

基于不完全信息的深度网络表示学习方法



富 坤¹ 赵晓梦¹ 付紫桐² 高金辉¹ 马浩然³

1 河北工业大学人工智能与数据科学学院 天津 300401

2 长春理工大学计算机科学技术学院 长春 130022

3 华中科技大学武汉光电国家研究中心 武汉 430074

摘 要 网络表示学习的目标是将网络中的节点嵌入到低维的向量空间,为下游任务提供有效特征表示。在现实场景中,大规模网络通常具有不完整的链路,而现有的大多数网络表示学习模型都是在网络是完整的假设下设计的,因此其性能很容易受到链路缺失的影响。针对该问题,文中提出了一种基于不完全信息的深度网络表示学习方法 DNRL(Deep Network Representation Learning)。首先采用转移概率矩阵将结构信息和属性信息进行动态融合,弥补了结构信息不完整带来的过大损失,然后采用一种具有强大特征提取能力的深度生成模型(变分自编码器)来学习节点的低维表示,并捕获网络数据中潜在的高非线性特征。在 3 个真实属性网络上的实验结果表明,与当前常用的网络表示学习模型相比,所提模型在不同程度链路缺失的节点分类任务中都明显地改善了分类效果,在可视化任务中更清晰地反映了节点的团簇关系。

关键词: 网络表示学习;属性网络;不完全信息;变分自编码器

中图法分类号 TP391

Deep Network Representation Learning Method on Incomplete Information Networks

FU Kun¹, ZHAO Xiao-meng¹, FU Zi-tong², GAO Jin-hui¹ and MA Hao-ran³

1 School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China

2 School of Computer Sciences and Technology, Changchun University of Science and Technology, Changchun 130022, China

3 Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430074, China

Abstract The goal of network representation learning(NRL) is embedding network nodes into low-dimensional vector space, for effective feature representation of the downstream tasks. Due to the difficulty of information collection in the real-world scenarios, large-scale networks often meet missing links between nodes. However, the most existing NRL models are designed on the foundation of complete information networks and that causes the poor robustness in incomplete networks. To solve this problem, a deep network representation learning(DNRL) method based on incomplete information networks is proposed. Firstly, a transfer probability matrix is used to dynamically mix the structural information and attribute information to cover the excessive loss caused by incomplete structural information. Then, a deep generative model variational autoencoder with powerful feature extraction capability is used to learn low-dimensional representation of nodes, and capture the potential high nonlinear features of nodes. Compared with the commonly used network representation learning methods, the experimental results on three real attribute networks show that the proposed model obviously improve effect in the node classification task with different degrees of link missing, visualization results clearly demonstrate the cluster relationship of nodes.

Keywords Network representation learning, Attribute network, Incomplete information, Variational autoencoder

1 引言

网络表示学习(Network Representation Learning, NRL)^[1-3]作为一种分析网络的基本工具,近年来引起了数据挖掘和机器学习领域的广泛关注。其主要目的在于寻找一个合适的映射函数,将大规模、高维、稀疏的向量映射到低维、稠密的向量空间中,同时保持原有网络节点的相似性。之后将网络表示

学习到的低维向量应用于下游的节点分类^[4]、可视化^[5]、链路预测^[6]、社区发现^[7]等网络分析任务中,这对解决现实网络中的实际应用问题具有重要意义^[8]。

目前,主流的网络表示学习方法主要分为基于拓扑结构的方法和基于拓扑结构与属性相结合的方法。第一类方法可以分为以下 3 种:1)基于矩阵特征向量的方法,如局部线性表示^[9]、拉普拉斯特征映射^[10];2)基于随机游走的方法,如

到稿日期:2020-09-30 返修日期:2021-03-02 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:国家自然科学基金青年科学基金(61806072)

This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China(61806072).

通信作者:富坤(fukun@hebut.edu.cn)

DeepWalk^[11], Node2Vec^[12]; 3) 基于神经网络的方法, 如 DNGR^[13], SDNE^[14]。第二类方法主要有 TADW^[15], AANE^[16]和 SNE^[17]等, 该类方法在学习网络节点表示的同时考虑了结构信息和属性信息, 同第一类方法相比, 其学到的低维嵌入表示更能保持网络中节点的相似性。现有的属性网络嵌入方法虽然可以综合利用结构信息和属性信息, 但它们往往假设输入的网络结构信息是完整的。然而, 现实中采集到的网络结构信息通常是不完整的, 而且可能存在不同程度的链路缺失, 这在以往的文献中很少被讨论。

数据缺失是网络研究中一个非常重要但经常被忽略的问题, 测量的有效性会受到缺失数据的影响, 而且偏差水平难以衡量^[18]。我们将从以下两个方面进行探讨: 一方面, 随着数据保护意识的增强和反爬虫技术的发展, 对完整数据集的获取变得越来越困难; 另一方面, 许多真实世界的网络本身就具有不完整的结构信息, 如犯罪份子会在社交平台故意隐藏他们的友谊关系。这些因素使不完全信息在实际应用中广泛存在, 导致数据处理步骤产生误差, 进而使后续网络分析任务受到极大的负面影响。更重要的是, 现实网络可能存在不同程度的结构信息缺失, 属性网络的结构信息和属性信息又是高度非线性的, 这给属性网络表示学习带来了更多的挑战。

针对上述问题, 本文设计了一种半监督的深度网络表示学习模型 DNRL, 用于处理具有不同程度不完全结构信息的属性网络。具体来说, 首先提出一种结合不完全结构信息和属性信息的动态策略; 其次为充分利用已有的先验信息, 采用半监督的变分自编码器来捕获网络数据中潜在的高非线性特征, 学习属性网络节点的低维表示。

2 相关工作

网络表示学习可以追溯到图的表示学习, 传统的图表示学习是降维的过程, 如主成分分析(PCA)、拉普拉斯特征映射(Laplace Eigenmaps, LE)^[19]等。由于这些方法涉及耗时的特征分解操作, 因此其时间复杂度较高, 不适用于大规模的网络表示学习。近年来, 随着大规模网络的快速发展, 网络表示学习受到了越来越多的关注, 相关的研究也层出不穷。其中, DeepWalk^[11]将随机游走生成的节点序列当作句子, 输入到 skip-gram 模型中学习节点表示。Node2Vec^[11]在可伸缩的节点邻域上设计了一个有偏差的随机游走, 通过使用广度优先(BFS)和深度优先(DFS)来探索不同的节点邻域, 然后将得到的节点序列当作句子输入到 skip-gram 模型中学习节点表示。除了基于随机游走的方法, LINE^[19]模型还对节点的一阶相似性和二阶相似性进行了概率建模。GraRep^[20]对 LINE 进行扩展, 提出了一种获取 k 阶关系信息的表示方法。DNGR^[13]首先运用 Random Surfing 方法获取网络的高维节点表示, 然后使用去噪自编码器学习节点的低维表示。SDNE^[14]使用深层神经网络建模学习网络节点的非线性表示。然而, 这些方法都只利用了拓扑结构信息, 而忽略了节点可利用的属性信息。在拓扑结构信息不完整的情况下, 大多数节点的链接非常稀疏, 这些方法不能产生令人满意的结果。

属性信息网络在许多领域中是普遍存在的。在设计网络表示学习模型时, 融入节点属性信息有助于获得更好的节点表示。TADW^[15]提出了一种结合网络拓扑结构信息和文本

信息的归纳矩阵分解方法。AANE^[16]基于矩阵分解将网络中的拓扑结构信息和属性信息整合到一起。SNE^[17]利用端到端神经网络模型来捕获网络结构和节点属性信息之间复杂的相互关系, 从而生成节点嵌入。SEANO^[21]是一种半监督学习框架, 该模型通过深度学习架构将节点的标签信息、属性信息、拓扑结构信息关联起来。ONE^[22]通过正交变换将属性嵌入的结果映射到结构嵌入所在的特征空间中, 并考虑了异常节点对属性网络嵌入的影响。

上述网络表示学习方法均假设网络是完整的。当网络中存在链路缺失时, 它们的性能会急剧下降, 因为它们缺乏处理缺失数据的能力。因此, 出现了一些针对不完全信息网络的研究工作。SINE^[23]模型提出了一种概率学习框架, 分别对节点上下文和节点属性关系进行建模, 来学习不完全信息网络的节点表示。ABRW^[24]模型考虑了结构信息不完整的情况, 提出了一种基于偏置随机游走的属性网络嵌入方法。但是, 它们都是浅层模型, 很难捕获到网络数据的高度非线性。因此, 对结构信息不完整的属性网络嵌入方法进行更有效的探索是非常有必要的。

3 问题定义

为了更好地阐述本文模型及具体算法, 首先给出了相关问题的定义。

定义 1(属性网络) 给定一个属性网络 $G=(V, E, A)$, 其中, V 表示属性网络的节点集合; $|V|=n$ 表示节点的数目; $E=\{e_{ij} | 0 < i, j \leq n\}$ 表示属性网络节点间的链接集合; A_i 表示属性网络中第 i 个节点对应的描述属性信息, 通常可以保证有足够的属性信息可用。

定义 2(不完全结构) 邻接矩阵 $W \in R^{n \times n}$, $W_{ij}=1$ 表示属性网络中节点 v_i 和节点 v_j 之间存在链接, 否则 $W_{ij}=0$ 。本文研究节点结构信息不完整的情况, 即网络存在不同程度的链路缺失, 此时会产生一些孤立节点, 导致邻接矩阵非常稀疏, 不利于特征提取。

定义 3(半监督属性网络表示学习) 网络表示学习又称为网络嵌入。给定含有少量已知标签信息的属性网络, 同时利用节点的标签信息、属性信息和网络拓扑结构信息, 学习网络中节点的低维向量表示 $Z \in R^{n \times d}$ (d 是学习到表示的维度, 远小于 n)。

4 算法描述

属性网络表示学习要获得良好的嵌入效果会面临着以下挑战: 1) 如何将不同语义空间上的不完整结构信息和属性信息进行融合; 2) 如何对高非线性的结构信息和属性信息进行特征提取。针对上述问题, 本文利用已获得的先验信息, 设计了一种半监督的深度网络表示学习方法 DNRL。下文分别从数据编码、特征提取和目标函数这 3 部分进行详细介绍。

4.1 数据编码

4.1.1 PPMI 矩阵

本文提出的深度网络嵌入模型 DNRL 的主要任务是将高维的网络节点映射到低维的潜在空间。属性网络的节点高维向量表示我们沿用 DNGR 中的方法, 仍然使用 **PPMI**^[25]矩阵作为输入。本文中的 **PPMI** 矩阵是一个反映节点间相互作

用的密集矩阵,输入 **PPMI** 矩阵可以确保模型能够捕获更高阶的接近度。**PPMI** 矩阵的构造方法如下:

$$PPMI(w, c) = \max\left(\log_2 \frac{P(w, c)}{P(w)P(c)}, 0\right) \quad (1)$$

其中, P 表示一个共现概率矩阵, c 表示对应单词 w 的上下文。采用 Random Surfing 模型^[13], 使用一个转移矩阵 T , 通过 k 步迭代得到共现矩阵 P 。Random Surfing 考虑重启的情况, 也就是说有 $1-\beta$ 的概率会回到原来的初始节点。共现矩阵 P 的行向量迭代公式如下:

$$p_k = \beta \cdot p_{k-1} T + (1-\beta) \cdot p_0 \quad (2)$$

其中, 行向量 p_k 的第 j 个下标表示通过 k 步转移之后到达节点 j 的概率, p_0 是一个只有一项为 1 其他都为 0 的独热向量。

虽然利用 **PPMI** 矩阵表示网络结构信息的方法是有效的, 但还有两个重要问题需要解决。

(1) 忽略了真实网络结构信息不完整的情况, 当有孤立节点产生时, 只根据结构信息计算转移矩阵, 产生的误差会影响后续的网络分析任务。

(2) 除了结构信息, 真实世界网络中的节点通常拥有与之相关联的属性信息。如何选择合适的方法对不同语义空间中的结构信息和属性信息进行融合值得思考。

针对上述问题, 本文采用结构-属性联合转移矩阵将不完整的结构信息和节点属性信息结合在一起。

4.1.2 结构-属性联合转移矩阵

本节介绍结构-属性联合转移矩阵 T 的计算过程。由于结构信息的不完整性, 本节采用线性的方法将结构转移矩阵和属性信息转移矩阵结合在一起, 有效实现了两者之间的相互补充制约。

转移矩阵 T 中的每个元素都是非负实数, 行向量给出了离散型概率分布, 表示从当前节点转移到下一个节点的概率。

定义 4(结构转移矩阵) 给定属性网络的邻接矩阵 $W \in R^{n \times n}$, 基于结构的转移矩阵 $T^W \in R^{n \times n}$ 的计算方法如下:

$$T_{i,j}^W = \frac{W_{i,j}}{\sum_{j \in n} W_{i,j}} \quad (3)$$

定义 5(属性相似度矩阵) 给定属性网络的属性矩阵 $A \in R^{n \times m}$, 其中每一行 $A_i \in R^m$ 对应于节点 v_i 的属性信息, 通过余弦相似度计算每一对节点之间属性信息的相似性, 然后将其存储到 $n \times n$ 维的矩阵中, 属性相似度矩阵 $S \in R^{n \times n}$ 的计算方法如下:

$$S_{i,j}^A = \text{CosSim}(A) = \frac{A_i A_j^T}{|A_i| |A_j|} \quad (4)$$

为了降低后续计算的复杂度, 我们只保留每个节点和最相似的 l 个节点的余弦值, 即只保留矩阵 S 中行向量的前 l 个值, 其他值都置 0, 具体方法如下:

$$S_{i,j}^A = \begin{cases} S_{i,j}^A, & \text{if } S_{i,j}^A \in \{\text{top } l \text{ of } S_i^A\} \\ 0, & \text{else} \end{cases} \quad (5)$$

定义 6(属性转移矩阵) 属性转移矩阵的构造方法与结构转移矩阵一致, 基于属性的转移矩阵 $T^A \in R^{n \times n}$ 的计算方法如下:

$$T_{i,j}^A = \frac{S_{i,j}^A}{\sum_{j \in n} S_{i,j}^A} \quad (6)$$

定义 7(结构-属性联合转移矩阵) 结构-属性联合转移矩阵的目的是合并和平衡这两个信息源, 利用预先计算好的

T^W 和 T^A , 联合转移矩阵 T 的计算式如下:

$$T_i = \begin{cases} T_i^A, & \text{if } T_i^W \text{ is all zeros} \\ \alpha T_i^W + (1-\alpha) T_i^A, & \text{else} \end{cases} \quad (7)$$

其中, T_i , T_i^W 和 T_i^A 分别对应各个转移矩阵的第 i 行, 如果一个节点没有任何链接, 即该节点是一个孤立节点, 则行向量 T_i^W 将全部为 0, 我们将直接把属性信息分配给它进行补偿。对于其他情况, 我们将用超参数 α (大于 0 且小于 1) 来平衡结构和属性信息。 α 值越大, 属性信息对最终的节点低维表示的影响就越小。

4.2 特征提取

为充分利用网络中已有的信息来提升嵌入效果, 本文引入了少量的先验标签信息, 将数据集随机分为两个子集 XL 和 $XunL$, 子集 XL 中的每个样本都包含标签信息, 而子集 $XunL$ 中的样本没有标签信息。在得到属性网络的高维向量表示后, 需要对其进行降维操作, 获得节点的低维特征表示, DNRL 模型采用半监督的变分自编码器 (VAE) 对 **PPMI** 矩阵 X 进行特征提取, 从而获得了网络的低维特征向量表示矩阵 $Z \in R^{n \times d}$ 。DNRL 模型如图 1 所示。

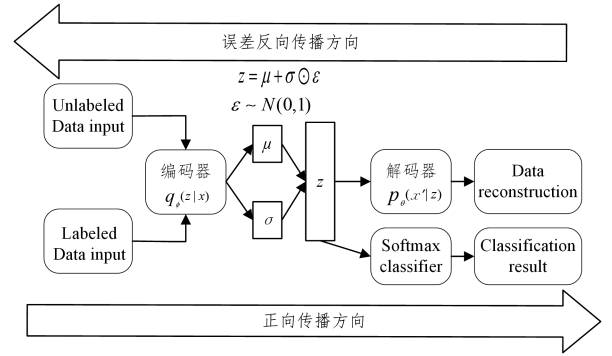


图 1 DNRL 模型的框架

Fig. 1 Framework of DNRL model

变分自编码器 (VAE)^[26] 是一种深度生成模型, 它将概率图模型和神经网络技术相结合, 拟合出原始数据服从的分布, 同时能够生成类似的数据, 是处理不完整数据的良好工具。VAE 由两个主要模块组成: 编码器和解码器。用 $q(z|x)$ 表示编码器网络, 用 $p(x'|z)$ 表示解码器网络, 它们的参数分别为 ϕ 和 θ 。不同于普通的自编码器, VAE 的编码器网络不是将输入 $x_i \in R^n$ 直接编码成低维的隐因子向量 $z_i \in R^d$, 而是输出用于采样隐因子向量 z 的均值和方差, 进而从具有该均值和方差的 Gauss 分布中生成 z 。最后将采样向量 z 作为解码器网络的输入, 生成近似于原始输入的输出 $x_i' \in R^n$ 。这样做的目的是通过假设隐因子的概率分布来降低每个隐因子的刚性误差。

在解码阶段, 除了学习重构网络的节点表示外, 还要根据可利用的先验信息重构先验节点的类标记。重构先验节点的类标记的目标是要训练一个分类器 $f: x \rightarrow y$, 使得每一个先验节点 x 的预测标签都尽可能地接近真实标签, y 为先验节点被预测的所属类别。最后, 联合优化变分自编码器的嵌入能力和分类器的预测效果, 采用误差反向传播机制, 学习网络节点的最优低维嵌入 Z 。

4.3 目标函数

DNRL 模型的优化目标是最大化变分下界函数和最小化

节点所属类别损失。其目标函数如式(8)所示:

$$L = -D_{KL}(q_\varphi(z|x) \parallel p_\theta(z)) + E_{q_\varphi(z|x)}[\log p_\theta(x|z)] + \gamma \sum_{i=1}^n y_i^T \text{lb } y_i' \quad (8)$$

其中,第一项散度函数 $D_{KL}(\cdot)$ 是一个正则化项,使后验分布 $q(z|x)$ 和先验分布 $p(z)$ 尽量接近,这里将 $p(z)$ 定义为标准正态分布 $N(0,1)$ 。第二项是重构损失, $p(x|z)$ 对应的是解码器部分,表示解码器使用基于高斯分布的随机采样获得潜在向量 z 来重构输入数据,重构的输出尽可能地接近输入。第三项是分类损失,其中 y_i 是节点 i 的真实标签, y_i' 是节点 i 的预测所属类别向量, n 为先验节点的个数, γ 是分类损失所占的比例系数。

为了简化计算,重构损失通常采用蒙特卡罗抽样^[27]方法进行近似求解。从分布 $q_\varphi(z|x)$ 中抽取 M 个样本,以样本的均值来近似期望 E_q ,此时重构损失可以转化为:

$$E_{q_\varphi(z|x)}[\log p_\theta(x|z)] = \frac{1}{M} \sum_{m=1}^M \log p_\theta(x|z^m) \quad (9)$$

为保证反向传播时模型可导,潜在向量通过重参数化技巧^[28]进行采样, $z = \mu + \epsilon \cdot \sigma$,即对于样本 z^m 来说,从分布 $N(0,1)$ 中采样变量 ϵ 对隐变量进行重参数化。此时,模型的损失函数可以转化为:

$$L = -D_{KL}(q_\varphi(z|x) \parallel p_\theta(z)) + \frac{1}{M} \sum_{m=1}^M \log p_\theta(x|z^m) + \gamma_1 \sum_{i=1}^k y_i^T \text{lb } y_i' \quad (10)$$

至此,为估计 DNRL 模型中的参数,使用 Adam 算法进行具体求解。

5 实验结果与分析

本节在真实网络数据集上与几种比较常用的方法进行比较,验证本文方法对具有不完整结构信息的属性网络的鲁棒性和有效性。

5.1 数据集与预处理

本次实验选取了 3 个真实属性网络数据集: Cora, Citeseer, Wiki。前两个是引文网络数据集, Cora 数据集包含 2708 份科学出版物,按出版物分为 7 类; Citeseer 数据集包含 3312 份科学出版物,分为 6 类。边表示文献之间的引用关系,每个节点的属性信息是对应论文的 bag-of-words 表示。Wiki 是一个网页网络,按网页类别分为 17 类。每一个网页表示一个节点,节点之间的边是连接到其他网页的超链接,将每个网页的文本内容作为该网页的属性信息。各个数据集的详细信息如表 1 所列。

表 1 真实网络数据集

Table 1 Datasets for real networks

数据集	节点数	边数	属性数	类别
Cora	2708	5278	1443	7
Citeseer	3312	4660	3703	6
Wiki	2405	12761	4973	17

在数据预处理中,我们需要对数据进行模拟,使其基于真实网络数据集,具有不同程度的不完整结构信息。具体而言,随机丢弃不同比例的链接,仅使用保留的链接作为结构信息,联合属性信息进行输入。注意,此时将会产生一些孤立的节点,这与之前的工作有所不同。

5.2 比较方法

我们将本文提出的 DNRL 与几种最常用的网络表示学习模型进行比较,这些模型可以分为以下几组。

(1) 纯属性。第一组模型只考虑节点的属性信息,将属性向量转化为低维表示,用于验证属性信息在节点分类任务中的有效性。我们选择 SVD 作为我们的基线模型。

(2) 纯结构。这组模型仅利用网络结构信息,而忽略了节点属性信息。DeepWalk 和 Node2Vec 使用随机游走生成节点序列,然后将其输入到 skip-gram 模型中,以获得潜在的节点表示。DNGR 采用随机冲浪模型获得节点高维表示,然后使用去噪自编码器学习节点表示。除此之外,为了与纯结构的模型进行性能对比,本文进行了只利用结构信息的分类对比实验 DNRL-W,即 DNRL 只使用邻接矩阵 $\mathbf{W}$ 生成转移矩阵,不考虑属性信息。

(3) 联合结构-属性。这组模型同时考虑了结构信息和属性信息。TADW 和 AANE 通过矩阵分解来学习节点表示。ABRW 利用属性信息对不完整结构信息进行补偿,然后使用随机游走方法进行网络表示学习。

为了进行公平的比较,所有模型的最终节点嵌入维度均设置为 $d=128$ 。对于所有的对比方法,我们将参数设置为原文中的默认值。DNRL 和 DNGR 有一些相同的参数,我们与 DNGR 保持一致,其中, $\beta=0.98, k=4$ 。本文中先验节点标记百分比设置为 0.3,平衡结构和属性信息的超参数 α 设置为 0.7。

5.3 节点分类

本文采用节点分类任务对网络表示学习模型的性能进行评估。为了模拟不同程度的不完全结构信息,我们从原始网络中移除一定比例的链接。具体而言,我们采用与先前工作相同的数据分割策略^[24,29],从输入图中随机抽取{30%, 50%, 70%, 90%}的链接来构造不完全信息网络。在得到节点表示后,随机抽取 50% 的标记节点来训练分类器,其余节点用于测试性能。为了度量分类结果,我们采用 Micro-F1 和 Macro-F1 作为度量指标。重复该过程 10 次,并报告其平均结果。表 2—表 4 是各模型在 3 个给定数据集上的分类结果,从表中可以观察到:

(1) DNRL 模型在 3 个数据集上的分类效果均优于其他模型,并且随着可用结构信息的增加,性能也逐步提升。此外,在只考虑结构信息的情况下, DNRL-W 模型同样比其他纯结构模型有更好的分类效果,从而验证了 DNRL 模型对具有不同程度的不完全结构信息网络的有效性。

(2) 由于 SVD 只关注网络属性信息,不受链路缺失的影响,因此可以观察到在大多数情况下其分类效果优于只使用结构信息的模型,这说明了属性信息对网络嵌入的有效性。

(3) 在网络链路变得稀疏时,只使用结构信息的模型分类性能急剧下降,而结合结构信息和属性信息的模型分类性能下降缓慢,这是因为在链路缺失的情况下,网络结构单独为节点分类任务提供的有用信息非常有限。因此,在网络嵌入过程中,可以使用属性信息来弥补结构信息不完整而带来的过大损失。

(4) 总体来说,模型 DNRL 对链路缺失比例的增加不太敏感,这也在一定程度上表明了正确的对结构信息和属

性信息进行融合可以更好地进行网络表示学习,进而获得 良好的分类效果。

表 2 Cora 数据集上的节点分类结果
Table 2 Node classification results on Cora

Method	30%		50%		70%		90%	
	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1
SVD	0.695	0.660	0.727	0.693	0.697	0.658	0.720	0.677
DeepWalk	0.480	0.447	0.662	0.651	0.755	0.751	0.801	0.795
Node2Vec	0.523	0.498	0.659	0.648	0.748	0.742	0.802	0.798
DNGR	0.452	0.410	0.621	0.602	0.714	0.690	0.785	0.779
TADW	0.760	0.743	0.807	0.791	0.843	0.831	0.855	0.844
AANE	0.683	0.654	0.724	0.691	0.694	0.658	0.714	0.673
ABRW	0.754	0.733	0.785	0.775	0.824	0.810	0.865	0.856
DNRL-W	0.546	0.520	0.683	0.666	0.772	0.763	0.822	0.810
DNRL	0.846	0.826	0.867	0.856	0.889	0.874	0.897	0.890

表 3 Citeseer 数据集上的节点分类结果
Table 3 Node classification results on Citeseer

Method	30%		50%		70%		90%	
	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1
SVD	0.722	0.627	0.723	0.626	0.737	0.647	0.717	0.623
DeepWalk	0.342	0.287	0.416	0.376	0.498	0.446	0.552	0.503
Node2Vec	0.362	0.298	0.440	0.404	0.506	0.461	0.573	0.518
DNGR	0.280	0.249	0.371	0.335	0.474	0.442	0.512	0.475
TADW	0.716	0.657	0.726	0.676	0.745	0.699	0.757	0.713
AANE	0.713	0.622	0.712	0.624	0.730	0.635	0.714	0.622
ABRW	0.666	0.608	0.655	0.604	0.705	0.660	0.698	0.643
DNRL-W	0.394	0.367	0.484	0.450	0.573	0.536	0.606	0.570
DNRL	0.757	0.735	0.786	0.760	0.783	0.760	0.797	0.772

表 4 Wiki 数据集上的节点分类结果
Table 4 Node classification results on Wiki

Method	30%		50%		70%		90%	
	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1
SVD	0.751	0.685	0.738	0.629	0.747	0.666	0.733	0.625
DeepWalk	0.502	0.384	0.574	0.462	0.639	0.542	0.659	0.576
Node2Vec	0.492	0.383	0.589	0.451	0.622	0.476	0.635	0.500
DNGR	0.446	0.371	0.559	0.443	0.602	0.487	0.645	0.541
TADW	0.662	0.490	0.694	0.538	0.702	0.553	0.709	0.618
AANE	0.735	0.548	0.760	0.584	0.748	0.581	0.760	0.617
ABRW	0.709	0.640	0.727	0.620	0.743	0.658	0.736	0.611
DNRL-W	0.566	0.502	0.643	0.589	0.691	0.579	0.699	0.570
DNRL	0.798	0.754	0.814	0.750	0.807	0.750	0.819	0.782

5.4 可视化

另一种评估节点表示质量的方法是可视化,该方法能直观评估 DNRL 模型的性能。由于篇幅的限制,我们只展示了所有方法对完整的 Cora 数据集进行嵌入表示的可视化结果。

实验通过 t-SNE 工具,将所有节点的嵌入表示映射到二维空间中,并绘制在图上,以便直接观察原始网络的群落结构。图中的每个点都代表原始网络中的一个节点,相同类别的节点用相同的颜色显示。可视化结果如图 2 所示。

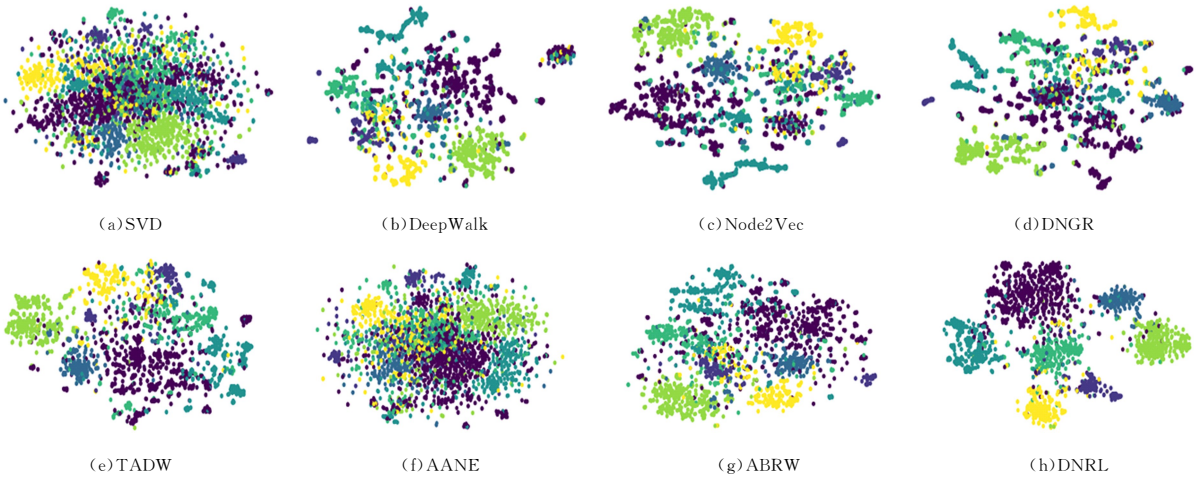


图 2 Cora 数据集的可视化结果(电子版为彩色)
Fig. 2 Visualization of Cora dataset

从图2中可以看出,与基线模型相比,本文方法可以实现更加紧凑的分组结构,且不同组之间有清晰的边界,有效反映了网络中节点的团簇关系。由此可得,DNRL模型的可视化效果更好,这也说明了DNRL模型对完整数据嵌入同样有效。

5.5 参数敏感性分析

DNRL模型在优化目标中有一个调节结构和属性信息比例的超参数 α ,通过实验分析参数 α 对节点分类性能的影响,分别在完整的Cora,Citeseer,Wiki数据集上对参数进行测试。实验中,我们将训练比例设置为50%。

图3给出了本文模型在3个数据集上选择不同 α 时的节点分类性能。实验中,参数 α 的取值为0到1之间的任意实数,其他参数设置不变。从图中可以看出,性能指标Micro-F1和Macro-F1在3个数据集上的波动不大,可见DNRL模型的性能稳定。

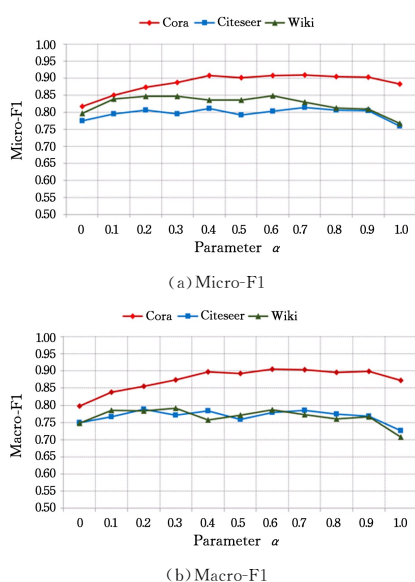


图3 参数 α 的影响

Fig. 3 Influence of α

结束语 本文针对属性网络表示学习过程中面临的结构信息不完整的情况,提出了一种半监督深度网络表示学习模型DNRL来进行网络节点特征提取。该模型首先通过构造联合概率转移矩阵将不完全结构信息和属性信息进行结合,采用Random Surfing模型将概率转移矩阵转换为共现矩阵,然后依据共现矩阵计算出PPMI矩阵,其行向量作为网络节点的高维向量表示,最后使用半监督变分自编码器对高维数据进行特征提取。在几个真实网络数据集上的节点分类和可视化的实验结果表明,DNRL模型的性能明显优于其他几种最常用的网络表示学习模型。此外,参数敏感性分析实验也验证了本文模型的鲁棒性。在未来的工作中,将进一步研究如何提升模型效率,使其适用于更大规模的网络。

参考文献

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

[2] ZHANG D K, YIN J, ZHU X Q, et al. Network representation

learning: A survey [J]. IEEE Transactions on Big Data, 2017 (99): 1.

[3] DING Y, WEI H, PAN Z S, et al. Survey of Network Representation Learning [J]. Computer Science, 2020, 47(9): 52-59.

[4] TANG J, AGGARWAL C, LIU H. Node classification signed social networks [C] // Proceedings of the 2016 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2016: 54-62.

[5] TANG J, LIU J, ZHANG M, et al. Visualizing Large-scale and High-dimensional Data [C] // Proceedings of the 25th International Conference on World Wide Web. 2016: 287-297.

[6] WANG H, ZHANG F, HOU M, et al. Shine: Signed heterogeneous information network embedding for sentiment link prediction [C] // Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. 2018: 592-600.

[7] HAN N, QIAO S J, YUAN C A, et al. A Fast Parallel Community Detection Algorithm for Mobile Social Networks [J]. Journal of Chongqing University of Technology (Natural Science), 2020, 34(1): 94-102.

[8] YIN Y, JI L X, HUANG R Y, et al. Research and development of network representation learning [J]. Chinese Journal of Network and Information Security, 2019, 5(2): 77-87.

[9] WANG J. Locally linear embedding [M] // Geometric Structure of High-Dimensional Data and Dimensionality Reduction. Berlin: Springer, 2012: 203-220.

[10] BELKIN M, NIYOGI P. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering [C] // Advances in Neural Information Processing Systems. 2002: 585-591.

[11] PEROZZI B, AI-RFOU R, SKIENA S. DeepWalk: Online Learning of Social Representations [C] // The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2014: 701-710.

[12] GROVER A, LESKOVEC J. Node2vec: Scalable Feature Learning for Networks [C] // The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 855-864.

[13] CAO S S, LU W, XU Q K. Deep neural networks for learning graph representations [C] // Proceedings of the 30th AAAI Conference on Artificial Intelligence. AAAI Press, 2016: 1145-1152.

[14] WANG D X, CUI P, ZHU W W. Structural deep network embedding [C] // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 1225-1234.

[15] YANG C, LIU Z Y, ZHAO D L, et al. Network representation learning with rich text information [C] // Proceedings of International Joint Conference on Artificial Intelligence. 2015: 2111-2117.

[16] HUANG X, LI J D, HU X. Accelerated attributed network embedding [C] // Proceedings of the 2017 SIAM International Conference on Data Mining. Philadelphia, PA: SIAM, 2017: 633-641.

[17] LIAO L Z, HE X N, ZHANG H W, et al. Attributed social network embedding [J]. IEEE Transactions on Knowledge and Data

Engineering, 2018, 30(12): 2257-2270.

[18] SMITH J A, MOODY J, MORGAN J H. Network sampling coverage II: The effect of non-random missing data on network measurement[J]. Social Networks, 2017, 48: 78-99.

[19] 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. New York: ACM, 2015: 1067-1077.

[20] CAO S S, LU W, XU Q K. Grarep: Learning graph representations with global structural information[C]// Proceedings of the 24th ACM International Conference on Information and Knowledge Management. New York: ACM, 2015: 891-900.

[21] LIANG J, JACOBS P, SUN J, et al. Semi-supervised embedding in attributed networks with outliers[C]// Proceedings of the 2018 SIAM International Conference on Data Mining. Philadelphia: SIAM, 2018: 153-161.

[22] BANDYOPADHYAY S, LOKESH N, MURTY M N. Outlier Aware Network Embedding for Attributed Networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 12-19.

[23] ZHANG D K, YIN J, ZHU X Q, et al. SINE: Scalable Incomplete Network Embedding[C]// IEEE International Conference on Data Mining. Washington DC, USA: IEEE Computer Society, 2018: 737-746.

[24] HOU C, HE S, TANG K. Attributed Network Embedding for Incomplete Structure Information[J]. arXiv: 1811. 11728, 2018.

[25] BOUMA G. Normalized(pointwise) mutual information in collocation extraction[J]. Proceedings of GSCL, 2009, 30: 31-40.

[26] YUAN F N, ZHANG L, SHI J T, et al. Theories and Application of Auto-Encoder Neural Networks: A Literature Survey [J]. Chinese Journal of Computers, 2019, 42(1): 203-230.

[27] REZENDE D J, MOHAMED S. Variational inference with normalizing flows [C] // International Conference on Machine Learning. PMLR, 2015: 1530-1538.

[28] LI Y, PAN Q, WANG S H, et al. Disentangled Variational Auto-Encoder for semi-supervised learning [J]. Information Sciences, 2019, 482(5): 73-85.

[29] YANG D J, WANG S Z, LI C Z, et al. From properties to links: Deep network embedding on incomplete graphs[C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM, 2017: 367-376.



FU Kun, born in 1979, Ph.D, associate professor. Her main research interests include social network analysis and network representation learning.