混合负采样的知识图谱嵌入

奚超亮,冷泳林
(渤海大学 信息科学与技术学院,辽宁 锦州 121000)

摘 要:知识图谱嵌入表示模型将实体与关系转化为低维的向量表示,来表达实体与关系之间的关联语义,是解决知识图谱补全问题的重要方法。传统模型采用随机负采样来构造负例三元组,容易产生低质量负样本,影响表示模型的特征学习能力。基于相似性的负采样方法,对实体点进行聚类,提高了负采样的质量。但针对知识图谱中的稀疏点,因无法控制聚类点数量,导致模型性能降低。经过对相似性负采样和样本点稀疏问题的研究,采用基于密度的聚类算法 DBSCAN(Density-Based Spatial Clustering of Applications with Noise)对聚类中的样本进行头尾实体的替换,并对 DBSCAN 中的领域聚类半径采取了自适应优化,找到合适的聚类中心,降低离群点的数量。同时对于聚类外的离群点进行过采样,构造离群点的相似点,解决稀疏点负采样的问题。最后,将该负采样方法与 TransE 结合,得到了混合负采样模型 TransE-DNS。研究结果表明:TransE-DNS 在链路预测和三元组分类任务上取得了更好的效果。

关键词:翻译模型;知识图谱;三元组分类;链路预测;DBSCAN clustering;负采样
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2023)09-0168-07
doi:10.3969/j.issn.1673-629X.2023.09.025

Knowledge Graph Embedding with Mixed Negative Sampling

XI Chao-liang,LENG Yong-lin
(School of Information Science and Technology,Bohai University,Jinzhou 121000,China)

Abstract:The embedding model of knowledge graph transforms entities and relationships into low dimensional vector representation to express the association semantics between entities and relationships,which is an important method to solve the problem of knowledge graph completion. The traditional embedding model adopts random sampling to construct negative triples,which is easy to produce low-quality negative samples,affecting the feature learning ability of representation models. The clustering-based negative samplings cluster entity points to improve the quality of negative sampling. However,for the sparse points of the knowledge graph,the clustering cannot control the number of clustering points,which leads to the degradation of the model performance. After researching on negative similarity sampling and sparse sample points,we adopt DBSCAN to replace the head and tail entities of the samples in the cluster and adaptively optimize the domain clustering radius in DBSCAN to find a suitable cluster center and reduce the number of outliers. At the same time,oversampling is conducted for outliers to build similarity points,which is used to solve the sparse point problem. Finally,the negative sampling method is combined with TransE to obtain the mixed negative sampling model Trans-DNS. The results show that TransE-DNS has achieved better results in link prediction and triple classification tasks.

Key words:translation model;knowledge graph;triple classification;link prediction;density-based spatial clustering of applications with noise clustering;negative sampling

0 引 言

知识图谱采用有向图的方式描述客观世界概念、实体及其关系。知识图谱技术作为人工智能三大主要技术之一,在 2012 年由谷歌公司提出,但其知识的描述和表示方法可以追溯到 1960 年的语义网,经过一系列演变,形成今天的知识图谱。目前,一些有代表性的知识图谱如 Freebase[1]、OpenKN[2]、Wordnet[3]、Probase[4]等从大量数据资源中抽取、组织和管理知识,为个性化推荐[5]、智能搜索与回答、内容分发提供强有力的知识支撑,推动人工智能各应用领域的快速发展。

随着知识图谱规模的不断扩大,知识图谱中的不完整数据也在增加,导致知识图谱质量不断下降。知识图谱的自动补全技术是提高知识图谱质量的一种有

收稿日期:2022-09-23 **修回日期**:2023-02-03
基金项目:辽宁省教育科学研究项目(LJ2020016);渤海大学国家安全研究院项目(XK202134-39)
作者简介:奚超亮(1996-),男,硕士研究生,CCF 会员(L4853G),研究方向为知识图谱嵌入、不平衡学习、深度学习;通信作者:冷泳林(1978-),女,博士,副教授,研究方向为知识图谱的存储与索引技术、不完整数据填充。

效手段。近年来,知识图谱利用自然语言的表示学习技术将实体关系映射到低维稠密向量空间,实现了知识的表示学习[6],推动了知识图谱补全及链路预测技术的发展。其中,以 TransE[7] 为主的翻译模型最为普遍,TransE 模型通过设置全局带参得分函数来定义实体与关系之间的嵌入表示,并且基于边界的训练目标来惩罚负样本,最终将知识库中的实体和关系映射到低维向量空间。此类知识图谱嵌入表示模型存在一个共性问题,原因是基于随机抽取的负采样方式不能很好地生成高质量的负样本,进而影响了训练模型的质量。Trans_SnS 模型[8] 提出了基于实体相似性的负采样方法,但该方法中使用的 K-means 聚类并不能有效确定聚类中心点的数量,并且没有很好地处理稀疏点,将稀疏数据划分到了错误的聚类中,导致稀疏数据负采样的质量较低。

通过对实体相似性负采样的进一步研究,该文采用基于密度的聚类算法 DBSCAN[9] 对相似性负采样方法进行优化,同时对 DBSCAN 中的 eps 聚类半径采取了自适应优化,并且结合 SMOTE[10] 思想对聚类中的离群样本点进行过采样,拟合相似实体点,来提高负样本的质量,最后将上述负采样方法同 TransE 模型结合得到 TransE_DNS 模型。实验选取了以下公开数据集(WN11,WN18,FB15K,FB13),分别在链路预测和三元组分类任务中对该模型进行评估,实验效果均获得了提升。

1　相关研究

1.1　翻译模型

知识图谱的嵌入表示方法中以翻译表示模型最为先进。2013 年 Bordes 等人[11] 提出了 TransE 模型,该模型把三元组关系看作头实体到尾实体的一种翻译操作,即三元组的头、尾和关系向量应满足 $h + r \approx t$。TransE 模型因其参数少、时间复杂度低等优点实现了在大规模稀疏知识图谱上较好的预测,也成为后续 Trans 系列模型的基础。2014 年 Wang 等人[12] 提出了 TransH 模型,通过把实体映射到关系所在的超平面上,实现了同一实体在不同关系上的不同向量表示,更好地解决了 TransE 在一对多,多对一和多对多关系上的嵌入表示问题。随后,Lin 等人[13] 提出的 TransR 模型认为关系和实体之间存在差异性,采用不同的向量空间分别映射关系和实体,然后通过一个投影矩阵完成从关系空间向实体空间的映射。虽然 TransR 对复杂关系建模效果很好,但其复杂度较高,很难应用于大规模知识图谱的表示。Ji 等人[14] 提出的 TransD 使用两个向量来表示实体,其中一个向量表示其含义,另一个向量用于构造实体到关系向量空间的映射矩阵,由于其动态地构建投影矩阵,相对于 TransR 大大减少了参数量和计算量。Ji 等人[15] 提出的 TranSparse 模型主要解决知识图谱中普遍存在的异构性和不平衡性问题。该模型提出了一种自适应稀疏矩阵实现对不同关系的投影,有效解决了大规模知识图谱的嵌入表示问题。Xiao 等人[16] 提出了 TransA 模型在 TransE 模型基础上更换度量函数,为实体和关系的每个维度添加权重来提升模型的表示能力。Hong 等人[17] 提出了一种结合实体领域信息的模型 CombiNe,该模型通过统计方法 TF-IDF,从实体的领域当中抽取重要实体邻居,通过短接联合表示的方式,提高了基于扩展信息的知识表示模型的性能。

以上模型主要是针对损失函数进行改进,但都忽略了负采样方式对模型效果的影响,该文主要通过改良负采样的方式,提高负采样的质量,来提升模型的效果。

1.2　负采样方法

Trans 系列的翻译模型在训练过程中通常采用均匀采样和伯努利采样。其中,均匀采样方法是通过均匀的随机替换样本头尾节点来生成负样本。由于数据集中的样本存在一对多和多对一的关系,导致均匀采样方法更容易生成假样本。伯努利采样针对三元组一对多的关系,使用更大概率替换头节点;反之,以更大概率替换尾节点。这种方法大大降低了生成假样本的可能性,弥补了均匀采样的缺点。以上两种采样方式在替换头尾节点时都采用随机替换的方式,优势在于降低了训练的时间复杂度。但是随着训练的进行,生成负样本质量过低,导致得分函数在此类低质量样本上得分较低,从而导致训练过程中梯度清零。

为提高负采样质量,近年来出现了以生成对抗网络、聚类模型为基础的负采样方法[18]。生成对抗网络的负采样方式以 KBGAN[19] 为首,选择基于平移距离的 KRL 模型作为负样本生成器和基于语义匹配的 KRL 模型作为对抗训练的鉴别器,生成器在一个候选负集合上产生一个概率分布,并选择概率最高的一个输入鉴别器。该鉴别器使正、负样本之间的边际损失最小化,提高了负采样的质量,学习最终的嵌入向量。

由于知识图谱嵌入中的负采样属于离散域的输出,KBGAN 并不能直接使用梯度下降策略,而是采用了强化学习策略进行训练,使生成器产生离散化负例,这种方式容易使训练模型不稳定。同时生成对抗模型的采样过程时间复杂度较大,不利于训练大规模的知识图谱。

2018 年 Wang 提出了 IGAN[20],将错误的正三元组输入神经网络,添加 Softmax 计算整个实体集的概率分布,通过鉴别器来得到较高质量的负三元组。

聚类采样 Trans_SNS 基于实体相似性负采样方法来提高负样本的质量,该模型使用 K-means 对实体进行聚类,利用聚类内部实体具有高度相似性,生成高质量负样本,进而提高 TransE 模型的性能,但该模型无法确定聚类中心点的数量。除此之外,当面对大规模稀疏知识图谱时,固定数量的聚类使一些离群点生成低质量的负样本,从而影响模型的效果。

2 知识图谱嵌入的负采样优化

2.1 基于实体向量的相似性分析

TransE 模型将知识图谱中的实体与关系嵌入到同一个向量空间中,其中每个三元组的头尾实体和关系之间满足 $h + r \approx t$ 的约束。如图 1 中 Stephen Curry 和 Seth Curry 同时属于 NBA 里的现役球员,那么在向量空间中,将 Stephen Curry 和 Seth Curry 作为头实体 h,尾实体 t 是 NBA,那么它们的向量表示趋近于相等。但由于 Stephen Curry 和 Seth Curry 分别代言 Nike 和 Armour,这又让他们的向量表示存在一定的区别。当实体间拥有更多相同约束时,它们的向量表示就越相似。如 Seth Curry 与 Durant 都是 NBA 球员,且同时效力 Brooklyn 俱乐部,因此 Seth Curry 与 Durant 这两个实体的向量表示更相似。反之,实体间约束越少,那么他们在向量空间中的距离越远,相似性越低。

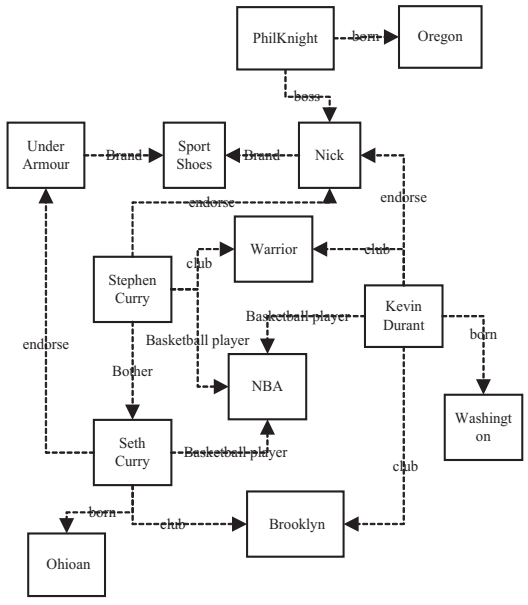

图 1 知识图谱局部关系

此外,当两个实体间没有直接约束关系时,如实体 Ohioan 和 Washington,他们分别是 Seth Curry 与 Durant 两个球员的出生地,没有直接联系。根据上文,Seth Curry 与 Durant 拥有较多的相同约束,那么他们相似性较高。同时 Ohioan 和 Washington 分别作为两个球员的出生地,与 Brooklyn 队和 NBA 都具有相同的间接联系。PhilKnight 出生于 Oregon 并且 PhilKnight 作为 NIKE 的老板,既不在 NBA 也不在 Brooklyn 队。因此,笔者认为相较于 Oregon,Washington 与 Ohioan 因具有较多的间接约束在向量空间中具有一定的相似性。

综上所述,在向量空间中,拥有较多相同约束条件的实体,他们之间的距离一定是较近的。拥有较多相同间接约束条件的实体,存在一些与他们有较多共同约束条件的实体点,而这些实体点相互之间又拥有较多的共同约束。因此,这些实体之间的联系也是较为紧密的,反之则较远。

综上,在实体向量空间中,相互之间直接距离较近的实体点,他们的语义一定相似;间接联系较多的实体点之间存在着隐关系,同样也具有较高的相似性。

2.2 DNS 负采样方法

负例样本的质量影响知识图谱的嵌入表示,高质量的负例样本应与被替换实体具有较高的相似性。为了得到高质量负例样本,提出了基于密度聚类的负采样算法(DBSCAN Negative Sample,DNS)。DNS 选择不受聚类中心点数量限制的 DBSCAN 算法将知识图谱中的实体按照紧密程度进行聚类,以此限定负例样本的选取范围。DBSCAN 通过邻域半径 eps 所给定的区域来划分实体聚类:

$$N_{\text{eps}}(x) = \{y \in X : \text{dist}(x,y) \leqslant \text{eps}\} \tag{1}$$

式中,y 表示实体点,X 表示实体集,$\text{dist}(x,y) \leqslant \text{eps}$ 判断 x,y 之间的距离是否小于 eps,这里选用欧氏距离来判定。

DBSCAN 的聚类数量由聚类的邻域半径 eps 和聚类内最小包含点数 minpts 决定。eps 越小,聚类精度越高,聚类内部产生的负样本质量也越高。但过高的聚类精度也会产生更多的离群点,从而降低聚类中样本的数量,引起模型过拟合。因此,DNS 算法通过自适应方式寻找合适的 eps,并设置了离群点数量的阈值不超过总训练集样本的四分之一。

同时,面对大型知识图谱的稀疏性问题,DNS 将向量空间中远离聚类的正样本,通过过采样的方式,生成相似度较高的负样本。并且针对过采样产生的假负样本影响模型修正的问题,通过随机选取部分真实样本,然后在真实样本中选择与假负样本相似的样本进行替换,来保证负样本采样的质量。算法 1 给出了 DNS 负采样方法的算法描述。

算法 1:DNS

输入:训练集 $S_{(h,r,t)}$,聚类最小包含点数 minpts,过采样样本数量 overCount,阈值 T,训练次数 epoch

输出:负采样训练集 $S'_{(h,r,t)}$

1:初始化:$S' \leftarrow [\,]$

2:if (epoch/T)%2! =1 then

3:eps←0.1//初始化聚类半径
4:overCount ←0//离群点数量
5: prev←0//前一次聚类数量
6:pres←DBSCAN(eps ,minpts)//当前聚类数量
7:while pres>prev and overCount <len(S)/4
8:do prev ← pres
9: classific←DBSCAN(eps,minpts) //DBSCAN 聚类
10:pres←get_class_num(classific)　　//读取聚类数量
11:outCount←get_outCount_num(classific) //读取离群点数量
12:eps←update(eps,pres)//更新聚类邻域半径
13:end
14:for each (h,r,t) in $S_{(h,r,t)}$
15:if　classific[(h,r,t)]! = -1 then//如果样本在聚类中
16:cluster←Sample(h,r,t)//取出同聚类的样本点集合
17:neg←Instead(cluster)//替换头或尾节点,构造负样本
18:else//如果样本为离群点
19:overSample ←CircleSmote(overcount,eps) //生成过采样样本集合
20:NegativeSample←NearSample(overSample) //选择与过采样样本较相似的真实样本
21:neg←Instead(NegativeSample)//替换头或尾节点,构造负样本
22:　end if
23: end for

算法 1 的第 2 行根据阈值 T 和 epoch 决定了重新聚类的迭代次数。第 3 至 13 行是寻找样本参数 eps 和生成聚类的过程。算法首先设定一个较低的 eps,然后根据 eps 和输入的参数 minpts 进行一次聚类,从而得到当前聚类数量 pres。当 pres 大于 prev 时,根据当前 eps 和 minpts 进行聚类,并得到当前聚类数量 pres 和离群点数量 outCount。最后,对 eps 进行更新。初始每次迭代,eps 累加 0.1,当 pres 大于 3 时,累加改为 0.03。第 16 至 18 行表示当需要负采样的样本点的 h 或 t 位于向量空间的某个聚类中,则在该聚类中随机抽取一个实体向量替换成 h' 或 t'。

传统的以 SMOTE 为主的过采样算法,都是基于 K 近邻随机选取若干样本点,通过少数类样本与近邻样本点的连线,在线上合成少数类样本点。但 SMOTE 算法是通过遍历所有样本点到少数类样本点的距离来选定 K 近邻,这种做法用在大规模的知识图谱上效率过低。DNS 改进了过采样算法来适用于大规模的知识图谱,第 19 至 22 行给出了离群点采样方法。对于不在聚类中的离群点,首先人工合成离群点的同类点($\triangle_1$, $\triangle_2$,…, $\triangle_n$),将离群 Outliers 视为圆心,将多数类样本的领域半径 eps 视为 Outliers 的邻域半径构造多个圆形区域,并在每个区域内进行随机过采样。之后随机抽选过采样样本点,寻找除离群点外,最接近该过采样样本点的真实样本点(∇_1, ∇_2,…, ∇_n)作为负样本,如图 2 所示。

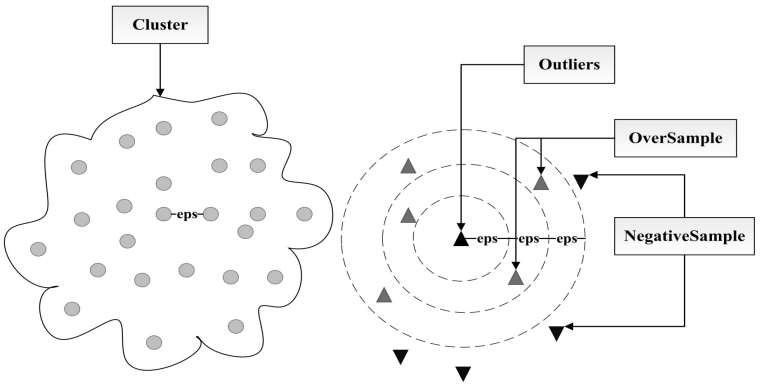

图 2　离群点过采样

为了避免在寻找真实样本时遍历实体集合,算法在每个聚类中随机抽取两个真实样本,生成一个较小的样本集合。将抽选的过采样样本与该集合中的真实样本进行比较,寻找一个最相似的真实样本作为负样本。

2.3　TransE_DNS 模型

该文以知识图谱嵌入模型 TransE 为基础,同时结合 DNS 负采样算法,提出了 TransE_DNS 训练模型。

在模型中,给定知识图谱 $G=(E,R)$,其中 $E=\{e_1,e_2,\cdots,e_n\}$ 表示知识图谱中的实体集合, $R=\{r_1,r_2,\cdots,r_m\}$ 表示知识图谱中的关系集合, n 和 m 分别表示实体与关系的数量。设得分函数为:

$$f_r(h,t)=\|h+r-t\|_{L_1/L_2},h,t\in E,r\in R \quad (2)$$

$f_r(h,t)$ 用来衡量三元组 $h+r$ 与 t 之间的距离,可以用 L_1 或 L_2 范数计算。如果三元组是正确的,则得分函数中 $h+r$ 与 t 得分较低,反之,表示三元组是错误的。因此,定义 TransE_DNS 模型的损失函数为:

$$L=\sum_{(h,r,t)\in s}\sum_{(h',r,t')\in s}\max(f_r(h,t)+\gamma-f_r(h'+r,t'),0) \quad (3)$$

其中, γ 为边界值表示正负样本之间的间距, (h,r,t) 是知识图谱中的真实样本, (h',r,t') 是负样本, h' 和 t'

为替换的头尾实体。当 $f_r(h,t)+\gamma-f_r(h'+r,t')$ 大于 0 时,损失函数 L 取原值,否则取 0,目标是使得最相近的正负例样本距离最大化。该文利用 Adam 适应性矩估计最小化损失函数。

算法 2 描述了 Trans_DNS 模型的完整训练过程。在训练过程中,DNS 负采样每迭代 T 次 epoch 后进行一次聚类。

算法 2:Trans_DNS

输入:训练集 $S_{(h,r,t)}$,实体集 E,关系集 R,边界值 γ,嵌入维度 K,学习率 α,聚类最小包含点数 minpts,过采样样本数量 overCount,阈值 T

输出:实体向量,关系向量

1:初始化参数:

2: $r \leftarrow \text{uniform}(-6/\sqrt{n},6/\sqrt{n})$ // $r \in R$

3: $r \leftarrow r/\|r\|$ // $r \in R$

4: $e \leftarrow \text{uniform}(-6/\sqrt{n},6/\sqrt{n})$ // $e \in E$

5: $e \leftarrow e/\|e\|$ // $e \in E$

6:loop

7: $S_{\text{batch}} \leftarrow \text{sample}(S,b)$ //从 S 中抽取大小为 b 的 mini-batch

8: $T_{\text{batch}} \leftarrow \varnothing$

9: for $(h,r,t) \in S_{\text{batch}}$ do

10:if (epoch/T)%2! = 1 then

11: $S'_{(h,r,t)} = \text{DNS}((h,r,t),\text{minpts},\text{overCount},T,\text{epoch})$ //通过 DNS 负采样

12:end if

13: $(h',r,t') \leftarrow \text{sample}(S'_{(h,r,t)})$ //从负样本集合中抽取负样本

14: $T_{\text{batch}} \leftarrow T_{\text{batch}} \cup \{((h,r,t),(h',r,t'))\}$

15:end for

16: $\sum_{(h,r,t)\in S}\sum_{(h',r,t)\in S'} \max(f_r(h,t)+\gamma-f_r(h'+r,t'),0)$ //更新实体向量与关系向量

17:end loop

第 1 至 6 行使实体集合中的 e 和关系集合中的 r 随机生成高维的实体和关系向量。

第 7 至 15 行表示从训练集 S 中抽取一个大小为 b 的 mini-batch 集合,根据当前的 epoch 和阈值 T 来判断是否通过 DNS 负采样生成负采样集合 S'。

第 16 行表示先将正样本与生成的负样本带入到损失函数中,使用 adam 优化最小化损失函数,更新实体向量与关系向量。

3 实验与分析

使用多个数据集,分别进行了链路预测和三元组分类的实验,从不同角度验证 Trans_DNS 模型的有效性。

3.1 数据集设置

选用知识工程中广泛使用的两个数据集 Freebase 和 WoreNet。Freebase 中包含了非常多的话题和类型知识,对知识图谱工程的相关研究具有重要的导向作用,实验选择 Freebase 数据集的两个子集 FB15K 和 FB13,其中 FB15K 是一个包含大规模常识性知识的知识图谱,该图谱中含有对称关系、非对称关系和反转关系。WordNet 是一个描述词汇之间关联特点的数据集,选择了 WN11 和 WN18 两个子集作为实验数据集,其中子集 WN18 被用于各种知识工程任务中。数据集的详细信息如表 1 所示。

表 1 实验中的数据集

Dataset	WN11	WN18	FB13	FB15K
Entity	38 696	40 943	75 043	14 951
Relation	11	18	13	1 345
Train	112 581	141 442	316 232	483 142
Valid	2 609	5 000	5 908	50 000
Test	10 544	5 000	23 733	59 071

3.2 链路预测

链路预测是一种根据知识图谱中的已存在实体去预测缺失事实的任务,它是一种有前途、广泛研究且旨在完成知识图谱补全的任务。对于确认的三元组 (h,r,t),其主要目的是预测缺失的 h 或 t。

在这个过程中,除了缺失的 h 或 t,其余实体被视为候补实体。利用候补实体替换三元组中的 h 或 t,生成候补三元组,并计算出其与测试三元组的得分。最后,根据候补三元组的得分进行升序排列。本组实验选用了 FB15K 和 WN18 作为数据集,将 MeanRank 和 Hits@10 作为评价指标。MeanRank 表示测试集中三元组匹配到正确结果的平均排序位次,Hits@10 表示根据得分序列,判断测试三元组的正确答案排在序列前 10 位次的占比。实际上,不完整的三元组补全后可能与已经存在的三元组重复,这会影响三元组的排序值。过滤掉这类三元组的操作称为 Filter,未过滤这类三元组则称为 Raw。经过各种模型的测试,Filter 的实验效果通常比 Raw 更好,能得到更好的 MeanRank 和 Hits@10。

在实验中,为了得到模型最佳的参数,对参数的设置进行了多次尝试。主要对以下参数进行设置和选择:训练周期 epoch 的取值范围设在{1 000,1 500,2 000},adam 的学习率 α 在{0.001,0.003,0.005,0.01,0.02}范围内,边界值 γ 在{1,2,2.5,3,3.5,4,4.5,5}范围内,嵌入维度 n 在{50,100,150,200}范围内,批处理 β 在{1 200,3 000,4 800,10 000}范围内,聚类最小包含点数 M 在{5,10,20,25,50}范围内,过采样样本数量 O 在{2,3,4,5,6}范围内,阈值 T 在{20,50,100}范围内。三元组得分计算均采用 L1 范数进行计算。

经过多次实验,WN18 和 FB15K 两组数据集的参数设置如表 2 所示。

表 2　链路预测参数设置

Dataset	epoch	α	γ	n	β	M	O	T
WN18	1 000	0.01	5	50	3 000	15	3	50
FB15K	2 000	0.003	2.5	100	10 000	10	5	50

链路预测结果如表 3 所示,因设备环境与参数等问题,对照实验达不到原文献的模型性能,因此直接采用原文献的实验结果,加粗部分为文中模型与表中模型对比下得到的最优解。从表中可以看出,文中模型在 WN18 数据集的 MeanRank 上得到了最优解,Hits@10 略低于表中最佳结果。从 FB15K 数据集的结果上看,在 MeanRank(unif)下得到了最优解,在 MeanRank(bern)下与最优解接近,Hits@10 与最佳效果仍有一定差距。实验结果表明,文中模型在针对关系复杂的 FB15K 数据集时,虽然能得到不错的平均排序得分,但正确实体排在前 10 的概率并不算高。笔者认为主要是以下两个原因:其一是因为 TransE-DNS 在 DNS 负采样时,虽然针对离群点,巧妙地通过过采样的方式构造相似实体点,寻找除离群点外,与之接近的真实样本点。但在这个过程中,因模型效率问题,不能遍历整个实体空间,只通过在每个聚类中随机抽选去进行比较。这导致可能会选择到一个与过采样样本点不够相似的真实样本点,同样这个点与离群点的相似度也很低;另一个原因是:庞大的实体向量空间内一定具有聚类密度的差异性,目前并不能很好地处理这种差异,所以虽然可以得到较高的 MeanRank,但 Hits@10 的精度不够高。

表 3　链路预测实验结果

Dataset	WN18				FB15k			
Model	MeanRank		Hits@10		MeanRank		Hits@10	
	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter
RECAL	1 080	1 163	37.2	52.8	828	683	28.4	44.1
LFM	469	456	71.4	81.6	283	164	26.0	33.1
TransE	263	251	75.4	89.5	243	125	34.9	47.1
TransM	292	280	---	---	197	94	---	---
TransH(unif/bern)	318/401	303/388	75.4/73.0	86.7/82.3	211/212	84/87	42.5/45.7	58.5/64.4
TransR(unif/bern)	232/238	219/225	78.3/79.8	91.7/92.0	226/198	78/77	43.8/48.2	65.5/68.7
KG2E_KL	362/342	348/331	80.5/80.2	93.2/92.8	183/174	69/59	47.5/48.9	71.5/74.0
TransD(unif/bern)	242/224	229/212	79.2/79.6	92.5/92.2	211/194	67/91	49.4/53.4	74.2/77.3
TransSparse(unif/bern)	233/223	221/211	79.6/80.1	93.4/93.2	216/190	66/82	50.3/53.7	78.4/79.9
Gtrans-SW(unif/bern)	247/215	234/202	79.1/80.2	92.9/93.5	207/189	66/85	50.6/52.9	75.1/75.3
TransE+GAN- scratch	---	244	---	92.7	---	90	---	73.1
TransE+GAN- pretrain	---	240	---	91.3	---	81	---	74
TransE-SNS(unif/bern)	220/207	208/195	80.2/80.6	94.0/94.6	198/210	56/95	48.9/52.5	80.1/83.0
TransE-DNS(unif/bern)	**182**/**176**	**170**/**165**	77.4/81.2	90.1/93.9	**176**/208	**53.4**/92	48.3/50.6	74.8/78.5

3.3　三元组分类

三元组分类用于验证 Trans_DNS 模型正确区分正负例三元组的性能。实验选择了 WN11、FB13 和 FB15K 三个数据集,其中由 Socher[21] 等提供的 WN11 和 FB13 测试集包含了正负例三元组。而 FB15K 中的测试集只有正例三元组,于是按照 FB13 负例三元组的生成方式,为 FB15 K 构造了负例三元组。

在三元组分类中,数据集中每个关系 r 都设置了阈值 θ_r。对于给定的三元组,如果其得分小于 θ_r,被归为正例,反之则归为负例。关系阈值 θ_r 由验证集获得最大分类精度时的阈值决定。

表 4　三元组分类任务参数设置

Dataset	epoch	α	γ	n	β	M	O	T
WN18	1 000	0.01	5	50	3 000	15	3	50
FB13	2 000	0.003	2	100	4 800	15	3	50
FB15K	2 000	0.003	2.5	100	10 000	10	5	50

三元组分类的参数如表 4 所示,其中 WN11 和 FB15K 在经过多次实验后,均采用了链路预测任务的参数。FB13 根据链路预测中参数的选择范围进行了

多次实验,并得到了最优参数。

三元组分类实验结果如表 5 所示,从表中可以得到 TransE-DNS(bern)在 FB13 上得到了最优解,且在 WN11 和 FB15K 上的性能优于大部分文献中的模型。总体来看,在三元组分类的实验中,Trans_DNS 得到了不错的实验结果,证明了 DNS 负采样优化了模型区分正负三元组的能力。

表 5 三元组分类实验结果

模型	WN11	FB13	FB15K
SE	53.0	75.2	---
LFM	73.8	84.3	---
NTN	70.4	87.1	68.2
TransE	75.9	81.5	79.8
TransE -NMM	79.5	77.2	---
TransH(unif/bern)	77.7/78.8	76.5/83.3	74.2/79.9
TransR(unif/bern)	85.5/85.9	74.7/82.5	81.1/82.1
TranSparse(unif/bern)	86.8/86.8	86.5/87.5	87.4/88.5
TransSparseDT	86.7	85.3	88.9
TransA	93.2	82.8	87.7
TransE+GAN- pretrain	85.4	85.2	---
TransE-SNS	83.2	87.1	86.6
TransE-DNS(unif/bern)	85.4/85.6	82.5/**88.5**	87.5/87.9

4 结束语

传统的知识图谱嵌入方法为了提升模型的训练速度,没有过多的从负采样的角度出发去优化模型,导致了大量的低质量负样本,对模型的训练没有帮助,最终影响了模型的性能。针对这个问题,该文从实体的相似度出发,先采用 DBSCAN 聚类的方式对大部分在向量空间中联系紧密的点进行聚类,再针对离群点采用过采样的方式生成假样本,抽选与其接近的真实样本点,解决了数据稀疏所导致的负采样效果不理想的问题。不足之处在于,没有处理好实体向量空间的局部密度差异性,这会导致整体的聚类效果变差。

未来将会尝试把 DNS 负采样扩展到其他知识表示模型中。同时,下一步的想法是,如何提取实体与关系之间更多深层次非线性特征,并采用多模态的聚类方式,强化实体点分类的精度,进一步提高负采样的质量,提升知识图谱嵌入模型的性能。

参考文献:

[1] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase:a collaboratively created graph database for structuring human knowledge[C]//2008 ACM SIGMOD international conference on management of data. Vancouver:ACM,2008:1247-1250.

[2] JIA Y,WANG Y,CHENG X,et al. OpenKN:an open knowledge computational engine for network big data[C]//The IEEE/ACM international conference on advances in social networks analysis and mining. Beijing: IEEE, 2014: 657-664.

[3] MILLER G A. WordNet:a lexical database for English[J]. Communications of the ACM,1995,38(11):39-41.

[4] WU W,LI H,WANG H,et al. Probase:a probabilistic taxonomy for text understanding[C]//2012 international conference on management of data (SIGMOD). Scottsdale:ACM, 2012:481-492.

[5] 王国霞,刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用,2012,48(7):66-76.

[6] 刘知远,孙茂松,林衍凯,等. 知识表示学习研究进展[J]. 计算机研究与发展,2016,53(2):247-261.

[7] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multirelational data [C]//International conference on neural information processing systems. South Lake Tahoe:[s. n.], 2013:2787-2795.

[8] 饶官军,古天龙,常 亮,等. 基于相似性负采样的知识图谱嵌入[J]. 智能系统学报,2020,15(2):218-226.

[9] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//The 2nd international conference on knowledge discovery and data mining. Portland: AAAI, 1996:226-231.

[10] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE:synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research,2002,16(1):321-357.

[11] BORDES A, WESTON J, USUNIER N. Open question answering with weakly supervised embedding models[C]//JointEuropean conference on machine learning and knowledge discovery in databases. Nancy:[s. n.], 2014: 165-180.

[12] WANG Zhen,ZHANG Jianwen,FENG Jianlin,et al. Knowledge graph embedding by translating on hyper-planes[C]//The 28th AAAI conference on artificial intelligence. Qu6bec City:AAAI,2014:1112-1119.

[13] LIN Yankai,LIU Zhiyuan,SUN Maosong,et al. Learning entity and relation embeddings for knowledge graph completion [C]//The 29th AAAI conference on artificial intelligence. Austin:AAAI,2015:2181-2187.

[14] JI Guoliang,HE Shizhu,XU Liheng,et al. Knowledge graph embedding via dynamic mapping matrix[C]//The 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing. Beijing:[s. n.],2015:687-696.

[15] JI Guoliang,LIU Kang,HE Shizhu,et al. Knowledge graph

(下转第 181 页)