

基于生成对抗网络和元路径的异质网络表示学习

蒋宗礼 樊珂 张津丽

北京工业大学信息学部 北京 100124

(jiangzl@bjut.edu.cn)



摘要 现实世界中的信息网络大多为异质信息网络,旨在表示低维空间中节点数据的网络表示方法已普遍用于分析异质信息网络,从而有效融合异质网络中丰富的语义信息和结构信息。但是现有的异质网络表示方法通常采用负采样从网络中随机选择节点,并且对节点和边的异质性学习能力不足。受生成式对抗网络和元路径的启发,文中提出了一种新型的异质网络表示方法。首先对采样方法使用元路径的策略进行改进,根据元路径不同的权重取样,使样本更好地体现节点之间存在的直接和间接关系,增强样本的语义关联。然后在生成对抗的博弈过程中使模型充分考虑节点和边的异质性并具备关系感知能力,实现对异质信息网络的表示学习。实验结果表明,与目前的表示算法相比,该模型学习到的表示向量在分类和链路预测实验中具有更好的性能表现。

关键词: 异质信息网络;网络表示学习;生成对抗网络;深度学习;元路径

中图法分类号 TP183

Generative Adversarial Network and Meta-path Based Heterogeneous Network Representation Learning

JIANG Zong-li, FAN Ke and ZHANG Jin-li

Department of Information Technology, Beijing University of Technology, Beijing 100124, China

Abstract Most of the information works in real world are heterogeneous information networks (HIN). Network representation methods aiming to represent node data in low dimensional space have been widely used to analyze heterogeneous information networks, so as to effectively integrate rich semantic information and structural information in heterogeneous networks. However, the existing heterogeneous networks representation methods usually use negative sampling to select nodes randomly from the network, and the heterogeneity learning ability of nodes and edges is insufficient. Inspired by the generative adversarial networks (GAN) and meta-path, we propose a new framework, which is improved by weighted meta-path based sampling strategy. The samples can better reflect the direct and indirect relationship between nodes and enhance the semantic association of samples. In the process of generation and confrontation, the model fully considers the heterogeneity of nodes and edges, and has the ability of relationship perception, so as to realize the representation learning of heterogeneous information networks. The experimental results show that, compared with the current representation algorithms, the representation vectors learned by the model have better performance in classification and link prediction experiments.

Keywords Heterogeneous information networks, Network representation learning, Generative adversarial network, Deep learning, Meta-path

1 引言

随着互联网的蓬勃发展,互联网技术也不断革新。现实生活中的众多实际应用场景,如社交网络、公路网络、学术网络以及通信网络等信息网络开始成为行业内研究的热点。

利用机器学习技术挖掘网络中的特征信息^[1],一个重要问题就是如何有效地表示网络数据。传统的信息网络表示方法通常构造高维稀疏矩阵,需要的计算和存储成本较高,并且缺乏灵活性^[2]。网络表示学习旨在学习一个映射函数,最大程度地保留信息网络的拓扑结构和属性信息,同时将高维稀

疏的网络表示成较低维度、特征稠密的向量,可直接作为网络特征输入到下游机器学习任务中,如用于判断两个人在社交网络中是否是朋友关系的链路预测任务和作为欧氏距离的点坐标进行可视化任务等^[3]。

同质信息网络是只含有一种节点和类型的信息网络,如图1所示。研究同质信息网络的表征方法主要有 DeepWalk^[4], node2vec^[5], LINE^[6]等,然而这些模型在应用到更为复杂的信息网络时,无法捕获到更多的语义信息。异质信息网络^[7]是一种特殊类型的信息网络,由多种类型的节点、链接关系以及额外的属性信息组成,具有大规模、异质性等特点。

以同质网络的研究方法作为出发点,许多学者尝试将同质网络表征的方法和异质网络的特性进行融合,如 Meta-path2vec^[8]等。

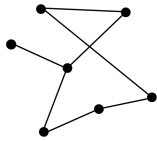


图 1 同质信息网络

Fig. 1 Homogeneous information network

受生成对抗网络^[9]的影响,逐渐出现了 GraphGan^[10]等算法,然而 GraphGan 只基于同质网络进行学习,现实生活中大多数的网络结构却是异质网络。

针对以上问题,本文提出了基于生成对抗网络和元路径的异质网络表示算法,该算法在采样的过程中,按照不同的比例分别对网络中直接相连和通过元路径间接相连的节点对进行抽样,提升间接关联节点在低维空间中的邻近程度,通过生成模型和判别模型的对抗学习异质网络的表征。本文主要的贡献可总结为以下 3 点:

(1)提出了一种新的异质网络表征学习方法 MHGan,将改进的生成式模型和判别式模型相结合,能够在异质信息网络中进行有效的学习,捕获异质网中丰富的语义信息。

(2)通过引入元路径改进抽样策略,使得样本的直接和间接关联度尽可能提高,同时生成器能够学习样本所在低维空间中的潜在分布。

(3)本文提出的方法在真实数据集的网络分析任务中取得了不错的结果,表明基于生成对抗和元路径的异质网络表示方法有着良好的表征学习能力。

2 相关研究

目前大多数网络表示方法研究的主要对象是同质信息网络。例如,受自然语言处理中词嵌入算法 Word2vec^[11-12]的启发,Perozzi 等^[4]率先提出 DeepWalk 算法,将自然语言处理的思想与网络表征巧妙结合,通过随机游走算法遍历网络中的节点,得到带有上下文环境的节点序列,模拟自然语言的状态,作为 skip-gram 计算模型的输入,最后得到节点的低维向量表示。以 DeepWalk 为基础,Grover 等^[5]提出的 node2vec 模型改进了随机游走的策略,通过对游走的深度和广度进行约束,进行带有偏置的游走,最后同样用 Word2vec 方法进行网络表征学习。与 DeepWalk 模型相比,node2vec 模型得到的表示向量保留了信息网络中更多的结构信息,在节点分类实验中的准确率有所提高。Tang 等^[6]提出 LINE 模型,其聚焦于网络节点的一阶相似度和二阶相似度,通过优化精心设计的目标函数,能够用于现实中大规模的信息网络表示。

以同质网络的研究方法作为出发点,许多学者尝试将同质网络表征的方法和异质网络的特性进行融合。例如,Dong 等^[8]提出的 Metapath2vec 模型采用基于元路径进行随机游走的策略,用于解决随机游走偏向于选择高可见类型节点的问题,然后用异质扩展的 skip-gram 处理游走获得的节点邻域,得到异质节点的向量表示。该方法能够有效地保留结构

和语义相互关系,但过渡依赖于元路径的选择,且元路径的选择对结果的影响较大,不具有普适性。

Goodfellow 等^[9]受博弈论中的二人零和博弈的启发,开创性地提出了生成式对抗网络(Generative Adversarial Network)。研究信息网络表征的生成式对抗网络主要有 GraphGan^[10]等。GraphGan 在对抗学习过程中,生成器尽可能地拟合节点 v_c 在网络中其他所有节点上的连续分布 $P_{\text{true}}(v|v_c)$,并且生成最可能与节点 v_c 连接的节点;判别器判别区分生成节点对和真实节点对,其输出为边 (v, v_c) 存在的概率,通过极大极小的博弈,最终拟合网络节点的真实分布,得到节点的向量表示。该方法研究的对象虽然是同质网络,但在忽略异质网络中节点的额外属性的情况下,也可用于异质网络的表征。

3 基于生成对抗网络和元路径的异质网络表示学习

3.1 符号定义

为了便于后文对算法进行详细的阐述,我们将本文用到的符号进行简单的解释,如表 1 所列。

表 1 符号定义
Table 1 Symbol definition

Symbol	Definition
$\mathcal{H}$	异质信息网络 HIN
V, E	节点和边的集合
A, R	节点和边类型的集合
ϕ, φ	节点类型和边类型的映射函数
T	网络模式
ρ, P	元路径和元路径集合
w, W	元路径权重和权重集合
l, d	元路径的长度和表示向量的维度
D, G	判别器和生成器
e^D, e^G	判别器和生成器的表示向量
M^D, M^G	判别器和生成器的关系矩阵
θ^D, θ^G	判别器和生成器的参数

3.2 问题定义

定义 1(异质信息网络, HIN^[7]) 假设网络 $\mathcal{H}=(V, E, A, R, \phi, \varphi)$,其中 V 和 E 分别为网络中节点和边的集合, A 和 R 分别是节点类型和边类型的集合。节点类型映射函数 $\phi:V \rightarrow A$,对于 $\forall v \in V$,都有 $\varphi(v) \in A$;边类型映射函数 $\varphi:E \rightarrow R$,对于 $\forall e \in E$,都有 $\varphi(e) \in R$ 。当节点的类型数目 $|A|>1$ 或者边的类型数目 $|R|>1$ 时,网络 $\mathcal{H}$ 为异质信息网络。

如图 2(a)所示的异质信息网络 HIN,有 User, Business, Star 3 种节点类型,节点之间的有向关系分别为 $User \rightarrow Business, Business \rightarrow User, Business \rightarrow Star$ 和 $Star \rightarrow Business$ 4 种类型,分别表示顾客去过商店、商店被顾客光顾、商店的评分和被评分的商店。

定义 2(网络表示学习) 对于信息网络 $\mathcal{H}=(V, E, A, R, \phi, \varphi)$,网络表示学习的目的是学习一个映射函数,能够将任意 $v \in V$ 映射为低维空间 $\mathbb{R}^d$ 中的一个向量,其中 $d \ll |V|$,使得此向量能够同时保存对应异质网中节点的结构和语义信息。

定义 3(元路径) 对于网络模式^[13] $T_{\mathcal{H}}=(A, R)$,其元路径为网络中两个链接对象之间的路径: $P=A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots A_l \xrightarrow{R_l} A_{l+1}$ 。节点对象 A_i 和 A_{i+1} 之间长度为 l 的复合关系

可形式化为 $R = R_1 \otimes R_2 \otimes \dots \otimes R_l$, 其中 l 表示复合关系的长度, $\otimes$ 表示关系之间的算子。

定义 4 (生成式对抗网络, GAN^[9]) Goodfellow 等开创性地提出了生成式对抗网络 (GAN)。生成式对抗网络包含一个生成模型 (generative model) 和一个判别模型 (discriminative model)。其中, 生成模型尽可能地生成和真实数据相似的假样本, 负责捕获样本数据的真实分布, 而判别模型一般情况下是一个二分类器, 判别输入的是真实数据还是生成的样本。训练时固定其中一方 (判别网络或生成网络), 更新另一个模型的参数, 交替迭代, 该模型的优化过程可形式化为式 (1) 的“极大极小博弈” (min-max) 问题。

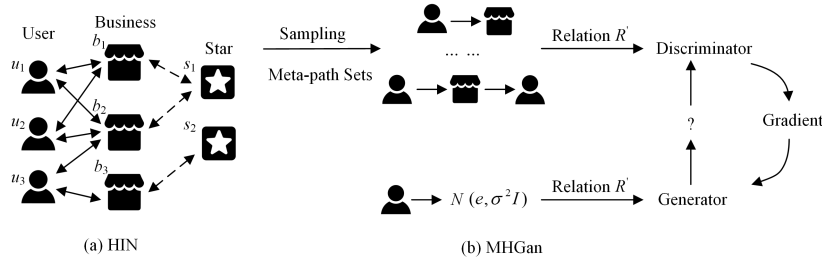


图 2 异质信息网络 HIN 和 MHGan 模型

Fig. 2 Heterogeneous information network HIN and MHGan model

步骤 1 对异质信息网络 $\mathcal{H}$ 定义其网络模式 $T = (A, R)$, 根据实际情况, 基于其网络模式定义元路径 p_i , 构建元路径集合 $P = \{p_1, p_2, \dots, p_n\}$, 并且构建元路径集合对应的权重集合 $W = \{w_1, w_2, \dots, w_n\}$, 使得对于每一个 $p_i \in P$, 都有 $w_i \in W$ 与之对应。

步骤 2 在训练判别器的过程中, 通常采用随机采样的方式, 在网络中对结构上直接相连的节点对进行采样作为正样本, 即通过单元路径进行随机采样, 如可用图 2(a) 中的单元路径 $User \rightarrow Business$ 抽样得到节点对 $u_1 \rightarrow b_1$, 该样本体现了网络节点的直接相关性, 但是忽略了不直接相邻的节点之间的间接相关性。如节点 u_1 和 u_2 之间虽然没有网络结构上的直接关联, 但通过元路径 $User \rightarrow Business \rightarrow User$ 可以抽样得到 $u_1 \rightarrow b_1 \rightarrow u_2$, 表示 u_1 去过被 u_2 光顾的同一家商店。由图 2(a) 可看出, $u_1 \rightarrow \dots \rightarrow u_2$ 之间存在多条不同的元路径, 包含不同的语义信息, 如 $User \rightarrow Business \rightarrow Star \rightarrow Business \rightarrow User$ 表示一位顾客去过的商店的评分和被另一位顾客光顾过的商店的评分一致, 即两位顾客去过同一级评分的商店。研究表明不同的元路径对异质信息网络表示的重要程度不同^[14-15], 因此在抽样过程中可将元路径的权重作为路径抽样的概率进行抽样, 而不是对单元路径进行简单随机抽样。

步骤 3 构建判别器 D 。给定两个节点 $u \in V, v$ (可能为真实节点或者假节点) 和一条特定的元路径 p , 判别器评估两个节点之间存在此元路径的概率, 即 $D(v|u, p; \theta^D)$, 其中 θ^D 为判别器 D 中的模型参数。我们将此概率量化为:

$$D(v|u, p; \theta^D) = \frac{1}{1 + \exp(-\mathbf{e}_u^D \mathbf{M}_p^D \mathbf{e}_v^D)} \quad (2)$$

其中, $\mathbf{e}_u^D \in \mathbb{R}^{d \times 1}$ 表示生成器中节点 u 的可学习的表征向量; $\mathbf{M}_p^D \in \mathbb{R}^{d \times d}$ 表示组成元路径的单元路径矩阵之间的顺序乘积, 即 $\mathbf{M}_p^D = \prod_{i=1}^j \mathbf{M}_{p_i}^D$ ($\mathbf{M}_{p_i}^D \in \mathbb{R}^{d \times d}$, j 表示 p 中单元路径的个

$$\min_{\theta^G} \max_{\theta^D} \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [1 - \log D(G(z))] \quad (1)$$

3.3 MHGan

在生成对抗网络的基础上, 本文针对异质信息网络提出一种新的表示学习模型 MHGan, 其整体设计如图 2(b) 所示, 主要分为两个部分, 分别为生成器 (generator) 和判别器 (discriminator)。对于一个给定的节点, 生成器尽可能地生成与此节点直接相关或者间接相关的节点作为判别器的输入, 而判别器则尽可能地区分与此节点存在相关性的真实节点和通过生成器生成的假节点。

数); $\mathbf{e}_v \in \mathbb{R}^{d \times 1}$ 表示输入样本的表示向量; θ^D 表示判别器 D 学习的所有节点向量和关系矩阵的总和。当节点 u 和 v 关于元路径 p 的关联性越强时, D 输出的值越大; 反之关联性越低, 输出的值越小。受条件式生成对抗网络的启发^[16], 以下 3 种情况均构成判别器损失的一部分。

情况 1 通过元路径相连。即在异质信息网络中的节点 u 和 v 通过元路径关系 p 真实连接, 如图 2(a) 中通过单元路径真实连接的关系 $u_1 \rightarrow b_1$ 表示用户 u_1 去过商店 b_1 , 又如关系 $u_1 \rightarrow b_1 \rightarrow s_1 \rightarrow b_2 \rightarrow u_2$ 表示顾客 u_1 和顾客 u_2 去过同一级评分为 s_1 的商店。此种情况的样本为从异质信息网络 $\mathcal{H}$ 中直接按照元路径抽样的正样本, 损失函数可表示为:

$$L_1 = \mathbb{E}_{\langle u, v, p \rangle \sim \mathcal{H}} -w \log D(v|u, p) \quad (3)$$

其中, $\langle u, v, p \rangle \sim \mathcal{H}$ 表示从异质网络 $\mathcal{H}$ 中提取的正样本, 节点 u 和 v 通过元路径 p 相连; w 是当前抽样元路径对应的权重。

情况 2 不通过元路径相连。即在异质信息网络中抽样的节点 u 和 v 对应的元路径不是 p , 而是 p' , 如图 2(a) 中的抽样关系 $u_1 \rightarrow u_2$, 由于节点 u_1 和 u_2 之间不存在直接连接性, 所以判别器倾向于将此种关系作为负样本。关于这种情况, 定义损失函数如下:

$$L_2 = \mathbb{E}_{\langle u, v \rangle \sim \mathcal{H}} -w' \log(1 - D(v|u, p')) \quad (4)$$

其中, $\langle u, v \rangle \sim \mathcal{H}$ 表示从异质网络 $\mathcal{H}$ 中抽样的节点 u 和节点 v , 且 u 和 v 之间存在的元路径为 p ; p' 从 $P' = P / \{p\}$ 提取得到, 其对应的权重为 w' 。

情况 3 通过生成器生成的假节点。给定节点 $u \in V$, 通过生成器 $G(u, p; \theta^G)$ 提供的节点 v 组成假的节点对, 如生成与图 2 的 HIN 类似的关系 $u_1 \rightarrow b_1 \rightarrow u'$ 。生成器可以根据已知的元路径关系尝试生成新的节点, 使得新节点能够尽可能接近通过正确的元路径 p 和节点 u 相连的真实节点。然而, 判别器倾向于将此种关系判断为负样本, 其损失函数如下:

$$L_3 = E_{(u,p) \sim \mathcal{H}} - \log(1 - D(v' | u, p)) \quad (5)$$

假样本 v' 是通过生成器 G 学习的分布直接生成的,因此式(5)中没有权重系数。

将以上 3 个部分集合起来训练判别器,如式(6)所示:

$$L^D = L_1 + L_2 + L_3 + \lambda^D \|\theta^D\|_2^2 \quad (6)$$

其中,正则项 $\lambda^D > 0$ 用于避免模型过拟合。通过最小化 L^D 来优化判别器的参数 θ^D 和元路径权重 W 。

步骤 4 构建生成器 G 。给定一个节点 $u \in V$ 和一条元路径 $p \in P$,生成器 $G(u, p; \theta^G)$ 倾向于生成一个假节点 v ,使其在元路径关系 p 的上下文中可能与节点 u 相连,也就是与真实的邻居节点相似,因此是潜在的,不一定是 V 中某一个真实的节点。为了满足此条件,生成器需要利用已知节点向量和元路径关系的矩阵,并且从连续的分布中生成相似度较高的样本,因此利用式(7)的高斯分布:

$$N(\mathbf{e}_u^G \mathbf{M}_p^G, \sigma^2 \mathbf{I}) \quad (7)$$

其中, $\mathbf{e}_u^G \in \mathbb{R}^{d \times 1}$ 为判别器中节点 $u \in V$ 的向量表示; $\mathbf{M}_p^G \in \mathbb{R}^{d \times d}$ 表示判别器中组成元路径 p 的单元路径矩阵之间的顺序乘积,即 $\mathbf{M}_p^G = \prod_{i=1}^j \mathbf{M}_{r_i}^G$ ($\mathbf{M}_{r_i}^G \in \mathbb{R}^{d \times d}$, j 表示 p 中单元路径的个数)。此连续分布也就是一个均值为 $\mathbf{e}_u^G \mathbf{M}_p^G$ 和协方差为 $\sigma^2 \mathbf{I} \in \mathbb{R}^{d \times d}$ 的高斯分布。直观上,均值表示通过关系 p 与目标节点 u 可能相连的假节点,协方差表示潜在的偏离程度。为了增强生成节点的能力,我们将多层感知机 MLP 集成到生成器中,可形式化为:

$$G(\cdot; \theta^G) = f(\mathbf{W}_n \cdots f(\mathbf{W}_1 \mathbf{e} + \mathbf{b}_1) + \mathbf{b}_n) \quad (8)$$

其中, $\mathbf{e}$ 是从式(7)中抽样所得, $\mathbf{W}$ 和 $\mathbf{b}$ 分别表示权重矩阵和偏置向量, f 表示激活函数。由于生成器想要尽可能骗过判别器,因此在训练生成器的阶段,判别器应该给生成器生成的样本更高的得分,如式(9)所示:

$$L^G = E_{(u,p) \sim \mathcal{H}} - \log D(v' | u, p) + \lambda^G \|\theta^G\|_2^2 \quad (9)$$

其中, v' 是生成器生成的样本,正则项 $\lambda^G > 0$ 用于避免模型过拟合。通过最小化 L^G 来优化生成器的参数 θ^G 。

MHGan 算法如算法 1 所示。

算法 1 MHGan 算法

输入: 异质信息网络 $\mathcal{H}$, 元路径集合 P , 判别器和生成器每次迭代的次数 n_D 和 n_G , 样本的个数 n_s , 训练的周期数 epoch

输出: 节点的向量表示

初始化元路径权重 W 、判别器和生成器的参数 θ^D 和 θ^G

1. While $i \leftarrow 0$ to epoch do
2. # 训练判别器
3. For $i \leftarrow 0$ to n_D do
4. 从 $\mathcal{H}$ 中抽样三元组 $\langle u, v, p \rangle \sim \mathcal{H}$
5. 对每个 $\langle u, v \rangle$ 抽样 n_s 个元路径 p'
6. 对每个 $\langle u, p \rangle$ 生成 n_s 个假样本 v'
7. 通过式(6)向后传播并更新参数 θ^D 和 W
8. End for
9. # 训练判别器
10. For $i \leftarrow 0$ to n_G do
11. 从 $\mathcal{H}$ 中抽样三元组 $\langle u, v, p \rangle \sim \mathcal{H}$
12. 对每个 $\langle u, p \rangle$ 生成 n_s 个假样本 v'
13. 通过式(9)向后传播并更新参数 θ^G
14. End for
15. End While
16. Output: θ^D 和 θ^G

4 复杂度分析

在 MHGan 迭代过程中,抽样阶段需保存节点的邻居信息,空间复杂度为 $O(|V|)$, $|V|$ 表示节点的个数;生成对抗网络需保存节点向量和关系矩阵,空间复杂度为 $O(|V| \cdot d) + O(d^2)$, d 表示表征向量的维度。抽样阶段由于无须遍历节点的所有邻居,只需要按照元路径选择其中一个邻居节点,因此时间复杂度为 $O(|V|)$;生成器和判别器在每次迭代过程中的更新主要与节点向量和关系矩阵的更新有关,其时间复杂度为 $O(n_s \cdot |V| \cdot d^2)$, n_s 表示样本的个数。因此模型每次迭代时的复杂度为 $O((n_G + n_D) \cdot n_s \cdot |V| \cdot d^2)$, n_G 和 n_D 分别表示每次迭代时生成器和判别器训练的次数。由于参数 d, n_s, n_G 和 n_D 都是常量,因此 MHGan 模型每次迭代时的复杂度和异质网络中的节点数量 $|V|$ 呈线性关系,同时也避免了传统表征方法的 softmax 算法,可用于大规模异质网络表征任务中。

5 实验

5.1 实验数据集

本文采用 DBLP 和 YELP 两个真实数据集, DBLP 数据集为学术会议网络, YELP 数据集为商业网络, 它们都是典型的异质信息网络。

DBLP 是计算机领域以作者为核心对研究成果进行分类的数据集。本文实验从 DBLP 数据集中提取了作者(author)、论文(paper)、会议(conference)和术语(term) 4 种节点类型, 总计 37791 个节点和 170794 条边。

YELP 是知名点评网站 YELP 公开的内部数据集。本文实验从 YELP 数据集中提取了用户(user)、商家(business)、服务(service)、星级(star)和预定(reservation) 5 种节点类型, 总计 3913 个节点和 38680 条边。

5.2 评价指标

实验采用准确率(Accuracy)、Micro-F1 分数、Macro-F1 分数和 AUC 指标来评价实验结果的有效性。

准确率(Accuracy)指对于给定的测试数据集, 分类正确的样本数与总样本数的比率。精确率(Precision)表示预测为正的样本中正确的比例。召回率(recall)表示样本的正例中被预测正确的比例。F1 分数(包括 Micro-F1 分数和 Macro-F1 分数)则表示精确率和召回率的调和平均数。以上指标的值越大表示分类的结果越好, 通过模型计算出的低维表征向量就越精确有效。AUC 指标是 ROC 曲线下的面积, 可从整体上衡量算法的准确性。链路预测算法在经过训练后可以得到网络中每一对节点的相似值(即边的相似值)。AUC 指标即基于测试集中边的相似值和不存在的边的相似值的比较(即以不存在的边作为基准), AUC 值越接近于 1, 则链路预测算法的表现越好。

5.3 实验设置

在实验中, 本文选取了被广泛采用并被证明有效的 Deepwalk, LINE, GraphGan, Metapath2vec 算法作为对比。

Deepwalk 受自然语言处理中 Word2vec^[11-12] 的思想启发, 使用图网络中节点和节点间的共现关系学习节点的向量

表示。通过 DFS 深度随机游走(RandomWalk)的方式在图中进行采样,获取足够数量的样本序列后用 skip-gram 模型学习网络节点的向量表示。

LINE^[6]也是一种基于邻域相似假设的方法,但与 DeepWalk 使用 DFS 构造邻域不同的是,LINE 可以看作一种使用 BFS 构造邻域的算法,对图中节点之间的相似度采用一阶和二阶相似度的方式进行定义。

GraphGan^[10]使用生成对抗网络来增强网络表示的学习能力,同时融合了生成模型和判别模型,并成功应用在同质信息网络的表示任务中。

Metapath2vec^[8]算法与 Deepwalk 类似,不同的是,其采用基于元路径的随机游走方式获取每个节点的异质邻域,然后用 skip-gram 模型完成节点的表示学习。

在实验过程中,MHGan 模型的参数设置如下:元路径的

复合关系长度阈值 l 为 4,即元路径中节点和边的最大数量为 5 和 4,并且只选择路径对称的结构,如图 2(a)的元路径 UBSBU;对于判别器和生成器在每个周期内迭代的次数 n_D 和 n_G ,分别取值于数组 $\{1,5,10,15,20\}$,通过梯度搜索得知, $n_D=15$ 和 $n_G=5$ 时为最优组合;每次抽样的样本个数 n_s 为 16;总迭代周期 epoch 为 32;网络节点的表征向量的维度 d 为 128;判别器和生成器的学习率均为 0.0001;生成器中高斯分布的方差 σ^2 为 1。

5.4 分类实验

在 DBLP 和 YELP 数据集上,使用完整数据集学习节点的向量表示,随后从有标签的节点中随机选择 10%~90% 的节点作为训练集,剩余部分节点作为测试集,使用逻辑回归作为分类器,实验的 Accuracy, Micro-F1 和 Macro-F1 结果如图 3 和图 4 所示。

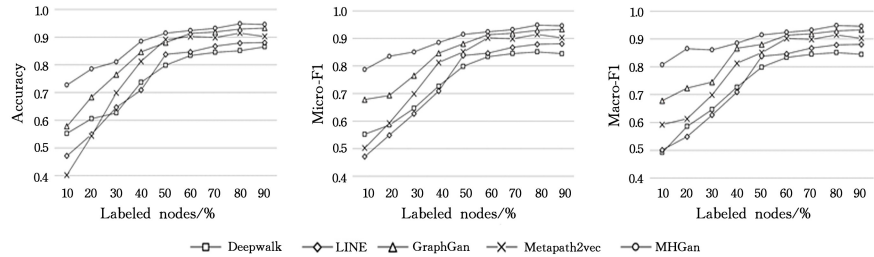


图 3 DBLP 节点分类结果
Fig. 3 Nodes classification result on DBLP

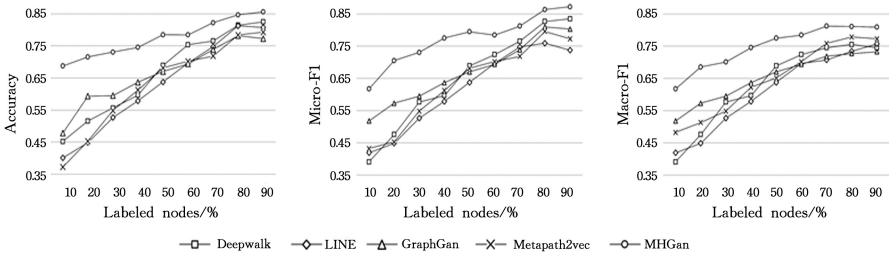


图 4 YELP 节点分类结果
Fig. 4 Nodes classification result on YELP

从图中可以看出,MHGan 算法在 DBLP 和 YELP 数据集上的向量表示在用于分类实验时优于传统的算法。LINE 由于计算节点的一阶相似度和二阶相似度,虽然可用于大规模的信息网络表示,但是表示向量的效果不够好。Metapath2vec 加入了元路径来捕获信息网络中的语义信息,在节点较多的 DBLP 数据上效果良好,但是在节点较少的 YELP 数据集上的性能较差。这是因为 Metapath2vec 算法依赖元路径的选择并且受网络复杂程度的影响,导致其性能不够稳定。GraphGan 由于针对的是同质网络,在处理异质网络时不能体现其中的异质性,表示学习的能力要弱于本文提出的引入元路径的 MHGan 算法。在 DBLP 数据集上,MHGan 的 Accuracy, Micro-F1 和 Macro-F1 指标分别有 2%~25.8%, 1.4%~16.1% 和 2.1%~19% 的提升;在 YELP 数据集上, MHGan 的 Accuracy, Micro-F1 和 Macro-F1 指标分别有 3.7%~34.8%, 4.4%~19.1% 和 4.1%~17.6% 的提升,从而验证了本文提出的算法能够更好地学习网络节点的特征信息。

在 YELP 数据集的分类实验中,经过多次迭代,最终构建出元路径集合对应的权重集合,以元路径 UBU 和组成其路径的单元路径为例,其对应的权重如表 2 所列。

表 2 UBU 路径权重
Table 2 UBU path weight

Path	Weight Range	Average Weight
UB	0.03~0.06	0.04
BU	0.05~0.08	0.06
UBU	0.15~0.22	0.17

其中单元路径 UB 和 BU 的语义信息对异质网络表征学习的重要程度相当,但元路径 UBU 表达的语义信息对表征学习的影响远大于路径 UB 和 BU 所表达的语义信息,这表明 MHGan 学习直接相关节点的表征时,更倾向于捕获间接相关节点之间的潜在联系。

5.5 链路预测实验

链路预测实验即预测网络中两个节点之间是否有相连的边。在本次任务中,我们预测 DBLP 数据集的 author-paper 边和 YELP 数据集的 user-business 边。实验过程中,我们

随机隐藏网络中 20% 的边作为测试集,剩下的节点和边用来训练网络表征模型,最后用学到的表征向量对测试集用内积的方式进行链路预测,采用 Accuracy 和 AUC 作为实验结果的指标。实验结果如表 3、表 4 所列。

表 3 DBLP 链路预测结果
Table 3 Link prediction results on DBLP

Methods	Accuracy	AUC
Deepwalk	0.547 4	0.723 1
LINE	0.664 7	0.775 3
GraphGan	0.553 2	0.682 5
Metapath2vec	0.477 5	0.692 6
MHGan	0.746 1	0.831 4

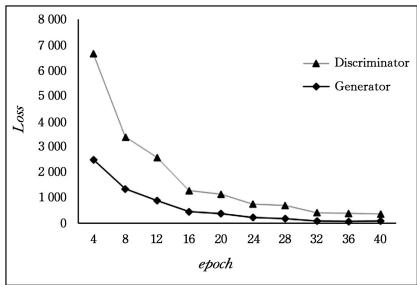
表 4 YELP 链路预测结果
Table 4 Link prediction results on YELP

Methods	Accuracy	AUC
Deepwalk	0.565 4	0.816 4
LINE	0.676 9	0.783 2
GraphGan	0.570 2	0.772 5
Metapath2vec	0.512 4	0.632 4
MHGan	0.735 8	0.823 0

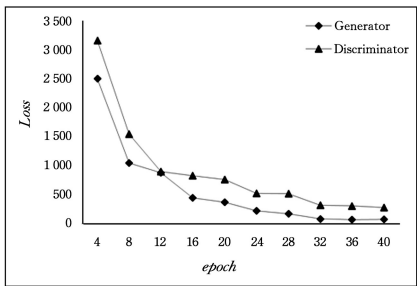
通过表 3 和表 4 可看出,在对 DBLP 和 YELP 数据集进行链路预测任务时,MHGan 相比其他基准算法都有更好的表现,并且结合了判别式模型的方法如 LINE 和 GraphGan 在多数情况下都优于生成式的模型,如 Deepwalk 和 Meta-path2vec。这说明在链路预测任务中,判别式模型相比生成式模型具有更好的表现。

5.6 参数敏感性

MHGan 算法中实验参数对表示学习的结果有着重要的影响,因此需要对敏感的参数进行分析。本文在保证数据集和其他参数一致的情况下,从算法迭代周期和表示向量的维度两个方面进行对比,结果如图 5 和图 6 所示。



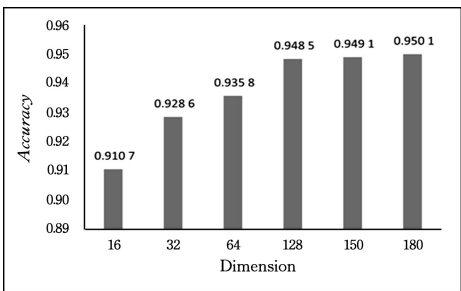
(a)DBLP-epoch 与 Loss



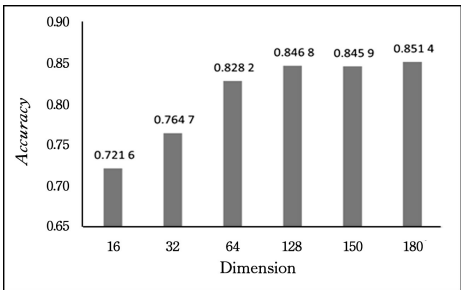
(b)YELP-epoch 与 Loss

图 5 算法中 epoch 和 Loss 的变化

Fig. 5 Change of epoch and Loss of algorithm



(a)DBLP-Dimension 与 Accuracy



(b)YELP-Dimension 与 Accuracy

图 6 算法的 Dimension 和 Accuracy 的变化

Fig. 6 Change of Dimension and Accuracy of algorithm

通过图 5 可以看出,随着迭代周期 $epoch$ 的增加,判别器和生成器在对抗的过程中损失逐渐减少并逐渐趋于稳定,但是随着周期的增长并且 $epoch$ 大于 32 后,模型在博弈的过程中已经接近平衡,没有较大的学习空间。考虑到算法的训练时间在实际应用中也较为重要,因此在算法的参数设置中 $epoch$ 取值为 32。图 6 中由于数据集 DBLP 和 YELP 节点的稠密程度不同,在节点较多的 DBLP 数据集上的实验证实了,节点维度设置较低时无法保存复杂网络中足够的网状结构关系和节点之间的语义联系,导致 MHGan 在分类任务中的准确度和较高维度时相比相差较多。而在结构较为简单的 YELP 上,较低的维度就能容纳可观的信息,使得分类准确度和较高维度时的准确度相差不多。并且从图 6 也可以看出,在达到一定的维度之后,算法学到的表示向量在分类任务中的提升不再明显,反而较高的维度使算法的复杂程度增加,需要的训练时间也相应增加,有过拟合的风险,并且会降低下游机器学习任务的处理效率,因此我们选择 128 维作为表示向量的维度。

结束语 本文提出了一种基于生成对抗网络和元路径的异质网络表示算法。通过在抽样阶段引入元路径,根据原路径的比重对信息网络中存在直接和间接联系的节点对进行样本提取,通过生成对抗网络中生成器和判别器的相互博弈进行网络节点的表示学习,尽可能地利用网络中的结构和语义信息。在真实数据集上的实验结果表明,本文提出的算法使学习到的表示向量在分类实验中的各项指标均有所提升。但本文算法所针对的异质网络均为静态网络,在训练的过程中不会发生改变。而在现实生活中,网络中节点的数量和边的存在性在不断地变化,因此如何实时对动态网络^[17-18]进行高效的表示学习是下一步的研究重点。

参 考 文 献

[1] TU C C, YANG C, LIU Z Y, et al. Network representation learning: an overview[J]. Sci. Sin. Inform., 2017, 47(8): 980-996.

[2] QI J S, LIANG X, LI Z Y, et al. Representation Learning of Large-Scale Complex Information Network: Concepts, Methods and Challenges [J]. Chinese Journal of Computers, 2018, 41(10): 2394-2420.

[3] ARSOV N, MIRCEVA G. Network Embedding: An Overview [J]. arXiv:1911.11726, 2019.

[4] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: Online learning of social representations [C] // Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014: 701-710.

[5] GROVER A, LESKOVEC J. node2vec: Scalable feature learning for networks [C] // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 855-864.

[6] TANG J, QU M, WANG M, et al. Line: Large-scale information network embedding [C] // Proceedings of the 24th International Conference on World Wide Web. 2015: 1067-1077.

[7] ZHOU H, ZHAO Z Y, LI C. Survey on representation learning methods oriented to heterogeneous information network [J]. Journal of Frontiers of Computer Science and Technology, 2019, 13(7): 1081-1093.

[8] DONG Y, CHAWLA N V, SWQMI A. metapath2vec: Scalable representation learning for heterogeneous networks [C] // Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 135-144.

[9] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C] // Advances in Neural Information Processing Systems. 2014: 2672-2680.

[10] WANG H, WANG J, WANG J, et al. Graphgan: Graph representation learning with generative adversarial nets [C] // Thirty-second AAAI Conference on Artificial Intelligence. 2018.

[11] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representation of words and phrases and their composition [C] // Advances in Neural Information Processing Systems. 2013: 3111-3119.

[12] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estima-

tion of word representations in vector space [J]. arXiv: 1301.3781, 2013.

[13] SHI C, LI Y, ZHANG J, et al. A survey of heterogeneous information network analysis [J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 29(1): 17-37.

[14] ZHANG J L, JIANG Z L, LI T, CHIN: Classification with META-PATH in heterogeneous information networks [C] // Proceedings of the 1st International Conference on Applied Informatics. 2018: 63-74.

[15] GUPTA M, KUMAR P, BHASKER B. HeteClass: A meta-path based framework for transductive classification of objects in heterogeneous information networks [J]. Expert Systems with Applications, 2017, 68: 106-122.

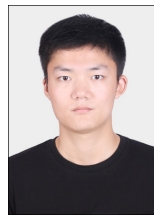
[16] SCOTT R, ZEYNEP A, YAN X C, et al. Generative adversarial text to image synthesis [C] // ICML. 2016: 1060-1069.

[17] BIAN R, YUN S, DOBBIE G, et al. Network Embedding and Change Modeling in Dynamic Heterogeneous Networks [C] // Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 861-864.

[18] WANG X, LU Y F, SHI C, et al. Dynamic Heterogeneous Information Network Embedding with Meta-path based Proximity [J/OL]. IEEE Transactions on Knowledge and Data Engineering. <http://yuanfulu.github.io/publication/TKDE-DyHNE.pdf>.



JIANG Zong-li, born in 1956, Ph.D., professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include network information search and processing.



FAN Ke, born in 1996, postgraduate. His main research interests include heterogeneous information network representation learning and so on.

(责任编辑:柯颖)