(S)=\min_D\max_G:p_{*,i}^-(S) \tag{10}$$

其中，$P_*^-(S)$表示获得非最优排序的总风险，$p_{*,i}^-(S)$表示正类三元组 tp_i和新生成的三元组 tp_i^* 获得非最优排序的风险。本文将对抗训练的目标通过上述过程设计为$P_*^-(S)$的最小化最大化博弈，将生成式对抗网络的训练目标与正类无标签学习的风险估计量的统一。

$$\min_D\max_G:\pi_pP_p^+(S)+\max\{0,P_u^-(S)+P_*^-(S)-\pi_pP_p^-(S)\} \tag{11}$$

这样的设计策略有助于更好地整合这两种方法，从而提高模型的性能。

在前文提及的基于生成式对抗网络的知识图谱补全算法中，其目标主要是训练生成器，利用训练好的生成器产生高质量的负样本，从而用于模型训练。与这些方法不同，本文的目标在于训练一个基于正类无标签学习的风险估计器。通过训练好的风险估计器对候选三元组进行排序预测，完成知识图谱补全任务。

4 实验

本文在广泛使用的标准数据集上对所提出的模型进行了评估，并验证了模型的性能和泛化性。然后进行了消融实验来验证各个模块对整体模型的贡献。

4.1 数据集

本文在两个标准数据集(FB15k-237[30]和 WN18RR[5])上对本文提出的模型进行了实验。FB15k-237 是 FB15k 的一个子集，包含真实世界的实体和关系。WN18RR 是 WordNET 的一个子集，由英语短语及其语义关系组成。为了防止数据泄漏，FB15k-237 和 WN18RR 在原有数据集上去除了反关系。细节如表 1 所列。

表 1 数据集统计信息

Table 1 Datasets statistics

数据集	实体	关系	训练集	验证集	测试集
WN18RR	40943	11	86835	3034	3134
FB15k-237	14541	237	272115	17535	20466

4.2 评价指标

在链接预测任务中，给定三元组$(h_i,r_i,t_i)\in TP$，破坏其中头实体(?,r,t)或尾实体(h,r,?)作为负例。模型训练的目标是对正确的三元组进行排序，评估其前 k 项预测的准确性。本文使用链接预测中常用的两个指标来评估本文的模型。

(1)平均排名(Mean Rank，MR)：得到的所有正确三元组的实体排名的平均值。

$$MR=\frac{1}{|TP|}\sum_{i=1}^{|TP|}rank_i \tag{12}$$

其中，$rank_i$表示第 i 个三元组的预测排序位置。该值越小表示正确三元组的平均预测排序位置越靠前，模型效果越好。

(2)平均倒数排名(Mean Reciprocal Ranking，MRR)：所有正确三元组的实体排名的倒数求平均值。

$$MRR=\frac{1}{|TP|}\sum_{i=1}^{|TP|}\frac{1}{rank_i} \tag{13}$$

MRR 主要用于衡量正三元组的最高排名，具有平滑性且受异常值影响更小。MRR 取值范围为 $MRR\in(0,1)$，其值越大表示模型性能越好。

(3)Hits@k：排名在前 k 位的正确实体所占比例的平均值。

$$Hits@k=\frac{1}{|TP|}\sum_{i=1}^{|TP|}\mathbb{I}(rank_i\leqslant k) \tag{14}$$

其中，k 的值一般取 1,3,10；$\mathbb{I}(\cdot)$是指示函数，如果条件为真($rank_i\leqslant k$)则值为 1，否则为 0。

4.3 链接预测

链接预测任务指预测给定三元组(?,r,t)缺失的头实体或者三元组(h,r,?)缺失的尾实体。实验结果使用 4.2 节提及的指标进行评估。本文将所提模型同知识图谱领域中最具代表性的基准模型在基准数据集上进行了比较，实验结果如表 2 所列。其中，正类先验概率π_p在$\{10^{-1},10^{-3},10^{-4},10^{-5},10^{-6}\}$之间进行搜索，生成器 G 和判别器 D 的初始学习率lr_G和lr_D分别在$\{10^{-2},10^{-3},10^{-4},10^{-5}\}$中搜索得到，嵌入维度 d 为{512,1024}。此外，头尾实体被替换的比例被设置为 1∶1。本文在数据集 WN18RR 和 FB15k-237 上的最优参数$\{\pi_p,lr_G,lr_D,d\}$分别为$\{10^{-4},10^{-5},10^{-5},1024\}$和$\{10^{-5},10^{-5},10^{-5},1024\}$。从实验结果中可以看出，基线模型虽然设计了复杂的评分函数，但是它们没有考虑到假阴性标签问题，导致模型性能受到了限制。虽然本文方法没有改进评分函数，但是通过引入正类无标签学习，缓解了假阴性标签问题。实验

结果表明，本文方法在各项指标上都取得了优势，证明本文方法能够有效缓解假阴性标签问题。

表 2 链接预测实验结果

Table 2 Link prediction experimental results

methods	WN18RR					FB15k-237				
	Hits@1	*Hits*@3	*Hits*@10	*MR*	*MRR*	*Hits*@1	*Hits*@3	*Hits*@10	*MR*	*MRR*
TransE	0.279	—	0.495	3936	0.206	0.217	—	0.497	209	0.310
DistMult	0.397	—	0.502	5913	0.433	0.224	—	0.490	199	0.313
ComplEx	0.409	0.469	0.530	7882	0.449	0.194	0.297	0.450	546	0.278
RotatE	0.428	0.492	0.571	3340	0.476	0.241	0.375	0.533	177	0.388
ConvE	0.419	0.470	0.531	4464	0.546	0.225	0.341	0.497	245	0.312
ATTH	**0.443**	0.482	0.573	—	0.486	0.252	0.384	0.540	—	0.348
KBGAN	—	—	0.469	—	0.215	—	—	0.458	—	0.277
NSCaChing	—	—	0.518	—	0.443	—	—	0.481	—	0.302
Our Method	0.435	**0.498**	**0.594**	**3273**	**0.492**	**0.319**	**0.453**	**0.607**	**156**	**0.415**

本文还将所提方法与类似改进了负样本抽样的知识图谱补全算法方法进行了比较。同样地，此前的方法中没有考虑假阴性标签问题，研究人员希望通过生成式对抗网络来提高负样本的质量，并直接使用生成式对抗网络生成的样本作为负样本训练模型。本文模型区别于上述模型，并非直接利用生成式对抗网络生成负样本，而是将其生成的样本视为无标签样本并通过正类无标签学习对三元组进行排序。实验结果表明，本文方法优于利用生成式对抗网络直接生成负样本的方法，在各项指标上取得了明显的优势。

4.4 与基准模型比较

本文模型没有对评分函数进行改进，而是选择直接使用基准模型的评分函数。本文使用了几何距离模型中的 TransE 和线性\双线性模型 DistMult 的两个版本，并在 FB15k-237 数据集上进行了实验，实验结果如表 3 所列。结果表明，本文模型能够有效提升基准模型的性能。

表 3 基于不同评分函数的模型

Table 3 Models based on different scoring functions

methods	*Hits*@1	*Hits*@3	*Hits*@10	*MRR*
TransE	0.217	—	0.497	0.310
Our Method+TransE	0.302	0.437	0.579	0.396
DistMult	0.224	—	0.490	0.313
Our Method+DistMult	0.319	0.453	0.607	0.415

4.5 消融实验

本文设计了消融实验以探究本文方法中各个模块对整体模型的贡献。如表 4 所列，本文设计了两种模块：(1)不包含正类无标签学习模块，直接使用生成式对抗网络生成的三元组作为负样本；(2)仅使用正类无标签学习模块，不包含生成式对抗网络生成的无标签三元组，而是采用均匀抽样获得无标签样本。最后，本文在 FB15k-237 数据集上进行了对比实验。实验结果表明，与不包含正类无标签学习模块相比，将抽样三元组视为无标签三元组并用正类无标签学习模块的方法在所有指标上都具有优势，这证明了本文的正类无标签学习模块可以有效缓解假阴性标签问题；与仅使用正类无标签学习模块的结果相比，加入生成式对抗网络生成无标签样本的方法也对模型有着明显的提升作用，这是因为生成的无标签样本在一定程度上缓解了知识图谱只包含少部分正样本而导致的数据稀疏性问题。通过消融实验，证明了本文提出的模块对整体模型的贡献，表明本文模型可以有效缓解假阴性标签问题并在一定程度上缓解了数据稀疏性问题。

表 4 消融实验

Table 4 Ablation experiment

方法	*Hits*@1	*Hits*@10	*MRR*
无正类无标签模块	0.276	0.490	0.374
无生成式对抗网络模块	0.291	0.547	0.342
Our Method	0.319	0.607	0.415

结束语 本文提出了基于生成式对抗网络和正类无标签学习的知识图谱补全算法，该算法利用最小化最大化博弈将生成式对抗网络和正类无标签学习的训练目标统一。首先利用生成式对抗网络生成大量无标签三元组，然后利用正类无标签学习处理生成的无标签三元组来缓解假阴性标签问题，并在一定程度上缓解了知识图谱的数据稀疏问题。本文进行了详尽的实验来探究本文模型的性能，并设计了消融实验来探究本文设计的每个模块对整体模型的贡献。在多个真实数据集上的实验结果表明，本文提出的方法具有较高的准确性和有效性。本文方法只能在有限程度上缓解假阴性标签问题。在未来，我们将会继续改进本文模型以进一步缓解假阴性标签问题。

参 考 文 献

[1] KROMPASS D, BAIER S, TRESP V. Type-Constrained Representation Learning in Knowledge Graphs[C]// The Semantic Web (ISWC 2015). Cham: Springer International Publishing, 2015: 640-655.

[2] NICKEL M, TRESP V, KRIEGEL H P. A three-way model for collective learning on multi-relational data[C]// Proceedings of the 28th International Conference on International Conference on Machine Learning. Madison, WI, USA: Omnipress, 2011: 809-816.

[3] BORDES A, USUNIER N, GARCIA-DURÁN A, et al. Translating embeddings for modeling multi-relational data[C]// Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2. Red Hook, NY, USA: Curran Associates Inc., 2013: 2787-2795.

[4] TROUILLON T, WELBL J, RIEDEL S, et al. Complex Embeddings for Simple Link Prediction[C]// Proceedings of The 33rd International Conference on Machine Learning. PMLR, 2016: 2071-2080.

[5] DETTMERS T, MINERVINI P, STENETORP P, et al. Convolutional 2D Knowledge Graph Embeddings[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[6] SUN Z，DENG Z H，NIE J Y，et al. RotatE：Knowledge Graph Embedding by Relational Rotation in Complex Space[C]// 7th International Conference on Learning Representations (ICLR 2019). New Orleans，LA，USA：OpenReview. net，2019.

[7] CHAMI I，WOLF A，JUAN D C，et al. Low-Dimensional Hyperbolic Knowledge Graph Embeddings[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics，2020：6901-6914.

[8] YAO L，MAO C，LUO Y. KG-BERT：BERT for knowledge graph completion[J]. arXiv：1909. 03193，2019.

[9] NIU G，ZHANG Y，LI B，et al. Rule-Guided Compositional Representation Learning on Knowledge Graphs[J]. Proceedings of the AAAI Conference on Artificial Intelligence，2020，34(3)：2950-2958.

[10] SHA X，SUN Z，ZHANG J. Hierarchical attentive knowledge graph embedding for personalized recommendation[J]. Electronic Commerce Research and Applications，2021，48：101071.

[11] ZHANG Y，YAO Q，SHAO Y，et al. NSCaching：Simple and Efficient Negative Sampling for Knowledge Graph Embedding [C]// 2019 IEEE 35th International Conference on Data Engineering(ICDE). 2019：614-625.

[12] YANG B，YIH W，HE X，et al. Embedding entities and relations for learning and inference in knowledge bases[J]. arXiv：1412. 6575，2014.

[13] CHAMI I，WOLF A，JUAN D C，et al. Low-Dimensional Hyperbolic Knowledge Graph Embeddings[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics，2020：6901-6914.

[14] DONG X，GABRILOVICH E，HEITZ G，et al. Knowledge vault：a web-scale approach to probabilistic knowledge fusion [C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge discovery and Data Mining. New York，NY，USA：Association for Computing Machinery，2014：601-610.

[15] GUO L，SUN Z，HU W. Learning to Exploit Long-term Relational Dependencies in Knowledge Graphs[C]// Proceedings of the 36th International Conference on Machine Learning. PMLR，2019：2505-2514.

[16] SCHLICHTKRULL M，KIPF T N，BLOEM P，et al. Modeling Relational Data with Graph Convolutional Networks[J]. arXiv：1703. 06103，2018.

[17] VASHISHTH S，SANYAL S，NITIN V，et al. Composition-based Multi-Relational Graph Convolutional Networks [C]// Eighth International Conference on Learning Representations. 2020.

[18] CRESWELL A，WHITE T，DUMOULIN V，et al. Generative Adversarial Networks：An Overview[J]. IEEE Signal Processing Magazine，2017，35(1)：53-65.

[19] YU L，ZHANG W，WANG J，et al. SeqGAN：Sequence Generative Adversarial Nets with Policy Gradient[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2017.

[20] CAI L，WANG W Y. KBGAN：Adversarial Learning for Knowledge Graph Embeddings[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，Volume 1. New Orleans，Louisiana：Association for Computational Linguistics，2018：1470-1480.

[21] WANG P，LI S，PAN R. Incorporating GAN for Negative Sampling in Knowledge Representation Learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[22] BEKKER J，DAVIS J. Learning from positive and unlabeled data：a survey[J]. Machine Learning，2020，109(4)：719-760.

[23] KIRYO R，NIU G，DU PLESSIS M C，et al. Positive-Unlabeled Learning with Non-Negative Risk Estimator[C]// Advances in Neural Information Processing Systems：Vol. 30. Curran Associates，Inc. ，2017.

[24] XU D，DENIL M. Positive-Unlabeled Reward Learning [C]// Proceedings of the 2020 Conference on Robot Learning. PMLR，2021：205-219.

[25] LUO C，ZHAO P，CHEN C，et al. PULNS：Positive-Unlabeled Learning with Effective Negative Sample Selector[J]. Proceedings of the AAAI Conference on Artificial Intelligence，2021，35(10)：8784-8792.

[26] DU PLESSIS M C，NIU G，SUGIYAMA M. Analysis of Learning from Positive and Unlabeled Data[C]// Advances in Neural Information Processing Systems. Curran Associates，Inc. ，2014.

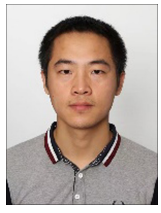
[27] PLESSIS M D，NIU G，SUGIYAMA M. Convex Formulation for Learning from Positive and Unlabeled Data[C]// Proceedings of the 32nd International Conference on Machine Learning. PMLR，2015：1386-1394.

[28] HSIEH Y G，NIU G，SUGIYAMA M. Classification from Positive，Unlabeled and Biased Negative Data[C]// Proceedings of the 36th International Conference on Machine Learning. PMLR，2019：2820-2829.

[29] RAN Z J，SUN L F，ZOU Y S，et al. Few-Shot Knowledge Graph Completion Model Based on Relation Learning Network [J]. Computer Engineering，2023，49(9)：52-59.

[30] TOUTANOVA K，CHEN D，PANTEL P，et al. Representing Text for Joint Embedding of Text and Knowledge Bases[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon，Portugal：Association for Computational Linguistics，2015：1499-1509.

HU Binhao，born in 1996，postgraduate. His main research interests include graph representation，knowledge graph and natural language processing.



ZHANG Jianpeng，born in 1988，Ph.D，assistant researcher. His main research interest is big data analysis.

（责任编辑：何杨）