

网络表示学习的研究与发展

尹赢¹, 吉立新¹, 黄瑞阳¹, 杜立新²

(1. 国家数字交换系统工程技术研究中心, 河南 郑州 450003;
2. 解放军 63898 部队, 河南 洛阳 471003)

摘要: 网络表示学习旨在将网络中的节点表示成低维稠密且具有一定推理能力的向量, 以运用于节点分类、社区发现和链路预测等社交网络应用任务中, 是连接网络原始数据和网络应用任务的桥梁。传统的网络表示学习方法都是针对网络中节点和连边只有一种类型的同质信息网络的表示学习方法, 而现实世界中的网络往往是具有多种节点和连边类型的异质信息网络。而且, 从时间维度上来看, 网络是不断变化的。因此, 网络表示学习的研究方法随着网络数据的复杂化而不断变化。对近年来针对不同网络的网络表示学习方法进行了分类介绍, 并阐述了网络表示学习的应用场景。

关键词: 大规模信息网络; 网络表示学习; 网络嵌入; 深度学习

中图分类号: TP391.4

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2019019

Research and development of network representation learning

YIN Ying¹, JI Lixin¹, HUANG Ruiyang¹, DU Lixin²

1. China National Digital Switching System Engineering & Technological Center, Zhengzhou 450003, China
2. The 63898 Troop of PLA, Luoyang 471003, China

Abstract: Network representation learning is a bridge between network raw data and network application tasks which aims to map nodes in the network to vectors in the low-dimensional space. These vectors can be used as input to the machine learning model for social network application tasks such as node classification, community discovery, and link prediction. The traditional network representation learning methods are based on homogeneous information network. In the real world, the network is often heterogeneous with multiple types of nodes and edges. Moreover, from the perspective of time, the network is constantly changing. Therefore, the research method of network representation learning is continuously optimized with the complexity of network data. Different kinds of network representation learning methods based on different networks were introduced and the application scenarios of network representation learning were expounded.

Key words: large-scale information network, network representation learning, network embedding, deep learning

收稿日期: 2018-10-25; 修回日期: 2018-12-20

通信作者: 尹赢, 15883880517@163.com

基金项目: 国家自然科学基金创新群体资助项目 (No.61521003)

Foundation Item: The National Natural Science Foundation for Creative Research Groups of China(No.61521003)

论文引用格式: 尹赢, 吉立新, 黄瑞阳, 等. 网络表示学习的研究与发展[J]. 网络与信息安全学报, 2019, 5(2): 77-87.

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.

1 引言

在大数据时代,网络数据在现实世界中无处不在。针对网络数据的研究与分析,已经受到社会各界的广泛关注。复杂而庞大的网络数据往往难以处理,将网络数据表示成一种高效合理的形式,并将其运用于节点分类、社区发现、链路预测和网络可视化等应用任务中,对解决现实网络中的实际应用问题具有重要意义。例如,在社交网络中,可以通过节点分类方法对用户进行分类来构建用户推荐系统;在通信网络中,可以通过社区发现来观察舆情传播进程;在生物网络中,可以通过链路预测推测蛋白质之间可能存在的相互作用关系以推动疾病的治疗。这些网络分析方法都依赖于网络的表示方式,通常利用网络邻接矩阵或者由网络中节点和连边构成的图来表示网络。但是,这样简单的表示方式往往存在以下问题。

1) 连边导致计算复杂度高

网络通常由节点集和连边集组成,节点代表实体对象,而连边则反映出节点间的关联关系。连边的存在使大多数网络处理器或网络分析算法是迭代的或组合爆炸的。例如,用两个节点之间的最短路径长度或平均路径长度表示它们之间的距离。要想使用传统的网络表示方法计算这样的距离,必须枚举出2个节点之间的所有可能路径,从本质上看,这是一个组合问题,计算复杂度很高。因此,传统的网络表示方法不再适用。

2) 数据样本相互耦合导致低并行性

以传统方式表示的网络数据给并行和分布式算法的设计和实现带来了很大的困难。研究的难点在于:网络中由连边连接的节点彼此耦合,若将不同的节点分布在不同的计算单元上进行处理,往往会造成服务器之间高昂的通信代价。虽然通过对大规模图形进行巧妙分割得到子图,使

并行化性得到一定的提高,但这种方法在很大程度上依赖于网络图形的拓扑特性。

3) 机器学习方法不适用

近年来,机器学习方法(特别是深度学习方法)在很多领域都有应用,并且取得了不错的效果。这些方法针对各类问题都提供了标准的、通用的、有效的解决途径。但对于传统网络数据的表示,大多数现有的机器学习方法往往是不适用的。虽然可以简单地用网络邻接矩阵中节点对应的行向量来表示该节点。但是这种表示往往具有很高的维度,使后续的网络处理和分析变得十分困难。

为了解决传统方法的缺陷,研究学者提出了一种新的网络表示方式:将网络中的节点表示成具有一定推理能力的向量映射到低维空间中,旨在以更加直观高效的方式尽可能地保留节点间的原始关系,从而便捷地作为机器学习模型的输入,运用于节点分类、链路预测和社区发现等社交网络分析应用中。将节点表示成向量映射到低维空间中,向量间的距离反映出原始网络中节点间的连边关系。这就解决了由网络中的连边带来的一系列网络分析问题。同时,这种网络表示方式不需要人工标注网络特征,而是自动从原始网络数据中提取出有用的网络表征,能够很好地保留原始网络中的信息实现网络重构。

2 网络表示学习的基本定义

定义1 网络表示学习:网络表示学习是表示学习的一个子集,又称网络特征学习。定义网络 $G=(V,E)$, V 表示网络节点集合, E 表示网络中连边的集合。网络 G 对应的顶点特征矩阵 A 是一个稀疏且庞大的矩阵,若顶点的属性特征为 n 维,则 A 的维度通常为 $V \times n$ 。通过网络表示学习,对网络中的任意节点 $v \in V$,都可以学习得到一个低维的实值向量表示 $r_v \in R^k$,且满足 k 远小于 $|V|$ 。

定义2 同质信息网络和异质信息网络^[1]:定义网络 $G=(V,E)$, V 表示网络中的节点集合, E

表示网络中边的集合。定义节点类型映射函数 $f: V \rightarrow A$ ，其中，每个节点 $v \in V$ 对应特定的类型 $f(v) \in A$ ；定义链接类型映射函数 $\gamma: E \rightarrow R$ ，其中，每条链接 $e \in E$ 对应特定的类型 $\gamma(e) \in R$ 。当 R 和 A 满足 $|A|=1$ 且 $|R|=1$ 时，称这样的网络为同质信息网络，同质信息网络包含单一的节点类型和边类型，如合作网、朋友圈等；当 R 和 A 满足 $|A|>1$ 或 $|R|>1$ 时，即边的类型数或者节点类型数大于 1，称这样的网络为异质信息网络，如文献信息网络等。如图 1 所示，由作者、文献、主题以及刊物组成的文献信息网络就是一个典型的异质信息网络。该网络有 4 类节点：作者、论文、刊物、主题，每篇论文与作者、刊物、主题之间都存在着链接，这些链接属于不同的关系类型。

3 网络表示学习方法介绍

随着网络的规模化和复杂化，如何分析网络中包含的丰富信息是一项非常重要的研究任务。网络表示学习将网络中的节点表示成向量映射到低维空间，可以将其作为机器学习算法的输入来挖掘网络数据中蕴含的丰富信息。现有的网络表示学习方法主要针对同质信息网络和异质信息网络。接下来分别对基于这两种网络的表示学习方法进行分析。

法进行分析。

3.1 针对同质信息网络的表示学习方法

基于网络表示学习方法所考虑的网络信息内容，本节将同质信息网络的表示学习方法分为两类：基于网络结构的网络表示学习和结合外部信息的网络表示学习。

3.1.1 基于网络结构的同质网络表示学习

网络结构可以划分为以不同粒度呈现的不同组。网络表示学习中常用的网络结构包括邻域结构、高阶节点邻近结构和网络社团。依据算法基于的不同方法将其分为 3 类：基于矩阵特征向量和矩阵分解的方法、基于浅层神经网络的方法和基于深度学习的方法。

基于矩阵特征向量和矩阵分解的方法：早期的网络表示学习算法大多是基于矩阵特征向量和矩阵分解的方法，Saul 等^[2-3]提出局部线性表示，通过邻居节点的线性组合来表示当前节点，算法假设节点与其邻居节点位于同一流形区域。当前节点可表示为

$$X_i = \sum_j w_{ij} X_j, \forall X_i \in V \quad (1)$$

其中， V 表示节点集合。算法将邻居节点表示的加权和与中心节点表示间的距离作为损失函数，通过

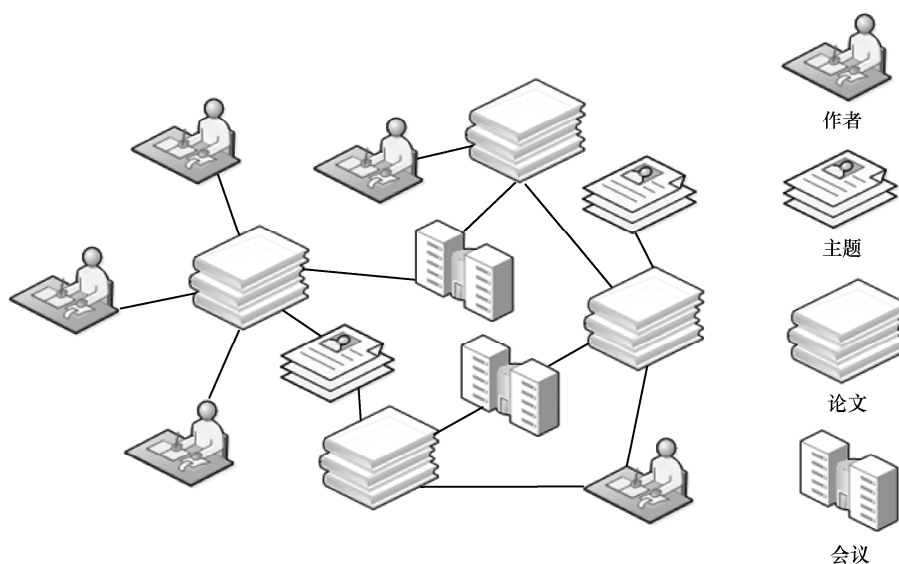


图 1 异质信息网络实例

优化式(2)获得保留邻域结构的节点表示矩阵 $\mathbf{X}^{N \times d}$ 。

$$\phi(\mathbf{X}) = \sum_i \left| Y_i - \sum_j w_{ij} X_j \right|^2 \quad (2)$$

利用节点的邻接矩阵对网络进行表示学习往往会带来很高的计算复杂度,有研究学者提出 Laplace 特征映射^[4]方法,该方法首先依据 K 近邻来构造出网络结构图,然后利用热核函数确定图中节点 i 和 j 之间的权重 w_{ij} 。最后,通过最小化式(3)表示的损失函数得到节点表示。其目标是将网络结构图中相似性高的节点在映射到低维空间后,保持相近的距离。

$$Loss(\mathbf{Z}) = \sum_i |z_i - z_j|^2 w_{ij} = trace(\mathbf{Z}^T \mathbf{L} \mathbf{Z}) \quad (3)$$

其中, $\mathbf{L}$ 是拉普拉斯矩阵且 $\mathbf{L} = \mathbf{D} - \mathbf{W}$, $\mathbf{D}$ 是对角阵且 $D_{ii} = \sum_j w_{ji}$ 。后来,He 等^[5]提出了局部保

留映射算法,对非线性的 Laplace 特征映射方法进行线性的近似。该算法提出一个转移矩阵 $\mathbf{A}$,使对于节点 x_i 的表示 z_i 有 $z_i = \mathbf{A}^T x_i$ 。Tu 等^[6]提出的图形分解算法是第一个在 $\mathcal{O}(|E|)$ 的时间复杂度上完成节点向量化的算法。与 Laplace 特征映射相比,图形分解算法通过对邻接矩阵进行分解得到节点的低维稠密向量表示,并通过对向量直接点乘或者求余弦相似度将两个表示向量进行重构,然后将重构后边的权重作为衡量表示效果的方法,其目标函数为

$$o(\mathbf{Z}, \lambda) = \frac{1}{2} \sum_{(i,j) \in E} (w_{ij} - \langle Z_i, Z_j \rangle)^2 + \frac{\lambda}{2} \sum_i \|Z_i\|^2 \quad (4)$$

网络表示学习不仅要考虑保留网络的局部结构,网络全局结构的保留也非常重要。Cao 等^[7]提出 GreRep 算法通过节点的 $k(k>2)$ 阶相似性来保留全局网络结构。对于网络的邻接矩阵 $\mathbf{A}$, k 阶概率转移矩阵为 $\mathbf{A}^k$, 矩阵中的元素 $A_{i,j}^k$ 表示当前节点 i 到邻居节点 j 的 k 步转移概率 $p_k(j|i)$ 。基于 Skip-gram 模型^[8], 算法的损失函数可表示为

$$Loss_k = \left(\sum_j p_k(j|i) \log \sigma(u_j^T u_i) \right) + \lambda E_{j^* \sim p_k(V)} [\log \sigma(-u_i^T u_{j^*})] \quad (5)$$

其中, $\sigma(x) = (1 + e^{-x})^{-1}$, $p_k(V)$ 是节点在网络中的分布, j^* 表示通过负采样得到的节点。此外, GraRep 将损失函数重新定义为矩阵分解问题,对于每个 k 步损失函数,梯度下降法 (SVD) 可以直接用于节点的表示学习。通过对每个函数的迭代更新,可以获得网络的全局表示。也有一些网络表示学习方法保留了网络的非对称性和非传递性等属性。Ou 等^[9]提出的 HOPE 保留了有向网络的非对称传递性,网络的非对称传递性是指:如果网络中节点 i 和 j 以及节点 j 和 v 之间存在有向连边,那么节点 i 和 v 之间很可能存在有向连边,但是不一定存在由 v 到 i 的有向连边。同时,文章将基于邻接矩阵得到的许多相似性度量方法(如 Katz 指标^[10]) 都统一为矩阵分解。依据网络中节点的高阶相似性,HOPE 构建了不同的非对称关系矩阵,然后直接利用 SVD 算法得到节点的低维表示。同时,HOPE 将原始的 SVD 问题转化为广义的 SVD 问题,使算法的时间复杂度大大减小,这也就使 HOPE 能适用于大规模网络。

基于浅层神经网络的方法:受自然语言处理中 Word2vec 算法^[11]的启发,Perozzi 等^[12]提出了 DeepWalk 算法用于学习网络中的节点表示,DeepWalk 通过随机游走遍历网络中的节点,从而得到一个有序的节点序列,该过程实际是对网络中的节点进行采样的过程,在随机游走过程中,节点越邻近,共现的概率越大;在通过随机游走得到节点序列以后,该算法利用 skip-gram 模型由单个节点预测前后序列,学习得到该节点的向量表示。Grover 等^[13]提出的 node2vec 改进了 DeepWalk 中的随机游走过程。DeepWalk 选取随机游走序列中下一个节点的方式是均匀随机分布的,而 Node2vec 通过引入两个参数 p 和 q ,将宽

度优先搜索和深度优先搜索引入了随机游走序列的生成过程。宽度优先搜索注重邻近节点并刻画了相对局部的一种网络表示，而深度优先搜索则反映了更高层面上节点间的同质性。Tang 等^[14]提出的 LINE 算法能够处理任意类型的大规模网络，包括有向和无向、有权重和无权重。该算法保留了网络中节点的一阶相似性和二阶相似性。一阶相似性是指有连边的节点之间的相似性，主要由连边权重决定，没有连边的节点间的相似性则为 0。二阶相似性是指节点的邻近网络结构的相似性，若 2 个节点没有共同的一阶邻居，则二阶相似性为 0。如图 2 所示，节点 5 和 6 的一阶相似性为 0，但二者具有很强的二阶相似性，节点 6 和 7 具有一阶相似性但不具备二阶相似性。一阶相似性和二阶相似性互补，使该算法能够很好地保留网络的局部结构和全局结构。

基于深度学习的方法：SDNE 算法^[15]利用深度神经网络对网络的表示学习进行建模，该半监督的深度学习模型由多重的非线性映射函

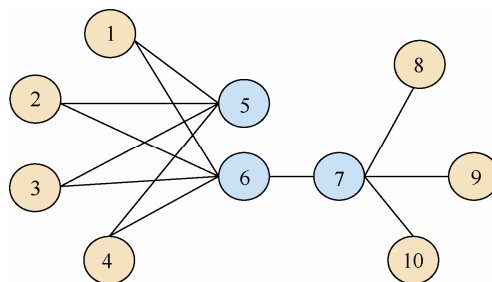


图2 一阶相似性和二阶相似性实例

数组成，将输入的节点映射到高度非线性空间中以获取网络的结构信息。模型如图 3 所示，主要由两部分组成：一是由拉普拉斯矩阵监督的能保留局部信息的一阶相似度模块，二是基于无监督深度自编码器的能保留全局信息的二阶相似度模块。

模型将一阶相似度和二阶相似度结合起来同时保留网络结构的局部信息和全局信息，其目标函数可表示为

$$L = L_{2nd} + \alpha L_{1st} + \nu L_{reg} \quad (6)$$

其中， νL_{reg} 用来防止过拟合，且有

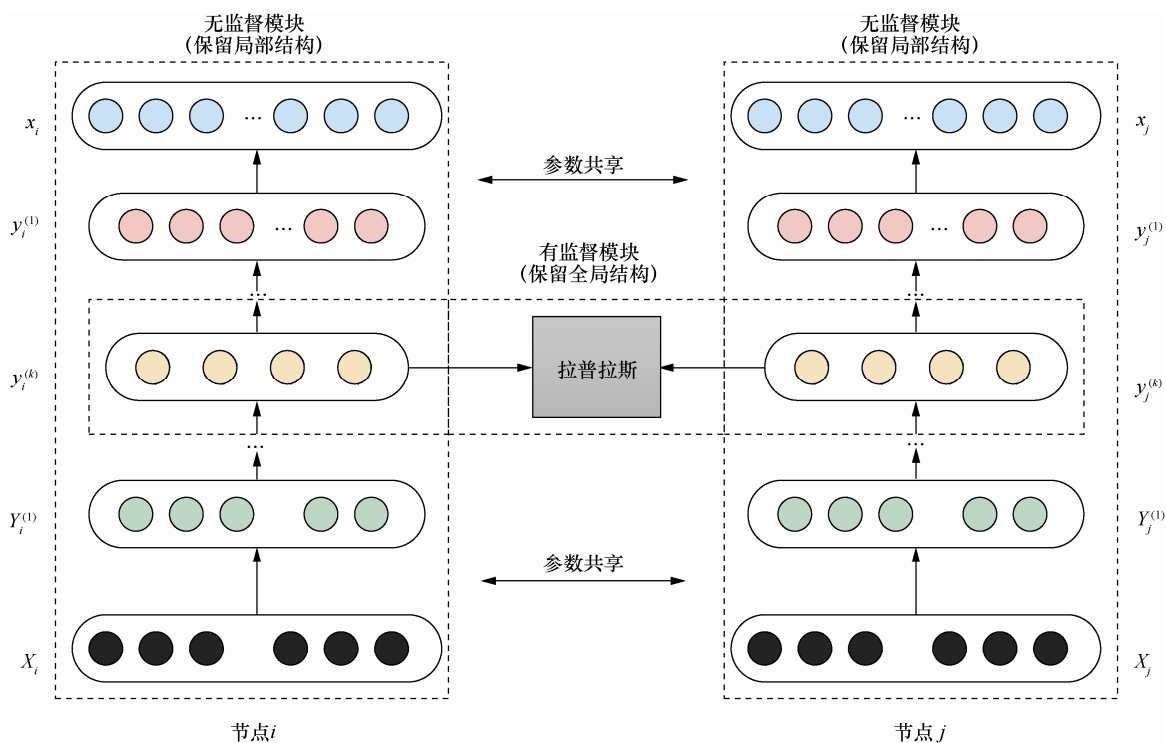


图3 SDNE 算法模型框架

$$L_{2nd} = \sum_i \|(\hat{x}_i - x_i) \odot b_i\|^2 \quad (7)$$

不同于传统的直推式网络表示学习算法, Hamilton 等^[16]提出一种适用于大规模网络的归纳式学习方法 GraphSAGE。该算法通过聚集采样到的邻居节点表示来更新当前节点的特征表示, 而不是像直推式方法将每个节点单独进行训练。在聚类过程中, 网络第 k 层聚集到的邻居即为宽度优先搜索 (BFS) 过程中第 k 层的邻居。同时, 作者在算法中利用了 3 种聚类策略: Mean、LSTM 和 MaxPooling。

GraphGAN^[17]引入对抗生成网络进行网络表示学习, 算法模型主要分为生成器和判别器两个部分。在对抗学习过程中, 生成器 $G(v|v_c; \theta_G)$ 尽可能地拟合节点 v_c 在网络中的连接分布 $p_{true}(v|v_c)$, 并且从节点集 V 中生成最可能与节点 v_c 连接的节点; 判别器 $D(v, v_c; \theta_D)$ 则用来判别节点对 (v, v_c) 是否存在连接, 最终输出节点对 (v, v_c) 存在连边的概率。生成器 $G(v|v_c; \theta_G)$ 和判别器 $D(v, v_c; \theta_D)$ 相互制衡: 生成器尽量完美地拟合

$p_{true}(v|v_c)$ 并生成类似于 v_c 的真实邻居的相关节点以达到欺骗判别器的目的; 而判别器则负责检测节点是 v_c 的真实邻居还是其对应生成器 $G(v|v_c; \theta_G)$ 生成的邻居。

总之, 判别器和生成器的训练过程是一个最大最小相互制约的过程, 算法的目标函数可表示为

$$\begin{aligned} & \min_{\theta_G} \max_{\theta_D} V(G, D) \\ &= \sum_{c=1}^V (E_{v \sim G(\cdot|v_c; \theta_G)} [\log(1 - D(v, v_c; \theta_D))] + \\ & \sum_{c=1}^V (E_{v \sim p_{true}(\cdot|v_c)} [\log D(v, v_c; \theta_D)]) \end{aligned} \quad (8)$$

常见的基于网络结构的同质网络表示学习方法如表 1 所示。

3.1.2 结合外部信息的同质网络表示学习

传统的网络表示学习方法都是基于网络拓扑结构的方法, 但除了网络拓扑结构, 网络还包括标签和文本等重要的外部信息, 这些外部信息也是网络表示学习的一个重要研究内容。

表 1 算法适用网络及实验数据集

算法	GreRep	HOPE	DeepWalk	Node2vec	LINE	SDNE	GraphGAN
网络属性	有向	✓		✓	✓	✓	
	无向	✓	✓	✓	✓	✓	✓
	有权	✓		✓	✓		✓
	标签						
实验数据集	BlogCatalog	✓	✓	✓	✓	✓	✓
	Flickr		✓		✓	✓	
	YouTube		✓		✓	✓	
	Wikipedia			✓	✓		✓
	PPI			✓			
	DBLP	✓					
	20-newsgroup	✓				✓	
	Cora		✓				
	SN-Twitter		✓				

为了适应节点分类等机器学习任务,Tu 等^[18]提出了针对带标签网络的半监督网络表示学习模型 MMDW。该模型基于矩阵分解,利用 SVM 分类器和节点的标签信息来寻找优化的分类边界。通过优化 SVM 最大间隔分类器和转换为矩阵因子分解形式的 DeepWalk 模型,MMDW 模型能学习到包含网络结构信息和标签信息的节点表示。

在网络数据中,节点除了包含标签信息,还包含一些文本信息。这些文本信息也可以作为网络信息的补充,有助于进行更加合理适用的网络表示学习。TADW^[19]基于矩阵分解将节点的文本信息引入网络表示学习中,TADW 模型框架如图 4 所示。

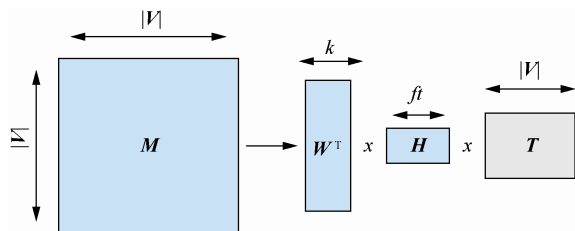


图 4 TADW 模型框架

模型将关系矩阵 M 分解为 T 、 W 、 H 这 3 个矩阵的乘积。其中, T 表示文本特征矩阵, H 和 W 表示参数矩阵,算法通过梯度下降法来迭代更新参数矩阵。模型的目标函数可表示为

$$\min_{W,H} \|M - W^T H T\|_F^2 + \frac{\lambda}{2} (\|W\|_F^2 + \|H\|_F^2) \quad (9)$$

但是 TADW 算法的计算代价很高,而且算法只是将节点属性进行简单结合,这就造成了网络中语义信息的大量丢失。Sun 等^[20]提出一种新的 CENE 算法,将节点内容视为一种特殊的节点来扩展网络,如图 5 所示。扩展的网络中包含两种节点:原始网络节点和构造的“节点内容”节点。基于不同节点,扩展的网络也包括两种边:原始节点之间的连边和原始节点

与“节点内容”之间的连边。算法使用逻辑回归函数学习扩展的网络,并通过负采样的方法优化目标函数。通过该算法得到的网络表示不仅可以保留网络结构特征,还可以保留节点和内容之间的语义信息。

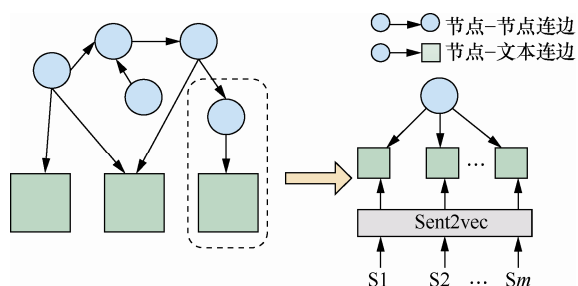


图 5 CENE 算法扩展的网络结构

Pan 等提出一个能结合网络结构、节点内容和节点标签这 3 种网络信息的深度学习模型 TriDNR^[21],模型结构如图 6 所示,其目标函数可表示为

$$L = (1 - \alpha) \sum_{i=1}^N \sum_{s \in S} \sum_{-b \leq j \leq b, j \neq 0} \log P(v_{i+j} | v_i) + \alpha \sum_{i=1}^N \sum_{-b \leq j \leq b} \log P(w_j | v_i) + \alpha \sum_{i=1}^{|L|} \sum_{-b \leq j \leq b} \log P(w_j | c_i) \quad (10)$$

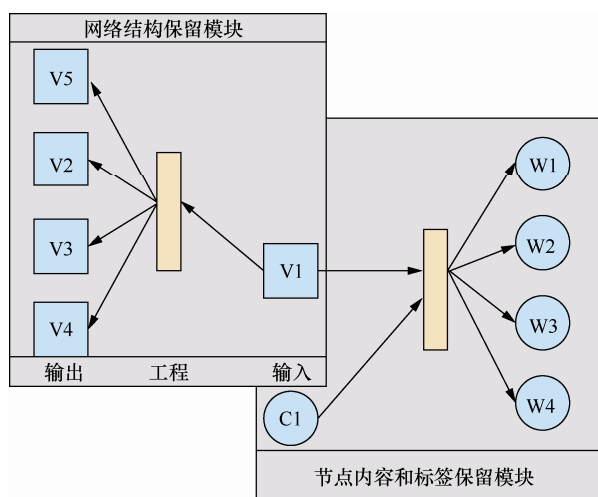


图 6 TriDNR 算法模型框架

类似于 DeepWalk 模型,该模型通过随机游走生成节点的表示向量来保留节点在结构上的相似性;然后用另一个神经网络学习节点上下文的相关性;同时,将节点标签作为输入,直接在标签和上下文之间建模来学习标签向量和单词向量。

常见的结合外部信息的同质信息网络表示学习算法如表 2 所示。

表 2 算法适用网络及实验数据集

算法	MMDW	TADW	CENE	TriDNR
网络属性	有向	✓	✓	✓
	有权			
	文本		✓	✓
	标签	✓		✓
数据集	DBLP		✓	✓
	CiteSeer	✓	✓	✓
	Cora	✓	✓	
	Wiki	✓	✓	

3.2 针对异质信息网络的表示学习方法

早期的网络表示学习算法是针对节点和边类型只有一种的同质信息网络,而将所有节点和连边视为同一类的同质网络忽略了网络中丰富的语义信息,并不能有效地反映现实世界中的网络。类型化、半结构化的异质信息网络建模可以捕获真实世界中最根本的语义信息^[22]。目前,针对异质信息网络的表示学习方法研究相对较少,主要分为 3 类:基于随机游走的方法、基于分解的方法和结合应用任务的方法。

1) 基于随机游走的方法

Yu 等^[23]受同质网络中 Node2vec 算法的启发,提出了 Metapath2vec 算法,该方法通过在异质信息网络中进行随机游走来获取节点的邻居节点集合。不同于同质信息网络中的随机游走, Metapath2vec 算法中的随机游走是在元路径^[24]的指导下进行的,其跳转概率为

$$p(v^{i+1}|v_t, \rho) = \begin{cases} \frac{1}{|N_{t+1}(v_t^i)|}, & (v^{i+1}, v_t^i) \in E, \phi(v^{i+1}) = t+1 \\ 0, & (v^{i+1}, v_t^i) \in E, \phi(v^{i+1}) \neq t+1 \\ 0, & (v^{i+1}, v_t^i) \notin E \end{cases} \quad (11)$$

基于元路径进行随机游走得到的节点序列,不仅包含网络的结构信息,还包含网络中的语义信息。在得到节点序列之后,算法利用 skip-gram 模型得到节点的向量表示,并采用负采样的方法更新式(12)的目标函数。

$$\log \sigma(X_c, X_c) + \sum_{m=1}^M E_u m \sim p(u) [\log \sigma - X_u m \cdot X_v] \quad (12)$$

Metapath2vec 算法在利用 softmax 函数进行归一化时,并没有考虑节点的类型,而是和同质网络表示学习算法 DeepWalk、Node2vec 的处理方式一样,是将所有节点的特征表示之和作为 softmax 函数的分母。作者在提出 Metapath2vec 的同时,提出了 Metapath2vec++^[23],该算法在 Metapath2vec 的基础上改进了 softmax 函数,如式(13)所示。Metapath2vec++对不同类型的节点分别进行归一化,相当于在浅层神经网络的输出层根据节点类型将整个异质信息网络分解成不同的同质信息网络。通过 Metapath2vec++得到的节点表示在运用节点分类、聚类 and 相似性搜索等应用任务时,具有更高的精度和可靠性。

$$p(c_i|v; \theta) = \frac{e^{X_{c_i} \cdot X_v}}{\sum_{u_i \in V_i} e^{X_{u_i} \cdot X_v}} \quad (13)$$

2) 基于分解的方法

从网络分解的角度对异质信息网络进行表示学习,往往有化繁为简的效果。Tang 等提出的 PTE 算法^[25],将包含词、文件、标签这 3 种节点类型的文本信息网络分解成 3 个不同的子网络:词-词网络、词-文件网络、词-标签网络。然后对这 3 个不同的网络进行表示学习,得到不同类型节

点的向量表示。Shi 等提出的 HERec^[26]模型基于元路径从异质信息网络中抽取同类节点序列,相当于从异质信息网络中抽取多个同质信息网络。基于不同元路径,该模型抽取出的同类节点序列有所不同。算法将基于不同元路径抽取出的同类节点分别进行表示学习,然后利用融合函数对节点的不同表示进行融合,并将向量整合到矩阵分解模型中,通过矩阵分解模型和融合函数一起对预测任务进行联合优化。该方法与基于卷积神经网络的监督方法相比调整参数更少。

3) 结合应用任务的方法

异质信息网络表示学习算法常和具体的数据挖掘任务相结合,Sun 等^[27]首先提出基于元路径评估异质信息网络中节点间相似性的算法 PathSim,该算法基于对称元路径,可测量异质信息网络中同类型对象间的相似程度。Sun 等^[28]提出的 Pathsclclus 算法利用用户提供的种子作为聚类的先验知识,提出一个基于聚类的生成模型来模拟节点间关系的生成,并通过评估元路径关系矩阵与基于先验知识的聚类结果的一致性来优化元路径权重和聚类效果。也有算法用于异质信息网络的节点分类,Luo 等^[29]提出的 HetPathMine 算法通过构建元路径选择模型来将异质信息网络中带标签的数据进行分类。Kong 等^[30]提出了一种适用于大规模网络的基于元路径的分类算法,该算法考虑了不同元路径所蕴含的不同语义,以捕获不同连接类型对实体对象间关联程度的影响,从而有效地对异质网络进行大规模集体分类。Fu 等^[31]提出的 HIN2vec 利用神经网络模型来实现异质信息网络中节点和边的表示。学习到的向量可以直接用于链路预测任务,并具有很好的效果。

4 应用场景

常见的网络表示学习应用场景主要有节点分类、链路预测、社区发现和可视化等。

1) 节点分类

在进行网络数据处理时,往往需要将网络中的节点进行合理分类。例如,在社交网络中,可以根据用户的兴趣爱好对用户分类以进行相关推荐。用户的兴趣爱好是将用户进行分类的类别标注信息,是对用户进行分类的依据。然而实际数据中的类别标注信息十分稀疏,可以利用网络表示学习算法来对节点进行编码,使在标注信息很少的情况下也能得到很好的分类效果。

2) 链路预测

链路预测是指对网络中丢失的边或者潜在的边进行预测,可以帮助分析数据缺失的网络以及网络的演化,其在现实生活中具有广泛的运用。例如,可以利用链路预测方法基于当前网络结构预测可能成为朋友的用户,从而对用户进行好友推荐。链路预测的常见评价指标为 AUC 值,当在样本集中随机挑选一个正样本以及一个负样本,分类算法根据计算得到的正样本分数值高于负样本分数值的概率就是 AUC 值。

3) 社区发现

社区发现是指将网络中相似的节点划分到同一社区,与节点分类的不同之处在于,社区发现不依靠任何标记的数据,而是直接对网络中的节点进行无监督的聚类。社区发现是一个自由度较高的网络应用任务,可以结合网络表示学习利用社区发现算法对网络中的用户进行自动分组。

4) 可视化

可视化是指利用一定的计算机技术将数据转换成图形或者图像直观地呈现出来,从而清晰地表达出信息。通过网络表示学习,可以在低维向量空间得到网络节点的表示向量,而这些向量可以直接用于网络的可视化,使网络可视化变得高效便捷。

5 结束语

网络表示学习的研究方法随着网络数据的复

杂化而不断变化,保留网络结构和网络属性是网络表示学习的基础。如果网络表示学习方法没有很好地保留网络结构和属性,那么当节点映射到低维向量空间中时,会造成严重的信息丢失。此外,也可以考虑将某些领域知识作为高级信息来辅助网络表示学习。从现有的网络表示学习工作来看,该领域是一个新兴的且非常具有前景的研究方向。近年来,网络表示学习取得了丰硕的研究成果,但仍存在以下挑战。

1) 现实网络是蕴含庞大数据量的大规模复杂网络,如何高效处理这些庞大的数据,是网络表示学习亟待解决的一个难点。

2) 动态网络的研究相对较少,如何从时间维度上综合考虑网络的重构是网络表示学习一个研究难点。

3) 现有的网络表示学习算法只能适用于一般的网络应用任务,如节点分类和链路预测等,这些方法主要针对一般的网络结构,对一些具体的网络应用任务可能并不适用。

参考文献:

- [1] SUN Y, HAN J. Mining heterogeneous information networks: a structural analysis approach[J]. ACM Sigkdd Explorations Newsletter, 2013, 14(2): 20-28.
- [2] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323-2326.
- [3] SAUL L K, ROWEIS S T. An introduction to locally linear embedding[J]. Journal of Machine Learning Research, 2008.
- [4] BELKIN M, NIYOGI P. Laplacian eigenmaps and spectral techniques for embedding and clustering[J]. Advances in Neural Information Processing Systems, 2002, 14(6): 585-591.
- [5] HE X, NIYOGI P. Locality preserving projections[C]//Advances in Neural Information Processing Systems. 2004: 153-160.
- [6] AHMED A, SHERVASHIDZE N, NARAYANAMURTHY S, et al. Distributed large-scale natural graph factorization[C]// International Conference on World Wide Web. ACM, 2013: 37-48.
- [7] CAO S, LU W, XU Q. GraRep: learning graph representations with global structural information[C]//ACM International Conference on Information and Knowledge Management. 2015: 891-900.
- [8] CHENG W, GREAVES C, WARREN M. From n -gram to skipgram to conccgram[J]. International Journal of Corpus Linguistics, 2006, 11(4):411-433.
- [9] OU M, CUI P, PEI J, et al. Asymmetric transitivity preserving graph embedding[C]//The 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016: 1105-1114.
- [10] KATZ L. A new status index derived from sociometric analysis[J]. Psychometrika, 1953, 18(1): 39-43.
- [11] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013.
- [12] PEROZZI B, AL-RFOU R, SKIENA S. DeepWalk:online learning of social representations[C]//ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2014: 701-710.
- [13] GROVER A, LESKOVEC J. node2vec: scalable feature learning for networks[C]//The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 855-864.
- [14] TANG J, QU M, WANG M, et al. LINE: large-scale information network embedding[C]//International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2015: 1067-1077.
- [15] WANG D, CUI P, ZHU W. structural deep network embedding[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 1225-1234.
- [16] HAMILTON W, YING Z, LESKOVEC J. Inductive representation learning on large graphs[C]//Advances in Neural Information Processing Systems. 2017: 1024-1034.
- [17] 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.
- [18] TU C, ZHANG W, LIU Z, et al. Max-margin DeepWalk: discriminative learning of network representation[C]//International Joint Conference on Artificial Intelligence. AAAI Press, 2016. 3889: 3895.
- [19] YANG C, ZHAO D, ZHAO D, et al. Network representation learning with rich text information[C]// International Conference on Artificial Intelligence. AAAI Press, 2015: 2111-2117.
- [20] SUN X, GUO J, DING X, et al. A general framework for content-enhanced network representation learning[J]. arXiv preprint arXiv:1610.02906,2016.
- [21] PAN S, WU J, ZHU X, et al. Tri-party deep network representation[C]//International Joint Conference on Artificial Intelligence. AAAI Press, 2016:1895-1901.
- [22] 孙艺洲, 韩家伟. 异构信息网络挖掘原理与方法[M]. 北京: 机械工业出版社, 2016,2.
- [23] SUN Y Z, HAN J W. Principles and methods of heterogeneous information network mining[M]. Beijing: Machinery Industry Press, 2016.
- [24] DONG Y, CHAWLA N V, SWAMI A, et al. metapath2vec: scalable representation learning for heterogeneous networks[C]// ACM Sigkdd International Conference on Knowledge Discovery & Data Mining. 2017:135-144.
- [25] SUN G, ZHANG X. Graph embedding with rich information through heterogeneous network[J]. arXiv preprint arXiv: 1710.06879, 2017.
- [26] TANG J, QU M, MEI Q. Pte: Predictive text embedding through large-scale heterogeneous text networks[C]//The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 1165-1174.
- [27] SHI C, HU B, ZHAO X, et al. Heterogeneous information network embedding for recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(2): 357-370.

- [27] SUN Y, HAN J, YAN X, et al. Pathsim: meta path-based top- k similarity search in heterogeneous information networks[J]. Proceedings of the VLDB Endowment, 2011, 4(11): 992-1003.
- [28] SUN Y, NORICK B, HAN J, et al. PathSelClus: integrating meta-path selection with user-guided object clustering in heterogeneous information networks[J]. ACM Transactions on Knowledge Discovery from Data, 2013, 7(3): 1-23.
- [29] LUO C, GUAN R, WANG Z, et al. Hetpathmine: a novel transductive classification algorithm on heterogeneous information networks[C]//European Conference on Information Retrieval. 2014: 210-221.
- [30] KONG X, YU P S, DING Y, et al. Meta path-based collective classification in heterogeneous information networks[J]. 2012: 1567-1571.
- [31] FU T Y, LEE W C, LEI Z. HIN2Vec: explore meta-paths in heterogeneous information networks for representation learning[C]//ACM, 2017: 1797-1806.



吉立新（1969- ），男，江苏淮安人，国家数字交换系统工程技术研究中心研究员，主要研究方向为通信网信息安全。



黄瑞阳（1986- ），男，福建漳州人，国家数字交换系统工程技术研究中心助理研究员，主要研究方向为文本挖掘、图挖掘。

[作者简介]



尹赢（1994- ），女，四川绵竹人，国家数字交换系统工程技术研究中心硕士生，主要研究方向为复杂网络、网络表示学习。



杜立新（1996- ），男，甘肃武威人，主要研究方向为大数据分析。