

doi: 10.12052/gdutxb.210062

概率条件下基于双目标交替优化的知识表示模型

张欣, 王振友

(广东工业大学 数学与统计学院, 广东 广州 510520)

摘要: 针对TransD模型参数多和实体两种表示间没有关联的问题, 提出一种改进的知识表示模型PTransD。通过减少实体投影数, 并对实体进行聚类来减少参数量, 同时利用K-L(Kullback-Leibler)散度限制实体投影和对应实体类, 使其概率分布相同。在模型训练时, 对三元组损失和K-L损失交替优化, 从类间距大的实体类中替换实体, 提高负例质量。最后, 在知识图谱数据集上进行三元组分类和链接预测实验。结果表明, 该模型的性能在各项指标上均有明显提高, 可以应用于知识图谱的完善和推理等。

关键词: 知识图谱; 表示学习; 交替优化; 三元组分类; 链接预测

中图分类号: TP3

文献标志码: A

文章编号: 1007-7162(2022)04-0024-08

A Knowledge Representation Model Based on Bi-Objective Alternate Optimization Under Probability

Zhang Xin, Wang Zhen-you

(School of Mathematics and Statistics, Guangdong University of Technology, Guangzhou 510520, China)

Abstract: Aiming at the problem that the TransD model has many parameters and the two representations of entities are not related, an improved knowledge representation model PTransD is proposed, which reduces the number of parameters by reducing the number of entity projections and clustering entities, while using K-L (Kullback-Leibler) The divergence limits the entity projection to the same probability distribution as the corresponding entity class. During model training, the triple loss and K-L loss are alternately optimized, and the entities in the classes with large spacing between the entities are replaced to improve the quality of negative examples. Finally, based on the experimental results of triple classification and link prediction on the knowledge graph data set, the performance has been significantly improved in various indicators. It can be applied to the perfection and reasoning of knowledge map.

Key words: knowledge graph; representation learning; alternate optimization; triple classification; link prediction

1960年, Quillian在进行自然语言理解的应用研究时提出了语义网络概念, 侧重描述概念间语义关系, 这是知识图谱的起源。语义网和链接数据是Tim Berners Lee分别在1998年和2006年提出的, 是知识图谱发展的基础。随后, 学者们构建了大量大规模知识库, 包括常识知识库Cyc^[1], 词典知识库WordNet^[2]和世界开源知识库FreeBase^[3]等。2012年, 知识图谱的概念由Google正式提出, 主要侧重强调数据或事物之间的关联, 随后亚马逊的Amazon Neptune、微软的多模数据库CosmosDB以及腾讯云的“星图”等陆续出现。

知识图谱是结构化的语义知识库, 以三元组 (h, r, t) 或图数据库的形式存储在网络中, 其中实体 h, t 和关系 r 以知识网络的形式来表示。知识图谱技术包括知识表示、知识图谱构建和知识图谱应用。知识表示学习是指对知识库中的实体和关系进行学习, 旨在将研究对象的语义信息表示为稠密低维实值向量。这能够有效解决计算效率低和数据稀疏问题, 可以用于知识获取、知识推理、推荐和问答等任务中^[4-6]。

知识表示学习主要包括基于平均距离的模型、

收稿日期: 2021-04-25

基金项目: 广东省基础与应用基础研究基金资助项目(2020B1515310001)

作者简介: 张欣(1998-), 女, 硕士研究生, 主要研究方向为数据分析、算法设计与分析

通信作者: 王振友(1979-), 男, 教授, 博士, 主要研究方向为医学影像学、最优化理论及应用、数值计算等, E-mail: zywang@gdut.edu.cn

语义匹配模型以及融合多源信息的模型等^[7]。受word2vec模型^[8]的“词向量在语义空间中具有平移不变性”启发,Bordes等^[9]提出了TransE模型。TransE简单高效,但是在处理复杂关系时性能不佳。针对TransE的局限性,研究者们相继提出了TransE的很多变体,其中包括TransR^[10]和TransD模型^[11]。TransD是在TransR基础上引入了投影向量的概念,这使模型参数量增加了一倍,且实体两种表示之间的关系不明晰,模型仍存在不可解释性等问题。

为此,本文针对TransD模型的缺陷,提出了一种概率分布下双目标交替优化的知识表示模型(以下简称PTransD)。首先,通过聚类算法构造 k 个实体类与实体投影一一对应,定义实体类中心为每一类实体语义向量的均值,采用“实体类中心与实体投影的欧氏距离最短则归属一类”的原则来判断实体投影所对应的实体类,这大大减少了模型参数的数目。然后本文采用概率代替欧氏距离来表示相似性,借鉴概率分布的原理,用K-L散度值来衡量该任务的损失。最后将三元组损失学习到的实体表示作为K-L散度损失的先验知识,双目标交替优化^[12]训练得到实体和关系的向量表示。在模型训练时,基于伯努利采样方法,从类间距大的实体类中选择实体进行负采样,并在数据集上进行实验分析,验证模型的有效性。

1 相关工作

平均距离模型是知识表示学习中的代表模型,采用基于距离的评分函数,用头实体通过关系进行翻译之后的实体和尾实体之间的距离来测量事实三元组的合理性,其中较有代表性的方法包括TransE、TransH^[13]、TransR、TransD、TransF^[14]以及TransGH^[15]等。

词向量模型word2vec使每个词可以映射到一个向量,以表示词对词之间的关系。Bordes等^[9]受到此现象启发,提出TransE模型。对于每个三元组 (h, r, t) ,关系向量 r 是从头实体向量 h 到尾实体向量 t 的一个翻译操作。TransE提出 $h + r - t \approx 0$ 的观点,要求正确的尾实体 t 分布在 $h + r$ 附近,定义得分函数为 $f(h, r, t) = \|h + r - t\|_{L_1/L_2}$, L_1 为1范数, L_2 为2范数。TransE模型参数少,计算简单,缺点是在遇到自反、一对多、多对一以及多对多等复杂关系时,不同实体在同一关系中会有相同的向量表示。

TransH克服了TransE的上述缺点,将实体和关系嵌入到统一的向量空间,把实体投影到关系的超

平面中进行翻译,即 $h_{\perp} = h - w_r^T h w_r$, $t_{\perp} = t - w_r^T t w_r$, w_r 为关系超平面的单位法向量; $h_{\perp}$ 为 h 在实体空间翻译以后在关系空间里的向量表示; $t_{\perp}$ 为 t 在实体空间翻译以后在关系空间里的向量表示。定义得分函数为 $f(h, r, t) = \|h_{\perp} + r - t_{\perp}\|$ 。TransH使实体在不同关系中有不同的表示,且没有增加模型的复杂度和训练难度。

TransE和TransH都是在同一空间中考虑实体和关系,然而从本质上看实体和关系是不同的客观事物,在同一空间中表示是不合理的。TransR提出不同的关系关注实体的不同属性,应具有不同的语义空间,将不同关系下的实体投影到不同的语义空间中进行翻译,即 $h_{\perp} = h M_r$, $t_{\perp} = t M_r$, M_r 为只和关系 r 有关的投影矩阵。

为解决投影矩阵只和关系有关的问题,TransD提出每个实体和关系都有两种表示,一种是捕获实体和关系的含义,表示为 h, r, t ;另一种是构造投影矩阵,表示为 h_{pj}, r_{pj}, t_{pj} ,则投影矩阵 $M_{rh} = r_{pj} h_{pj}^T + I^{m \times n}$, $M_{rt} = r_{pj} t_{pj}^T + I^{m \times n}$,其中 $h, h_{pj}, t, t_{pj} \in \mathbb{R}^n$ 和 $r, r_{pj} \in \mathbb{R}^m$, $I^{m \times n}$ 表示行列下角标值相等时元素为1,否则为0的一个 $m \times n$ 矩阵。TransD中实体两种表示之间没有关系,但实验结果显示同一类实体具有相似的投影向量,故实体投影和实体语义向量间应存在一定联系。TransD仅使用了向量运算,计算复杂度较小,适用于大规模知识图谱。

TransR认为每个关系都有一个关系空间,忽略了关系空间的结构,例如关系“出生地”可以推断出“国籍”。针对这一缺陷,TransF建模投影矩阵的基子空间来减轻关系投影的负担,将投影矩阵分解为关系空间的坐标矩阵和基张量的积,即 $M_{rh} = \sum_i \alpha_r^{(i)} U^{(i)} + I^{m \times n}$, $M_{rt} = \sum_i \beta_r^{(i)} V^{(i)} + I^{m \times n}$, $U^{(i)}, V^{(i)}$ 分别为第 i 个头尾实体的关系基空间, $\alpha_r^{(i)}, \beta_r^{(i)}$ 分别为对应的关系系数。

TransH、TransR和TransD都考虑了复杂关系的映射属性,但计算量大。TransGH在TransH的基础上提出广义关系特定超平面的概念,使用一组基向量来代替单个法向量,得到投影实体向量为 $h_{\perp} = h - \sum_i (w_i^T)^T h w_i^T$, $t_{\perp} = t - \sum_i (w_i^T)^T t w_i^T$,其中 w_i^T 表示关系超平面法向量的第 i 个基向量。

表1列出了以上所有提及的知识表示模型的复杂度。PTransD在时间复杂度相等的情况下,空间复杂度远小于TransD,而相比于TransH, PTransD的复杂度相差不大,且实验结果和性能较好,验证详见第3节。

表1 各嵌入模型的复杂度

Table 1 Complexity of several embedding models

模型名称	参数	时间复杂度 ¹⁾
TransE ^[9]	$O(N_e m + N_r n)(m = n)^{2)}$	$O(N_t)^3)$
TransR ^[10]	$O(N_e m + N_r(m+1)n)$	$O(2mnN_t)$
TransD ^[11]	$O(2N_e m + 2N_r n)$	$O(2nN_t)$
TransH ^[13]	$O(N_e m + 2N_r n)(m = n)$	$O(2mN_t)$
TransF ^[14]	$O(N_e m + N_r(m+2v)n), v \ll m$	$O(2vmnN_t)$
TransGH ^[15]	$O(N_e m + N_r(1+v)n)(m = n), v \ll m^4)$	$O(2vmN_t)$
PTransD	$O((N_e + k)m + 2N_r n)^5)$	$O(2nN_t)$

1) 时间复杂度表示一次迭代中乘法操作的数量。

2) N_e 和 N_r 分别代表实体和关系的数量, m 和 n 分别表示实体空间和关系空间的维数。

3) N_t 代表一个知识图谱中三元组的数目。

4) v 表示一种关系的向量数。

5) k 代表实体聚类簇数或实体投影向量数目。

2 PTransD模型

基于TransD模型存在的问题,本文详细地介绍了在此基础上改进的知识表示模型PTransD。PTransD使用聚类算法和概率分布相似的原理来克服TransD的模型参数多和实体两种表示之间无联系的缺陷,并将得分函数的距离模型和概率分布相似模型集成一个模型,从而增强模型的代表能力。

2.1 实体表示与聚类结合

模型参数越多,模型的拟合能力越强,容易出现过拟合,导致模型泛化能力差,因此本文减少了投影向量个数。假定实体投影向量个数为 k ,对实体语义向量 $\mathbf{e}$ (包括 $\mathbf{h}$ 和 $\mathbf{t}$)聚成 k 类,使实体类和实体投影形成一一对应的关系。聚类标准为计算每个 $\mathbf{e}$ 到所有实体投影 $\mathbf{e}_{pj}$ 之间的欧氏距离,每个 $\mathbf{e}$ 都对应一个距离最近的 $\mathbf{e}_{pj}$,对应同一个 $\mathbf{e}_{pj}$ 的实体自成一类。

为便于数学表达,记 $\mathbf{e}_{pj}$ 和对应实体集组成一个邻域 $U(\mathbf{e}_{pj})$, $\mathbf{e}_{pj}$ 为该邻域的邻域中心,对应实体集中的元素 $\mathbf{e}$ 被称作邻域向量,记作 $\mathbf{e} \in U(\mathbf{e}_{pj})$,两两邻域之间互不相交。 k 个实体投影分别为 $\mathbf{e}_{pj}^0, \mathbf{e}_{pj}^1, \dots, \mathbf{e}_{pj}^k$,在实体空间中对应 k 个邻域,所有实体就被自然地分为了至多 k 类。如图1所示,在第 i 个邻域 $U(\mathbf{e}_{pj}^i)$ 中,它们所对应的投影向量都为 $\mathbf{e}_{pj}^i$ 。

所有实体聚类完成后,头尾投影矩阵分别表示为

$$\mathbf{M}_h = \mathbf{r}_{pj}(\mathbf{e}_{pj}^i)^T + \mathbf{I}^{m \times n}, \quad \mathbf{h} \in U(\mathbf{e}_{pj}^i) \quad (1)$$

$$\mathbf{M}_t = \mathbf{r}_{pj}(\mathbf{e}_{pj}^j)^T + \mathbf{I}^{m \times n}, \quad \mathbf{t} \in U(\mathbf{e}_{pj}^j) \quad (2)$$

$$i, j \in \{1, 2, \dots, k\}$$

定义关系空间中被投影的头实体 $\mathbf{h}_\perp$ 和尾实体

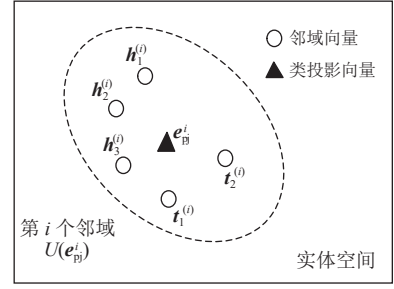


图1 实体空间实体分布示意图

Fig.1 Schematic diagram of entity distribution in entity space

$\mathbf{t}_\perp$ 的表示为

$$\mathbf{h}_\perp = \mathbf{M}_h \mathbf{h} \quad (3)$$

$$\mathbf{t}_\perp = \mathbf{M}_t \mathbf{t} \quad (4)$$

特别地,当 $m \geq n$ 时,式(3)、式(4)可以进一步表示为

$$\mathbf{h}_\perp = \mathbf{M}_h \mathbf{h} = (\mathbf{e}_{pj}^i)^T \mathbf{h} \mathbf{r}_{pj} + [\mathbf{h}^T, \mathbf{0}^T]^T, \mathbf{h} \in U(\mathbf{e}_{pj}^i) \quad (5)$$

$$\mathbf{t}_\perp = \mathbf{M}_t \mathbf{t} = (\mathbf{e}_{pj}^j)^T \mathbf{t} \mathbf{r}_{pj} + [\mathbf{t}^T, \mathbf{0}^T]^T, \mathbf{t} \in U(\mathbf{e}_{pj}^j) \quad (6)$$

本文采用距离函数模型,被投影到关系 r 平面的头实体 $\mathbf{h}_\perp$ 和尾实体 $\mathbf{t}_\perp$ 满足 $\mathbf{h}_\perp + \mathbf{r} - \mathbf{t}_\perp \approx \mathbf{0}$,得分函数为

$$f(\mathbf{h}, \mathbf{r}, \mathbf{t}) = -\|\mathbf{h}_\perp + \mathbf{r} - \mathbf{t}_\perp\|_{L_1/L_2} \quad (7)$$

其可以使用 L_1 距离或者 L_2 距离。

2.2 实体空间内的概率分布

对三元组 $(\mathbf{h}, \mathbf{r}, \mathbf{t})$ 来说,实体 $\mathbf{h}$ 、 $\mathbf{t}$ 及关系 $\mathbf{r}$ 都有两种表示,用于构建投影矩阵的投影向量 $\mathbf{h}_{pj}$ 、 $\mathbf{t}_{pj}$ 、 $\mathbf{r}_{pj}$ 解决了“投影矩阵只与关系有关”的问题,但三元组的两种表示之间的关联信息并没有在TransD的模型假设中出现。TransD的实验结果显示,同一类的实体具有相似的投影向量。因此,不妨提前对实体投影做出约束,在实体空间中,相距越近的实体 $\mathbf{e}$,越有可能属于同一类,对应的实体投影 $\mathbf{e}_{pj}$ 的距离也应该越近。本文提出猜想:若不同类之间的距离相近,对应的实体投影也应该越近。

测量类间距离有很多种方法,包括平均距离法、最短距离法、重心距离法等,但这些方法都涉及大量的两点距离计算,故本文提出实体类中心的概念,使用实体类中心代替整个实体类进行类间距离计算。在实体空间中,每一类实体语义向量可以确定一个实体类中心,采用算术平均值的方法计算 k 个实体类中心,第 i 类的实体类中心定义为

$$\mathbf{e}^i = \frac{\sum_{\mathbf{e} \in U(\mathbf{e}_{pj}^i)} \mathbf{e}}{N(U(\mathbf{e}_{pj}^i))} \quad (8)$$

式中: $N(U(e_{pi}^i))$ 为属于第 i 个邻域 $U(e_{pi}^i)$ 的实体总数。若第 i 类实体集为空, 则实体类中心定义为该类的投影向量 e_{pi}^i , 以便后续计算。如图1中, 第 i 类邻域的实体类中心为

$$e^i = \frac{h_1^{(i)} + h_2^{(i)} + h_3^{(i)} + t_1^{(i)} + t_2^{(i)}}{5}$$

实体类中心距离越近, 对应的投影向量也越近。常规的做法是使用欧式距离来表示这种相似性, 但是在高维空间中, 每个坐标对欧式距离所做的贡献往往是不同的, 本文把这种距离关系转换成一种概率来表示相似性。

在实体空间2个实体类中心 e^i 和 e^j 中, e^i 以条件概率 p_{ji} 选择 e^j 作为它的临近点。考虑以 e^i 为中心点的高斯分布, 若 e^j 越靠近 e^i , 则 p_{ji} 越大; 若2点相距越远, p_{ji} 越小。因此, p_{ji} 的定义为

$$p_{ji} = \frac{\exp(-\|e^i - e^j\|^2 / 2\sigma_i^2)}{\sum_{d \neq i} \exp(-\|e^i - e^d\|^2 / 2\sigma_i^2)} \quad (9)$$

式中: σ_i 为以 e^i 为中心点的高斯分布的方差, $p_{ii} = 0$ 。

实体类中心 e^i 和 e^j 分别对应的投影向量为 e_{pi}^i 和 e_{pj}^j , e_{pi}^i 以条件概率 q_{ji} 选择 e_{pj}^j 作为它的临近点。类似地, 考虑以 e_{pi}^i 为中心点的高斯分布, 若 e_{pj}^j 距离 e_{pi}^i 越近, 则 q_{ji} 越大, 反之越小。将所有高斯分布的方差均设置为 $1/\sqrt{2}$, q_{ji} 的定义为

$$q_{ji} = \frac{\exp(-\|e_{pi}^i - e_{pj}^j\|^2)}{\sum_{d \neq i} \exp(-\|e_{pi}^i - e_{pd}^d\|^2)} \quad (10)$$

由于 p_{ji} 和 p_{ij} 并不相等, 对应的投影向量的概率 q_{ji} 和 q_{ij} 也不相等, 不具有对称性, 所以为了得到一个更加通用、合理的联合概率分布, 使对于任意的 i, j , 都有 $p_{ji} = p_{ij}$, $q_{ji} = q_{ij}$ 。将式(9)进行对称化处理, 得到新的定义为

$$p_{ij} = \frac{p_{ji} + p_{ij}}{2k}, p_{ij} = p_{ji} \quad (11)$$

同时, 对式(10)改进为

$$q_{ij} = \frac{\exp(-\|e_{pi}^i - e_{pj}^j\|^2)}{\sum_{d \neq i} \exp(-\|e_{pi}^i - e_{pd}^d\|^2)}, q_{ij} = q_{ji} \quad (12)$$

若考虑 e^i 与其他所有实体类中心之间的概率, 则可以构成一个概率分布 P_i , 且满足 $p_{ij} \geq 0$, $\sum_{j=0}^{k-1} p_{ij} = 1$ 。所有实体类中心的概率分布合成一个概率矩阵 $P = (P_1^T, P_2^T, \dots, P_k^T)^T$ 。类似地, 实体投影向量之间也构成了一个概率矩阵 $Q = (Q_1^T, Q_2^T, \dots, Q_k^T)^T$, 满足 $q_{ij} \geq 0$, $\sum_{j=0}^{k-1} q_{ij} = 1$ 。

2.3 模型训练

2.3.1 双训练目标

所有知识表示模型都采用三元组损失函数作为目标函数进行训练, 本文在三元组损失函数的基础上增添K-L散度损失函数作为辅助, 完整的损失函数表示为

$$\begin{aligned} L &= L_{\text{score}} + L_{\text{K-L}} = \\ &\sum_{(h,r,t) \in \mathcal{A}, (h',r',t') \in \mathcal{A}'} \xi(f(h,r,t), f(h',r',t')) + D_{\text{K-L}}(P \| Q) = \\ &\sum_{(h,r,t) \in \mathcal{A}} \sum_{(h',r',t') \in \mathcal{A}'} \max(f(h',r',t') + \gamma - f(h,r,t), 0) + \\ &\sum_i \sum_j p_{ij} \lg\left(\frac{p_{ij}}{q_{ij}}\right) \end{aligned} \quad (13)$$

在式(13)中, 第1个目标表示三元组损失, 目的是为了区分正确三元组和错误三元组, 其中 $\mathcal{A}$ 表示正确三元组的集合, $\mathcal{A}'$ 代表错误三元组的集合, ξ 表示 hinge 损失函数, γ 为正确三元组得分与错误三元组得分之间的距离, 在最小化 L_{score} 过程中, 需要满足限制条件

$$\forall (h,r,t) \in \mathcal{A} \cup \mathcal{A}', \|h\|_2 \leq 1, \|t\|_2 \leq 1, \|r\|_2 \leq 1, \|h_{\perp}\|_2 \leq 1, \|t_{\perp}\|_2 \leq 1, \forall i \in \{0, 1, \dots, k-1\}, \|e^i\|_2 \leq 1, \|e_{pi}^i\|_2 \leq 1$$

第2个目标表示K-L散度损失, 目的是对实体空间的实体类中心和对应的实体投影进行相似性度量。 $D_{\text{K-L}}(P \| Q)$ 为衡量概率分布 P 和 Q 之间的差异。当 $P=Q$ 时, 取最小值为0。2个目标函数相辅相成, 共同优化模型。

2.3.2 负采样策略

训练模型时, 需要损坏知识图谱中的三元组来构建负例三元组。TransE提出的方法是均匀采样(随机替换头尾实体), 但这种抽样方法在处理一对多、多对一以及多对多的复杂关系时, 构建的三元组不是负例的概率较大。针对上述缺点, TransH提出基于伯努利分布的采样, 以不同的概率来替换头尾实体, 降低引入错误负例的概率。本文在伯努利分布采样的基础上, 选择类间距大的类中的实体来替换头尾实体, 以便提高模型对实体的区分度。

1) 以不同的概率替换

在生成负例时, 根据关系的类型不同来设置不同的替换策略。对于一对多关系, 以更高的概率来替换头实体; 对于多对一关系, 以更高的概率来替换尾实体; 对于多对多关系, 相当于多个多对一关系或者一对多关系, 按前两种关系的替换策略来进行。

首先提出2个概念: 在一个关系的所有三元组中, tph 表示每个头实体对应的尾实体的平均数量,

hpt表示每个尾实体对应的头实体的平均数量。当 $tph < 1.5$ 且 $hpt < 1.5$ 时,则认为关系 r 是一对一的;当 $tph > 1.5$ 且 $hpt > 1.5$ 时,则认为关系 r 是多对多的;当 $tph < 1.5$ 且 $hpt \geq 1.5$ 时,则认为关系 r 是多对一的;当 $tph \geq 1.5$ 且 $hpt < 1.5$ 时,则认为关系 r 是一对多的。

其次用参数为 $p = \frac{tph}{tph + hpt}$ 的伯努利分布来抽样。对于给定的一个三元组,以概率 p 替换头实体,以概率 $1 - p$ 替换尾实体,来生成负例三元组。

2) 选择类间距大的类的实体

在前文中,所有实体被划分为了 k 类,每一类实体都具有相似的类型和属性。给定三元组 $(\hat{h}, r, t)$,假设抽样时依据伯努利分布需要替换头实体,选择同一邻域内的实体 $\hat{h}'$ 来替换 $\hat{h}$,依据式(11)可知,翻译到关系空间的 $\mathbf{h}'_{\perp}$ 和 $\mathbf{h}_{\perp}$ 也靠近,产生的负例 $(\hat{h}', r, t)$ 很可能是正确的,导致预测标签错误。因此,尽可能选择类间距大的类中的实体来替换。

不妨假设实体从实体空间翻译到关系空间,并没有改变它们之间的相对距离关系。选择类间距大的类的实体进行替换,确保翻译到关系空间的两个实体也相距较远。对于需替换头实体的三元组,计算头实体所对应的实体投影 $\mathbf{e}_{pj}$ 到所有实体投影的距离,并按距离将对应的类别降序排列,表示为 $\{l_1, l_2, \dots, l_k\}$ 。选取第 l_1 类的实体来替换头实体,若第 l_1 类中没有实体,则依次递推选择第 l_2 类至第 l_k 类中的实体。

对于三元组 $(\hat{h}, r, t)$, $\mathbf{h} \in U(\mathbf{e}_{pj}^i)$ 且 $\mathbf{t} \in U(\mathbf{e}_{pj}^j)$, r 属于一对多关系,应以更大概率替换头实体,由于 $U(\mathbf{e}_{pj}^i)$ 与 $U(\mathbf{e}_{pj}^m)$ 的实体类中心越远,类间距越大,故随机选择 $\mathbf{h}^* \in U(\mathbf{e}_{pj}^m)$ 来替换 $\mathbf{h}$ 。

在关系空间 r 中, $\mathbf{h}_{\perp} + \mathbf{r} \approx \mathbf{t}_{\perp}$,其他正确的尾实体也应分布在 $\mathbf{t}_{\perp}$ 周围,但 $\mathbf{t}^*_{\perp}(\mathbf{h}^*_{\perp} + \mathbf{r})$ 与 $\mathbf{t}_{\perp}$ 相距较远,损坏的三元组 $(\hat{h}^*, r, t)$ 是负例的正确率较高。

2.3.3 实现交替优化算法

模型训练迭代包含2个阶段:三元组损失和K-L散度损失。每次迭代中,首先训练三元组损失2次,得到的实体向量表示作为K-L散度损失的输入,再继续训练1次,这种交替学习的方法在更加关注三元组损失目标的同时,更好地协同优化模型。算法1给出了PTransD的学习算法。

算法1 Learning PTransD

Input: Training set $\mathcal{A} = \{(\hat{h}, r, t)\}$, entities and rel. sets E and L , margin γ , entity and relation embedding dim m and n , number of entity clusters k .

Output: The well trained embedding model.

```
1: initialize all parameters
    $\mathbf{r}, \mathbf{r}_{pj} \leftarrow \text{uniform}(-6/\sqrt{n}, 6/\sqrt{n})$  for each  $r \in L$ 
    $\mathbf{r} \leftarrow \mathbf{r}/\|\mathbf{r}\|$  for each  $r \in L$ 
    $\mathbf{e}, \mathbf{e}_{pj} \leftarrow \text{uniform}(-6/\sqrt{m}, 6/\sqrt{m})$  for each  $e \in E$ 
2: for number of training iteration do
   // triple loss objection
3: for 2 steps do
4:  $\mathbf{e}, \mathbf{e}_{pj} \leftarrow \mathbf{e}/\|\mathbf{e}\|, \mathbf{e}_{pj}/\|\mathbf{e}_{pj}\|$ 
5:  $\Delta_{\text{batch}} \leftarrow \text{sample}(\mathcal{A}, B)$  // sample a minibatch of size  $B$ 
6:  $T_{\text{batch}} \leftarrow \emptyset$  // initialize the set of pairs of triples
7: foreach  $(\hat{h}, r, t) \in \Delta_{\text{batch}}$  do
8:  $(\hat{h}', r, t') \leftarrow \text{sample } \mathcal{A}'$  // sample a corrupted triple
9:  $T_{\text{batch}} \leftarrow T_{\text{batch}} \cup \{(\hat{h}, r, t), (\hat{h}', r, t')\}$ 
10:   end for
11: Update embedding w.r.t
      
$$\sum_{(\hat{h}, r, t) \in \mathcal{A}, (\hat{h}', r, t') \in \mathcal{A}'} \nabla \xi(f(\hat{h}, r, t), f(\hat{h}', r, t'))$$

12:   end for
   // K-L divergence loss objection
13:  $S^i \leftarrow U(\mathbf{e}_{pj}^i) \setminus \mathbf{e}_{pj}^i$  // get entity set of  $U(\mathbf{e}_{pj}^i)$ 
14:  $\mathbf{e}^i \leftarrow \text{Eq. (8)}$  // get center vectors of entity classes
15:  $\mathbf{P}, \mathbf{Q} \leftarrow \text{Eq. (11), (12)}$  // calculate probability matrix
16: Update embedding w.r.t  $\nabla KL(\mathbf{P} \parallel \mathbf{Q})$ 
17: end for
```

3 试验和结果分析

本节介绍PTransD模型的实验部分,通过在知识图谱上进行三元组分类和链接预测来评估模型的性能。首先介绍这两项工作的评价指标和实验结果,然后与其他模型方法的实验结果进行对比分析。

3.1 数据集

WordNet是世界著名的大型英语词典知识库,其名词、动词、形容词和副词被各自组成同义词网络,并通过关系连接,可用于语义消歧;FreeBase是一个完全结构化的大型知识库,其内容主要来自其社区成员的贡献和多种多样的数据库。本文在WordNet的子集(WN18和WN11)和Freebase的子集(FB15K和FB13)上进行实验。统计资料如表2所示,可以看出,WN18包含的实体较多,而FB15K包含的关系类别较多。

表 2 数据集的统计
Table 2 Statistics of datasets

数据集	实体	关系	训练集	有效集	测试集
WN18	40 943	18	141 442	5 000	5 000
FB15K	14 951	1 345	483 142	50 000	59 071
WN11	38 696	11	112 581	2 609	10 544
FB13	75 043	13	316 232	5 908	23 733

3.2 链接预测

在知识图谱中,链接预测的任务是进行实体关系学习,具体地,就是预测一个关系事实三元组 (h,r,t) 中所缺失的 h 和 t 。在该任务中,缺失位置的实体是从知识图谱中筛选出的一组候选实体,按得分进行排序,而不是只得到一个最好的实体。在训练过程中,对每个三元组 (h,r,t) ,用知识图谱中的所有实体来替换头实体或尾实体,并计算得分,然后将替换后的三元组按得分进行降序排列,通过观察原三元组在该序列中的位置来评估模型的性能,位置越靠前说明模型性能越好。

知识图谱中存在一些一对多、多对一以及多对多的复杂关系,一些损坏三元组也存在于知识图谱中,但这些三元组是正确的,排名靠前是合理的。如果直接将这些损坏三元组认定为负例进行训练,会降低模型的表示能力。为了避免这种情况产生,将这种正确的损坏三元组从训练集、验证集和测试集中去除,该实验设置称为“Filt”,而没有经过去除处理的实验设置称为“Raw”。

3.2.1 评价指标

对所有训练的三元组按得分进行综合排列,用 2 个常用评价指标衡量模型优劣。一是平均排序 (Mean Rank),表示正确实体在所有候选实体中的平均排名,排名值越低,正确实体在排列中的位置越靠前,模型性能越好;二是 HIT@10,表示正确实体排在前十名的概率,概率越大,模型预测越准确。

3.2.2 实验设置

在这个任务中,使用 WN18 和 FB15K 作为数据集,并都采用 Adadelata SGD 算法^[16]作为优化方法,设

置超参数: $\varepsilon = 1 \times 10^{-6}$, $\rho = 0.95$ 。在训练 PTransD 时,在 $\{0.25,0.5,1,2\}$ 中选择边际 γ ,在 $\{20,50,80,100\}$ 中选择实体维度 m 和关系 n ,在 $\{100,200,1\ 000,1\ 400\}$ 中选择 batch 的大小 B ,在 $\{20,50,100,200\}$ 中选择聚类个数 k 。最佳的参数由验证集确定。

“unif”表示均匀采样,“bern”表示基于伯努利分布的采样。在“unif”设置下:在 WN18 上, $\gamma = 1, m = 50, n = 50, B = 200, k = 50$;在 FB15K 上, $\gamma = 0.5, m = 50, n = 100, B = 1\ 400, k = 200$ 。在“bern”设置下:在 WN18 上, $\gamma = 1, m = 50, n = 80, B = 1\ 000, k = 50$;在 FB15K 上, $\gamma = 0.25, m = 100, n = 50, B = 1\ 400, k = 100$ 。对于这 2 个数据集,本实验将所有训练三元组迭代 500 次。

3.2.3 实验结果

PTransD 的链接预测实验结果如表 3 所示,表中加粗的数字表示在同一指标下最优模型的实验结果。结果标明:(1) 相对于原模型 TransD, PTransD 模型的 2 个指标值有部分提升, HIT@10 指标值提升更明显;(2) 对比 2 个数据集上的指标值, PTransD 在 FB15K 上的结果较好,证明该模型在关系复杂且信息稠密的知识图谱上性能更优;(3) PTransD 相对于其他模型来说, HIT@10 值较高,证明其学习的能力更好。

为了验证 PTransD 确实能够较好地处理各种复杂关系,进一步对不同关系类型的三元组进行实验。选择具有更多关系类型的 FB15K 数据集来进行验证。在 1 345 个关系中, 1-1 关系占 24%, 1- n 关系占 23%, n -1 关系占 29%, m - n 关系占 24%, 各关系的比例十分均衡。实验结果如表 4 所示,表中加粗的数字表示在同一指标下最优模型的实验结果。结果表明:

表 3 链接预测实验结果
Table 3 Results of link prediction

模型名称	WN18实验结果				FB15K实验结果			
	Mean Rank		HIT@10/%		Mean Rank		HIT@10/%	
	Raw	Filt	Raw	Filt	Raw	Filt	Raw	Filt
Unstructured ^[17]	315	304	35.3	38.2	1 074	979	4.5	6.3
SE ^[18]	1 011	985	68.5	80.5	273	162	28.8	39.8
SME(Linear) ^[19]	545	533	65.1	74.1	274	154	30.7	40.8
SME(Bilinear) ^[19]	526	509	54.7	61.3	284	158	31.3	41.3
TransE ^[9]	263	251	75.4	89.2	243	125	34.9	47.1
TransH ^[13]	401	388	73.0	82.3	212	87	45.7	64.4
TransGH ^[15]	210	197	81.6	95.3	186	64	54.1	80.1
TransR ^[10]	238	225	79.8	92.0	198	77	48.2	68.7
TransF ^[14]	—	198	—	95.3	—	62	—	82.3
TransD ^[11]	224	212	79.6	92.2	194	91	53.4	77.3
PTransD(unif)	239	224	80.1	92.0	205	87	55.6	81.2
PTransD(bern)	220	210	82.4	94.1	192	72	55.6	82.4

(1) 相比于TransD模型, PTransD模型在复杂关系上的HIT@10值明显提高;

(2) 相比于其他模型, PTransD模型能较好地区分1-1关系以及 m - n 关系, 在另外两种关系中性能表现也较好。

表 4 FB15K各类关系的HIT@10值
Table 4 HIT@10 of each type of relations in FB15K

模型名称	预测头实体的HIT@10				预测尾实体的HIT@10				%
	1-1	1- n	n -1	m - n	1-1	1- n	n -1	m - n	
Unstructured ^[17]	34.5	2.5	6.1	6.6	34.3	4.2	1.9	6.6	
SE ^[18]	35.6	62.6	17.2	37.5	34.9	14.6	68.3	41.3	
SME (Bilinear) ^[19]	30.9	69.6	19.9	38.6	28.2	13.1	76.0	41.8	
TransE ^[9]	43.7	65.7	18.2	47.2	43.7	19.7	66.7	50.0	
TransH ^[13]	66.8	87.6	28.7	64.5	65.5	39.8	83.3	67.2	
TransGH ^[15]	87.0	95.8	47.9	80.8	86.8	55.7	94.8	84.3	
TransR ^[10]	78.8	89.2	34.1	69.2	79.2	37.4	90.4	72.1	
TransF ^[14]	88.1	94.9	53.2	82.8	88.8	62.1	93.4	85.8	
TransD ^[11]	86.1	95.5	39.8	78.5	85.4	50.6	94.4	81.2	
PTransD(unif)	89.7	92.7	44.2	82.0	87.8	52.1	94.2	84.7	
PTransD(bern)	89.5	94.5	45.7	82.8	88.9	52.5	94.1	85.8	

3.3 三元组分类

三元组分类的目标是判断一个给定的三元组 (h, r, t) 是否正确。训练过程中, 知识图谱中的三元组被判断为“正确”, 每个正例三元组进行负采样得到的三元组被判断为“错误”。在关系 r 下的所有三元组可通过计算最大分类精度值得到一个阈值 σ_r , 用来判断一个三元组“正确”或“错误”。如果其得分函数不超过阈值 σ_r , 就预测为正确, 反之为错误。

3.3.1 评价指标

三元组分类任务使用准确率 A 作为评价指标, 公式为

$$A = \frac{T_{\text{pos}} + T_{\text{neg}}}{N_{\text{pos}} + N_{\text{neg}}}$$

(14)

式中: T_{pos} 和 T_{neg} 分别为预测正确的正例三元组个数和负例三元组个数, N_{pos} 和 N_{neg} 表示训练集中的正例三元组个数和负例三元组个数。 A 越高, 模型进行三元组分类任务的能力越强。

3.3.2 实验设置

这个任务采用Adadelata SGD算法作为优化方法, 并设置超参数 ε 为 1×10^{-6} , ρ 为0.95。在训练PTransD时, 在 $\{1, 2, 4\}$ 中选择边际 γ , 在 $\{20, 50, 100\}$ 中选择实体维度 m 和关系 n , 在 $\{20, 120, 480, 960, 4\ 800\}$ 中选择batch的大小 B , 在 $\{20, 50, 100, 200\}$ 中选择聚类个数 k 。最佳的参数由验证集确定。WN11上的最佳配置为 $\gamma = 2, m = 100, n = 100, B = 960, k = 50$, 并且使用 L_1 作为相似性度量; FB13上的最佳配置为 $\gamma = 1, m = 50, n = 50, B = 4\ 800, k = 200$, 并且使用 L_1 作为相似性度

量; FB15K上的最佳配置为 $\gamma = 1, m = 100, n = 100, B = 4\ 800, k = 100$, 并且使用 L_1 作为相似性度量。

3.3.3 实验结果

表5列出了不同模型的三元组分类精度, 表中加粗的数字表示在同一指标下最优模型的实验结果。在3个数据集上, PTransD模型都比TransD模型分类能力更好, 这说明PTransD模型更适用于大规模知识图谱。

表 5 不同模型的三元组分类精度
Table 5 Accuracy of triple classification of different models

模型名称	三元组分类精度			%
	WN11	FB13	FB15K	
SE ^[18]	53.0	75.2	—	
SME(Bilinear) ^[19]	70.0	63.7	—	
TransE ^[9]	75.9	81.5	79.8	
TransH ^[13]	78.8	83.3	79.9	
TransGH ^[15]	87.3	85.2	91.2	
TransR ^[10]	85.9	82.5	82.1	
TransD ^[11]	86.4	89.1	88.0	
PTransD(unif)	88.0	89.0	90.0	
PTransD(bern)	88.2	89.9	91.2	

4 结 论

本文提出了一种概率分布下基于双目标交替优化的知识表示模型PTransD。针对翻译的模型TransD参数多的问题, PTransD限制实体投影个数, 对实体进行聚类, 将“实体语义向量和实体投影两种

表示属于一一对应的关系”转变成“实体类和实体投影属于一一对应的关系”。针对实体两种表示之间的关系无约束的问题,PTTransD在对实体进行聚类的基础上,通过求平均值计算每类的实体类中心,利用概率代替欧氏距离来衡量实体类中心和实体投影的概率分布相似性,加强对实体投影的约束。采用交替优化的方法获得三元组损失和概率分布下的K-L散度损失,并共同训练模型。为了验证方法的有效性,在WordNet和FreeBase的大规模真实数据集上对链接预测和三元组分类任务进行了综合测评。实验结果表明,PTTransD模型有较好的性能,可以应用于知识图谱的完善和推理中。

在将来的研究中继续改进PTTransD模型,针对关系空间中关系的两种表示之间的相关性,引入关系路径;还可将PTTransD模型应用于涉及关系抽取、知识推理的任务中。

参考文献:

- [1] LENAT D B, PRAKASH M, SHEPHERD M. CYC: using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks [J]. *AI Magazine*, 1985, 6(4): 65.
- [2] MILLER G A. WordNet [J]. *Communications of the ACM*, 1995, 38(11): 39-41.
- [3] BOLLACKER K, COOK R, TUFTS P. Freebase: a shared database of structured general human knowledge[C]//Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence. Vancouver: AAAI Press, 2007: 1962-1963.
- [4] JI S, PAN S, CAMBRIA E, *et al.* A survey on knowledge graphs: representation, acquisition and applications [EB/OL]. (2021-04-01)[2021-04-20]. <https://arxiv.org/abs/2002.00388>.
- [5] GAO Y, LI Y, LIN Y, *et al.* Deep learning on knowledge graph for recommender system: a survey[EB/OL]. (2020-03-25) [2021-04-20]. <https://arxiv.org/abs/2004.00387>.
- [6] HUANG X, ZHANG J, LI D, *et al.* Knowledge graph embedding based question answering [C]//Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. New York: ACM, 2019: 105-113.
- [7] DAI Y, WANG S, XIONG N N, *et al.* A survey on knowledge graph embedding: approaches, applications and benchmarks [J]. *Electronics*, 2020, 9(5): 750.
- [8] MIKOLOV T, SUTSKEVER I, CHEN K, *et al.* Distributed representations of words and phrases and their compositionality [J]. *Advances in Neural Information Processing Systems*, 2013, 26: 3111-3119.
- [9] BORDES A, USUNIER N, GARCIA-DURAN A, *et al.* Translating embeddings for modeling multi-relational data[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe: NIPS, 2013: 2787-2795.
- [10] LIN Y, LIU Z, SUN M, *et al.* Learning entity and relation embeddings for knowledge graph completion[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Austin: AAAI Press, 2015: 2181-2187.
- [11] JI G, HE S, XU L, *et al.* Knowledge graph embedding via dynamic mapping matrix[C] //Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing: ACL, 2015: 687-696.
- [12] WANG H, ZHANG F, ZHAO M, *et al.* Multi-task feature learning for knowledge graph enhanced recommendation[C]//Proceedings of the 2019 World Wide Web Conference. San Francisco: ACM, 2019: 2000-2010.
- [13] WANG Z, ZHANG J, FENG J, *et al.* Knowledge graph embedding by translating on hyperplanes[C] //In Proceedings of the Twenty-eighth AAAI Conference on Artificial Intelligence. Québec City: AAAI Press, 2014: 1112-1119.
- [14] DO K, TRAN T, VENKATESH S. Knowledge graph embedding with multiple relation projections[C]//2018 24th International Conference on Pattern Recognition. Beijing: IEEE, 2018: 332-337.
- [15] ZHU Q, ZHOU X, TAN J, *et al.* Learning knowledge graph embeddings via generalized hyperplanes[C] //Proceedings of the 18th International Conference on Computational Science. Wuxi: Springer, 2018: 624-638.
- [16] ZEILER M D. Adadelat: an adaptive learning rate method[EB/OL]. (2012-12-22) [2021-04-20]. <https://arxiv.org/abs/1212.5701v1>.
- [17] BORDES A, GLOROT X, WESTON J, *et al.* Joint learning of words and meaning representations for open-text semantic parsing[C]//Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics. La Palma: PMLR, 2012: 127-135.
- [18] BORDES A, WESTON J, COLLOBERT R, *et al.* Learning structured embeddings of knowledge bases[C]//In Proceedings of the Twenty-eighth AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press, 2011: 301-306.
- [19] BORDES A, GLOROT X, WESTON J, *et al.* A semantic matching energy function for learning with multi-relational data [J]. *Machine Learning*, 2014, 94(2): 233-259.

(责任编辑: 张玮欣)