软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(6):2749−2764 [doi: 10.13328/j.cnki.jos.006486]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种融合伴随信息的网络表示学习模型*

杜航原[1], 王文剑[1], 白 亮[2]

[1](山西大学 计算机与信息技术学院, 山西 太原 030006)

[2](计算智能与中文信息处理教育部重点实验室 (山西大学), 山西 太原 030006)

通信作者: 王文剑, E-mail: wjwang@sxu.edu.cn

摘 要: 网络表示学习被认为是提高信息网络分析效率的关键技术之一, 旨在将网络中每个节点映射为低维隐空间中的向量表示, 并使这些向量高效的保持原网络的结构和特性. 近年来, 大量研究致力于网络拓扑和节点属性的深度挖掘, 并在一些网络分析任务中取得了良好应用效果. 事实上, 在这两类关键信息之外, 真实网络中广泛存在的伴随信息, 反映了网络中复杂微妙的各种关系, 对网络的形成和演化起着重要作用. 为提高网络表示学习的有效性, 提出了一种能够融合伴随信息的网络表示学习模型 NRLIAI. 该模型以变分自编码器 (VAE) 作为信息传播和处理的框架, 在编码器中利用图卷积算子进行网络拓扑和节点属性的聚合与映射, 在解码器中完成网络的重构, 并融合伴随信息对网络表示学习过程进行指导. 该模型克服了现有方法无法有效利用伴随信息的缺点, 同时具有一定的生成能力, 能减轻表示学习过程中的过拟合问题. 在真实网络数据集上, 通过节点分类和链路预测任务对 NRLIAI 模型与几种现有方法进行了对比实验, 实验结果验证了该模型的有效性.

关键词: 网络表示学习; 伴随信息; 变分自编码器 (VAE); 图卷积网络 (GCN); 互信息

中图法分类号: TP181

中文引用格式: 杜航原, 王文剑, 白亮. 一种融合伴随信息的网络表示学习模型. 软件学报, 2023, 34(6): 2749–2764. http://www.jos.org.cn/1000-9825/6486.htm

英文引用格式: Du HY, Wang WJ, Bai L. Network Representation Learning Model Integrating Accompanying Information. Ruan Jian Xue Bao/Journal of Software, 2023, 34(6): 2749–2764 (in Chinese). http://www.jos.org.cn/1000-9825/6486.htm

Network Representation Learning Model Integrating Accompanying Information

DU Hang-Yuan[1], WANG Wen-Jian[1], BAI Liang[2]

[1](School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)

[2](Key Laboratory of Computational Intelligence and Chinese Information of Ministry of Education (Shanxi University), Taiyuan 030006, China)

Abstract: Network representation learning is regarded as a key technology for improving the efficiency of information network analysis. It maps network nodes to low-dimensional vectors in a latent space and maintains the structure and characteristics of the original network in these vectors efficiently. In recent years, many studies focus on exploring network topology and node features intensively, and the application bears fruit in many network analysis tasks. In fact, besides these two kinds of key information, the accompanying information widely existing in the network reflects various complex relationships and plays an important role in the network's construction and evolution. In order to improve the efficiency of network representation learning, a novel model integrating the accompanying information is proposed with the name NRLIAI. The model employs the variational auto-encoders (VAE) to propagate and process information. In addition, it aggregates and maps network topology and node features by graph convolutional operators in the encoder, reconstructs the

* 基金项目: 国家自然科学基金 (61902227, 62076154, U1805263, 61773247); 山西省自然科学基金 (201901D211192); 山西省高等学校科技创新项目 (2019L0039)
本文由"面向开放场景的鲁棒机器学习"专刊特约编辑陈恩红教授、李宇峰副教授、邹权教授推荐.
收稿时间: 2021-03-09; 修改时间: 2021-07-16; 采用时间: 2021-08-27; jos 在线出版时间: 2022-07-22
CNKI 网络首发时间: 2022-11-15

© 中国科学院软件研究所 http://www.jos.org.cn

network in the decoder, and integrates the accompanying information to guide the network representation learning. Furthermore, the proposed model solves the problem that the existing methods fail to utilize the accompanying information effectively. At the same time, the model possesses a generative ability, which enables it to reduce the overfitting problem in the learning process. With several real-world network datasets, this study conducts extensive comparative experiments on the existing methods of NRLIAL through node classification and link prediction tasks, and the experimental results have proved the feasibility of the proposed model.

Key words: network representation; accompanying information; variational auto-encoder (VAE); graph convolutional network (GCN); mutual information

在现实世界的众多应用场景中, 广泛存在着大量以信息网络为载体的关联型数据, 例如, 货币在不同实体间频繁流动形成的金融网络, 各类工业设备与传感器互联而成的工业物联网, 生物组织中蛋白质之间相互作用形成的蛋白质互作用网络, 以及社交用户相互关联形成的电子社交网络等. 这些场景中, 数据样本之间不再彼此独立, 而是通过错综复杂的关联相互作用和影响, 信息网络成为其最自然和直接的表达形式——数据样本被记录为网络中的节点, 样本间的关联被表示为网络节点之间的边. 一直以来, 信息网络对于关联型数据的表达和分析发挥着重要的基础性作用. 然而, 随着相关产业的迅速发展, 网络型数据的规模呈指数式增长, 数据样本间的关联关系则导致网络数据分析所需的算力以“双指数”趋势增长. 究其根本, 是信息网络中的“边”为大规模网络分析处理带来了巨大的困难与挑战[1]. 首先, “边”的存在使得在处理关联样本时涉及大量的迭代和遍历计算, 导致组合爆炸; 其次, “边”在有效描述关联关系的同时也使样本间相互耦合, 不利于在网络上实现高效的并行化计算; 第三, 目前的机器学习方法大都以向量空间中的相似性计算为基础, 无法直接应用到网络拓扑空间中.

从网络形成机制维度考虑, 网络产生于一个蕴含相似性表达的隐含向量空间[2]. 如果能将信息网络由拓扑空间嵌入到该隐含向量空间, 将网络节点表示为该空间中的向量, 就可以利用向量间的相似性度量取代网络中的边, 进而克服数据处理中一系列由“边”引发的瓶颈问题. 这一将网络拓扑空间嵌入到隐含向量空间的过程被称为“网络表示学习”或“网络嵌入”. 近年来, 网络表示学习因其在网络大数据分析中的关键性地位, 成为了学术界和产业界的关注焦点[3]. 许多研究工作致力于打破网络中“边”的约束, 为网络中的节点学习一个有效的向量表示, 为下游网络分析任务提供更好的支持, 并且已经在节点分类[4]、网络可视化[5]、链路预测[6]、社区发现[7]等多种任务中发挥了积极作用. 在编码器-解码器视角下, 各种网络表示学习方法都可以归结为对 2 个映射函数不断优化的系统[8]: 一个编码器, 将网络节点映射为低维隐空间中的向量或嵌入; 以及一个解码器, 由学得的低维表示重构网络. 这一系统隐含的基本目标在于, 要求解码器能够尽可能完整的由编码后的低维嵌入恢复原有网络的结构和性质, 以确保该低维嵌入能够提供下游任务所需的全部信息.

从实现策略上来看, 网络表示学习方法主要包含以下 5 种类型: ① 矩阵分解方法[9,10], 源于维度约减技术, 以矩阵形式表达网络节点间的关联, 认为网络中的高维节点表示只与低维隐空间中的少量因素相关, 通过对该矩阵进行分解将高维节点嵌入到低维隐空间中. 此类方法只适用于小型网络, 对于节点数量庞大的网络会面临存储和计算效率较低的问题. ② 随机游走方法[4,11−13], 源于自然语言处理中的词向量学习, 通过有限步的随机游走将信息网络转化为节点序列的集合, 以节点对的出现频率表达它们之间的相关性, 利用上下文节点学习各节点的向量表示, 此类方法在大规模网络中具有较高的效率. ③ 边建模方法, 由节点间的连接关系学习节点表示, 使用的连接关系包括: 节点间的一阶及二阶相似度[5]、边向量[14]、节点连接概率[15,16]等, 这类方法的局限性在于只考虑了有限范围内的节点连接信息, 难以捕捉全局网络结构. ④ 深度学习方法, 使用深度神经网络抽取复杂的结构特性和高度非线性的节点表示, 其中最具代表性的方法是深度图网络[17,18]和图卷积网络 (GCN)[19−22]. ⑤ 混合方法, 综合运用上述方法学习节点表示, 例如文献 [23] 基于度惩罚准则融合谱嵌入和 DeepWalk[4]方法捕捉网络中的宏观结构特性. HARP 算法[24]结合随机游走方法和边建模方法的优点, 能够由原始网络的小规模采样网络中学习节点表示.

早期的网络表示学习研究以无监督设定为主, 即训练数据不包含节点标签信息, 将网络表示学习视为一个独立于下游任务的通用过程. 近年来, 研究者们发现在一些场景中能够获得部分节点标签供网络表示学习使用. 这些类别标签与网络结构和节点属性具有紧密联系, 能对网络表示学习过程产生重要影响, 为此一些半监督网络表示学习方法着眼于利用可获取的部分标签信息提升节点表示的有效性. 这些方法通常将网络表示学习与某一监督学

© 中国科学院软件研究所 http://www.jos.org.cn

习任务相结合, 构建二者的统一优化目标, 以保证获得的节点表示既能表达完整信息, 又能在不同类别间具有可区分性. DDRW[12]在 DeepWalk[4]基础上利用部分节点标签附加支持向量分类优化目标, 为分类任务学习具有判别性的节点表示. SemiNE[25]通过两个阶段在半监督场景下获得节点低维表示: 首先, 利用 DeepWalk 以无监督方式学习节点表示; 接着, 利用标签信息构建一个神经网络对该表示进行修正. Pan 等人[13]将网络表示学习的信息源归为 3 类: 网络结构、节点属性和节点标签, 在一个神经网络框架下利用段落向量模型抽取节点属性和标签信息, 同时通过 DeepWalk 方法学习网络结构信息. 与之类似, LANE 算法[26]基于 3 类信息分别为节点构建相似度矩阵, 并通过谱嵌入方法映射为 3 个低维表示, 利用协方差度量矩阵间的相关性, 进而将 3 个低维表示中的相关信息映射至 1 个联合低维表示中. 文献 [27] 提出一种属性网络的半监督表示学习方法, 首先构建一个用于保持网络上下文结构的目标函数来学习节点嵌入, 再通过一个深度神经网络将节点嵌入和节点属性映射到隐空间中, 对二者进行拼接并以此预测节点标签, 最终通过不断降低预测标签和真实标签间的误差损失训练模型. He 等人[28]将网络表示学习视为将节点间距离信息由原空间映射到低维隐空间的过程, 通过度量学习技术使节点在低维隐空间中的距离与其在原空间中的拓扑距离具有一致性, 同时引入标签信息, 确保具有相同标签的节点具有相近的低维表示, 在此基础上构建了自注意力深度量化模型控制节点表示的规模, 使表示学习的空间和时间效率得到提高.

事实上, 现实网络中除了拓扑结构和节点属性之外, 还蕴含着丰富的伴随信息, 并且这种伴随信息不仅限于作为监督信息的节点标签, 典型的还包括节点间的积极关系和消极关系. 例如, 在电子商务网络中, 用户对他人关于某一商品评价的信任可以抽象为积极关系; 在社交网络中, 不同观点持有者之间的反对或厌恶是一种典型的消极关系. 这些伴随信息反映了网络中复杂微妙的各种关系, 对网络的形成和演化起着不可忽略的作用, 为进一步提升网络表示学习的有效性提供了重要信息. 要实现对这些伴随信息的有效融合和利用, 面临以下 3 个重要问题: ① 如何对伴随信息进行有效的形式化描述, 使网络表示学习的信息源进一步丰富和完善; ② 如何挖掘网络结构、节点属性和伴随信息之间的一致性关系, 实现三者的有效融合; ③ 如何处理表示学习过程中由于训练样本不足及伴随信息有限导致的过拟合问题. 在以往的研究中, 这些问题很少得到关注. 为此, 本文提出了一种融合伴随信息的网络表示学习方法 (network representation learning with integration of accompanying information, NRLIAI) 用于解决上述问题. 该方法以变分自编码器 (variational auto-encoders, VAE)[29]为框架, 包含编码器和解码器两个组成部分.

• 编码器将网络编码为隐空间中的低维表示, 使用图卷积网络 (graph convolutional network, GCN) 融合拓扑结构信息学习网络节点在隐空间中的分布, 并从学得的分布中采样获得节点向量表示.

• 解码器将节点向量表示重构为拓扑空间中的网络结构, 同时利用一个 Softmax 层在隐空间中生成节点类别分布, 并使该分布与先验伴随信息中的相应判别信息不断趋近.

通过设计一致性的优化目标, 使得网络在隐空间中的低维表示能较好保持原有结构和特性, 又能利用伴随信息提供的判别性信息对低维表示的学习过程进行指导, 同时该方法具有一定的生成能力, 能一定程度减轻由于训练样本不足及伴随信息稀疏导致的过拟合. 为验证本文提出模型的有效性, 我们在几个真实网络数据集上进行了节点分类和链路预测实验, 与现有的几种网络表示学习方法进行了对比分析.

本文主要贡献:

(1) 给出了多种类型先验伴随信息的形式化描述, 设计了一种融合伴随信息的网络表示学习模型, 能使网络表示学习过程在保持网络结构和特性的基础上, 充分挖掘伴随信息的辅助指导作用, 构建性能更优的网络向量表示.

(2) 设计了一个具有生成能力的统一优化模型, 实现了网络结构、节点属性以及伴随信息的有效传播和融合, 同时提高了算法在训练样本不足及伴随信息稀疏场景下的抗过拟合能力.

(3) 通过真实网络上的节点分类和链路预测任务, 在多种先验伴随信息条件下将本文提出的方法与几种基线方法进行比较, 实验结果表明, 本文方法能够获得优于其他方法的性能.

本文第 1 节首先介绍网络表示学习的问题定义, 并提出先验伴随信息的形式化描述方式, 在此基础上对融合伴随信息的网络表示学习方法的原理进行详细说明. 第 2 节通过仿真实验对该方法的有效性进行验证. 第 3 节对本文工作进行总结, 并对未来研究工作进行展望.

© 中国科学院软件研究所 http://www.jos.org.cn

1 融合伴随信息的网络表示学习模型 NRLIAI

本文提出融合伴随信息的网络表示学习模型 NRLIAI, 模型总体结构如图 1 所示. 在对先验伴随信息进行形式化描述的基础上, 以变分自编码器为信息处理框架, 由编码器将网络节点及其拓扑关系映射为低维隐空间中的分布, 解码器将分布中的采样向量重构为网络空间结构, 并融合先验伴随信息作用于隐空间中获得的类别分布, 最终构建一致性优化目标获得有效的网络向量表示. 我们首先给出网络表示学习的问题定义, 并提出了先验伴随信息的形式化描述方法, 在此基础上对本文提出的网络表示学习模型 NRLIAI 进行详细说明.

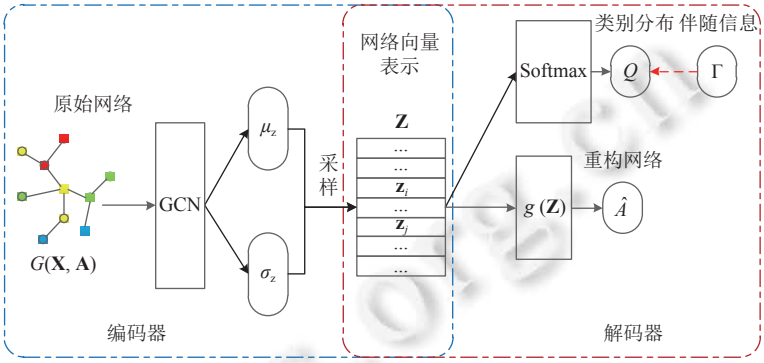


图 1 NRLIAI 模型总体结构

1.1 问题定义

本文关注的问题是构建一个网络表示学习模型, 将网络中的数据映射至低维隐空间中, 为网络节点学习一个有效的低维向量表示, 且这些向量间的相对关系能反应原网络中存在的结构关系和节点属性, 同时与先验伴随信息中蕴含的网络特性保持一致. 形式化定义如下.

定义 1. 网络表示学习. 给定一个无向无权网络 $G(\mathbf{X},\mathbf{A})$, 其中 $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\}$ 表示网络中节点构成的集合, $\mathbf{x}_i\in\mathbb{R}^m\ (1\leqslant i\leqslant N)$ 是第 i 个节点的属性向量, $\mathbf{A}=\left[a_{ij}\right]_{N\times N}$ 为网络的邻接矩阵, 若节点 $\mathbf{x}_i$ 和 $\mathbf{x}_j$ 之间存在连边则矩阵元素 $a_{ij}=1$, 否则 $a_{ij}=0$. 网络表示学习要解决的问题是, 获得一个映射关系 $f:\mathbf{x}_i\rightarrow\mathbf{z}_i\in\mathbb{R}^d\ (d\ll N)$, 将网络节点映射为低维空间中的向量表示, 且向量间的关系尽可能完整的保持原有网络中节点间的关系.

在现实网络中, 除了网络结构和节点属性, 还存在着丰富的伴随信息. 这些伴随信息与网络结构和属性紧密相关, 相互影响彼此的形成; 同时, 二者具有截然不同的信息形式, 前者常常表现为离散的、稀疏的、信息高度浓缩的低维数据, 后者则为以网络结构关联的高维数据. 网络中常见的伴随信息主要包含: 节点标签、节点间约束关系、节点文本信息以及网络语义信息等. 本文讨论的伴随信息主要包括 2 个类型: 标签信息和结对约束关系. 其中, 标签信息包含正标签和负标签, 它们对网络节点的先验类别进行描述, 即节点与类别间的隶属关系; 结对约束关系包含积极关系和消极关系, 是对网络节点彼此之间相互影响的表达. 对于节点文本信息以及网络语义信息等其他形式更为复杂的伴随信息, 若经过处理后可通过节点间关联矩阵进行形式化表达, 原则上都可使用本文提出的 NRLIAI 模型在表示学习过程中融合, 具体处理手段不在本文讨论范围内.

定义 2. 正标签矩阵. 给定一个网络 $G(\mathbf{X},\mathbf{A})$, 其节点包含 K 个类别, 使用标签集合 $\mathbf{L}=\{l_1,l_2,\ldots,l_K\}$ 对这些类别进行标记, 矩阵 $\mathbf{Y}\in\mathbb{R}^{N\times K}$ 称为该网络的正标签矩阵, 用于记录网络节点的正标签信息, 其矩阵元素由公式 (1) 定义.

$$y_{ik}=\begin{cases}1, & \text{若}\mathbf{x}_i\text{的类别标签为}l_k\\ 0, & \text{否则}\end{cases} \tag{1}$$

定义 3. 负标签矩阵. 给定一个网络 $G(\mathbf{X},\mathbf{A})$ 及其节点标签集合 $\mathbf{L}=\{l_1,l_2,\ldots,l_K\}$, 矩阵 $\mathbf{Y}^-\in\mathbb{R}^{N\times K}$ 称为该网络的负标签矩阵, 用于记录网络节点的负标签信息, 其矩阵元素由公式 (2) 定义.

$$y_{ik}^-=\begin{cases}1, & \text{若}\mathbf{x}_i\text{的类别标签不为}l_k\\ 0, & \text{否则}\end{cases} \tag{2}$$

© 中国科学院软件研究所 http://www.jos.org.cn

定义 4. 结对约束矩阵. 给定一个网络$G(\mathbf{X},\mathbf{A})$及其节点标签集合$\mathbf{L}=\{l_1,l_2,\ldots,l_K\}$, 集合$C^+$中的节点之间存在积极关系, 集合$C^-$中的节点之间存在消极关系, 则可使用结对约束矩阵$\mathbf{R}\in\mathbb{R}^{N\times N}$来表达网络中的结对约束关系, 其矩阵元素由公式 (3) 定义.

$$r_{ij}=\begin{cases}1, & 若\left(\mathbf{x}_i,\mathbf{x}_j\right)\in C^+\\ -1, & 若\left(\mathbf{x}_i,\mathbf{x}_j\right)\in C^-\\ 0, & 否则\end{cases} \tag{3}$$

为便于融入网络表示学习的数据处理过程, 在以上定义的基础上, 我们利用一个伴随信息矩阵Γ对上述先验伴随信息进行统一记录, 如公式 (4) 所示.

$$\Gamma=\begin{cases}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}, & 给定正标签信息\\ \dfrac{1}{K-1}\left(\mathbf{Y}^{-}\mathbf{Y}^{-\mathrm{T}}\right), & 给定负标签信息\\ R, & 给定积极或消极关系\end{cases} \tag{4}$$

由公式 (4) 可知, Γ将不同类型的伴随信息统一为关于节点间相对关系的描述.

1.2 NRLIAI 模型的构建

如图 1 所示, 本文提出的 NRLIAI 模型以 VAE 作为信息传播框架, 包含编码器和解码器两个组成部分. 其中, 编码器用于将网络$G(\mathbf{X},\mathbf{A})$映射为低维隐空间中的分布, 解码器用于将分布中采样的向量重构为高维网络关系, 同时融合先验伴随信息对向量表示进行调节和修正.

(1) 编码器

对于信息网络$G(\mathbf{X},\mathbf{A})$, 编码器可形式化的表示为公式 (5) 所示的映射关系.

$$q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{A})=\sum_{i=1}^{N}q(\mathbf{z}_i|\mathbf{X},\mathbf{A}) \tag{5}$$

其中, ϕ为编码器中的待训练参数, $\mathbf{Z}=\{\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_N\}$表示节点向量表示构成的集合, 对于网络中的任意节点$\mathbf{x}_i$, 其在低维隐空间中的向量表示$\mathbf{z}_i$可由公式 (6) 所示的分布$q(\mathbf{z}_i|\mathbf{X},\mathbf{A})$中采样获得.

$$q(\mathbf{z}_i|\mathbf{X},\mathbf{A})=\mathcal{N}\left(\mathbf{z}_i|\mu_i,diag\left(\sigma_i^2\right)\right) \tag{6}$$

其中, μ_i和σ_i^2为$\mathbf{z}_i$所属分布的期望与方差. 图卷积网络能够同时对节点属性和拓扑结构进行端对端学习, 具有广泛适用性和较强的非线性表达能力, 为此我们使用两个图卷积网络为向量表示的分布产生期望和方差, 即$\mu=GCN_\mu(\mathbf{X},\mathbf{A})$和$\log\sigma=GCN_\sigma(\mathbf{X},\mathbf{A})$. 这两个 *GCN* 具有相同的网络结构, 如公式 (7) 所示.

$$GCN(\mathbf{X},\mathbf{A})=Gconv(ReLU(Gconv(\mathbf{A},\mathbf{X};W_0));W_1) \tag{7}$$

其中, $Gconv(\cdot)$为图卷积网络层[21], W_0和W_1分别为第 1 层和第 2 层中的连接权重矩阵, 其中W_0在两个 *GCN* 中是共享的.

(2) 解码器

在解码器中, 由公式 (6) 的分布中采样得到节点的向量表示, 即$\mathbf{z}_i\sim\mathcal{N}\left(\mathbf{z}_i|\mu_i,diag\left(\sigma_i^2\right)\right)$, 并通过一个非线性解码函数将这些低维向量重构为高维网络关系. 与文献 [22] 类似, 本文使用公式 (8) 所示的内积函数作为解码函数.

$$p(\mathbf{A}|\mathbf{Z})=\prod_{i=1}^{N}\prod_{j=1}^{N}p\left(a_{ij}=1|\mathbf{z}_i\mathbf{z}_j\right)=\sigma\left(\mathbf{z}_i^{\mathrm{T}}\mathbf{z}_j\right) \tag{8}$$

同时, 利用一个 Softmax 层计算节点向量表示的分布概率:

$$\mathbf{Q}=\mathrm{Softmax}(\mathbf{Z}) \tag{9}$$

其输出结果$q_{ik}\in\mathbf{Q}$表示$\mathbf{z}_i$属于类别l_k的概率, $\mathbf{Q}$描述了低维隐空间中节点向量表示的类别分布情况.

1.3 模型的优化

为了获得信息网络在低维隐空间中的有效表示, 我们认为学习模型应该包含 2 个优化目标: 第一, 在 VAE 结构中获得的向量表示应当尽可能保持原网络的结构与特性; 第二, 低维隐空间中节点向量表示的类别分布应与先

© 中国科学院软件研究所 http://www.jos.org.cn

验伴随信息中蕴含的判别信息趋于一致.

在 VAE 框架中, 模型的第 1 个优化目标可表达为最大化公式 (10) 所示的证据下界 (evidence lower bound, ELBO)[29].

$$\log p_\phi(\mathbf{X}) \geqslant \mathcal{L}_{\mathrm{ELBO}}(\mathbf{X},\phi) = E_{q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{A})}\left[\log \frac{p(\mathbf{A},\mathbf{Z})}{q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{A})}\right] \tag{10}$$

进一步地, 公式 (10) 可变换为:

$$\begin{aligned}\mathcal{L}_{\mathrm{ELBO}}(\mathbf{X},\phi) &= E_{q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{A})}\left[\log p(\mathbf{A},\mathbf{Z}) - \log q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{A})\right]\\ &= E_{q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{A})}\left[\log \frac{p(\mathbf{A},\mathbf{Z})}{p(\mathbf{Z})} + \log p(\mathbf{Z}) - \log q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{A})\right]\\ &= E_{q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{A})}\left[\log p(\mathbf{A}|\mathbf{Z}) - \log \frac{q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{A})}{p(\mathbf{Z})}\right]\\ &= E_{q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{A})}\left[\log p(\mathbf{A}|\mathbf{Z})\right] - \mathrm{KL}\left[q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{A}) \,\|\, p(\mathbf{Z})\right]\end{aligned} \tag{11}$$

其中, $\mathrm{KL}[q(\cdot)\|p(\cdot)]$表示两个分布间的 Kullback-Leibler (KL) 散度. 由此, 我们将第 1 个优化目标$\mathcal{L}_{\mathrm{VAE}}$定义为:

$$\mathcal{L}_{\mathrm{VAE}} = E_{q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{A})}\left[\log p(\mathbf{A}|\mathbf{Z})\right] - \mathrm{KL}\left[q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{A}) \,\|\, p(\mathbf{Z})\right] \tag{12}$$

其中, $q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{A})$和$p(\mathbf{A}|\mathbf{Z})$分别由公式 (5) 和公式 (8) 定义, 假设 $\mathbf{Z}$ 的先验分布为高斯分布, 即 $p(\mathbf{Z}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{z}_i|0,\mathbf{I})$, 在训练时可利用重参数技巧[29]进行梯度的反向传播.

NRLIAI 模型的第 2 个优化目标要求网络节点在低维隐空间中的类别分布与先验伴随信息中蕴含的判别信息趋于一致, 为此我们使用互信息定义这一优化目标.

$$\mathcal{L}_\Gamma = I(\Gamma;\mathbf{Q}) \tag{13}$$

$I(\Gamma;\mathbf{Q})$表示伴随信息Γ与后验类别分布$\mathbf{Q}$之间的互信息, 用于度量二者之间的统计相关性. 公式 (13) 的取值越大, 表明节点向量表示的类别分布对伴随信息的依赖性越强, 二者的一致性程度就越高. 依据互信息定义, 可将公式 (13) 进一步变换为:

$$\mathcal{L}_\Gamma = \mathrm{KL}(p(\Gamma|\mathbf{Q})\|p(\Gamma)) = \sum_{q\in Q}\sum_{\tau\in\Gamma} p(q)\,p(\tau|q)\log\frac{p(\tau|q)}{p(\tau)} = \sum_{k=1}^{K}\sum_{i,j=1}^{N} q_{ik}\hat{\tau}_{ij}\log\frac{\hat{\tau}_{ij}}{\tau_{ij}} \tag{14}$$

其中, q_{ik}为 Softmax 层输出的分布概率, τ_{ij}为伴随信息矩阵Γ中的元素, $\hat{\tau}_{ij}$由公式 (15) 计算.

$$\hat{\tau}_{ij} = \sum_{k=1}^{K} q_{ik}q_{jk} \tag{15}$$

最终, 我们将 NRLIAI 模型的优化目标定义为:

$$\max_\phi \mathcal{L} = \max_\phi \left[\gamma\mathcal{L}_{\mathrm{VAE}} + (1-\gamma)\mathcal{L}_\Gamma\right] \tag{16}$$

其中, 参数$\gamma\in(0,1)$用于调整$\mathcal{L}_{\mathrm{VAE}}$和$\mathcal{L}_\Gamma$对于学习过程的影响. 若$\gamma=1$, 则网络中的伴随信息不会被融合到模型的信息传播过程中, 此时模型退化为一个变分图自编码器 (variational graph auto-encoders, VGAE)[22].

1.4 算法实施流程

我们使用 TensorFlow 提供的 Adam-Optimizer 优化器实现对式 (16) 的优化, 为了提升在大规模网络上的训练效率, 我们将网络数据集划分为一系列批处理集合, 引入 mini-batch 策略进行模型参数更新. 模型训练完成后, 使用确定参数后的编码器将网络数据映射至隐空间中的分布, 由该分布中采样得到的向量即为融合伴随信息的网络表示学习结果. NRLIAI 模型的算法执行流程总结如算法 1.

算法 1. 融合伴随信息的网络表示学习模型 NRLIAI.

输入: 网络$G(\mathbf{X},\mathbf{A})$, 节点类别数 K, 伴随信息Γ, 参数γ, 迭代次数 L;
输出: 网络向量表示$\mathbf{Z}=\{\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_N\}$.

© 中国科学院软件研究所 http://www.jos.org.cn

1. 使用预训练网络对编码器的连接权重 W_0 和 W_1 进行初始化
2. For $l = 1,\ldots,L$ do
3. For $i = 1,\ldots,N$ do
4. $\mu_i = GCN_{\mu}(\mathbf{x}_i, \mathbf{A}_i)$
5. $\log\sigma_i = GCN_{\sigma}(\mathbf{x}_i, \mathbf{A}_i)$
6. $q(\mathbf{z}_i|\mathbf{X},\mathbf{A}) = \mathcal{N}\left(\mathbf{z}_i|\mu_i, diag\left(\sigma_i^2\right)\right)$, 并从中采样获得 $\mathbf{z}_i$
7. End for
8. $p(\mathbf{A}|\mathbf{Z}) = \sigma\left(\mathbf{z}_i^{\mathrm{T}}\mathbf{z}_j\right)$
9. 利用公式 (12) 计算 $\mathcal{L}_{\mathrm{VAE}}$
10. $\mathbf{Q} = \mathrm{Softmax}(\mathbf{Z})$
11. 利用公式 (14) 计算 $\mathcal{L}_{\Gamma}$
12. $\mathcal{L} = \gamma\mathcal{L}_{\mathrm{VAE}} + (1-\gamma)\mathcal{L}_{\Gamma}$
13. 反向传播并更新模型连接权重
14. End for
15. 利用训练完成的模型获得网络的低维向量表示

相比现有的其他网络表示学习方法, NRLIAI 模型主要具备优势有两个.

(1) 模型可以通过融合不同类型的伴随信息, 为网络表示学习过程提供附加判别信息. 相比无监督学习方式, 伴随信息的引入使获得的网络表示能更好地保持原有网络的结构和特性; 相比传统的面向节点分类任务的半监督网络表示学习方法, NRLIAI 模型能对节点标签以外的伴随信息进行融合, 因而对于多种下游任务有更强的适用性.

(2) 模型以 VAE 作为信息处理和传播的框架, 具有一定的生成能力, 因此, 模型对样本噪声的容忍度以及训练样本不足时的抗过拟合能力都得到了提高.

2 实验与结果

我们在几个真实网络数据集上利用本文提出的 NRLIAI 模型学习网络的向量表示, 并将其用于节点分类和链路预测任务, 通过与几种基准方法进行比较, 对 NRLIAI 模型的有效性和先进性进行了验证.

2.1 评价指标

对于节点分类任务, 我们使用 Micro-$F1$ 分数和 Macro-$F1$ 分数作为节点分类结果的评价指标, 其定义如公式 (17) 和公式 (18) 所示.

$$\text{Micro}-F1 = \frac{2\times\sum_{k=1}^{K} precision(k)\times\sum_{k=1}^{K} recall(k)}{\sum_{k=1}^{K} precision(k)+\sum_{k=1}^{K} recall(k)} \tag{17}$$

$$\text{Macro}-F1 = \frac{\sum_{k=1}^{K} F1(k)}{K} \tag{18}$$

其中, $F1(k)$ 表示分类结果中第 k 个类别的 $F1$ 分数, 定义如公式 (19) 所示.

$$F1(k) = \frac{2\times precision(k)\times recall(k)}{precision(k)+recall(k)} \tag{19}$$

$precision(k)$ 和 $recall(k)$ 表示分类结果中第 k 个类别的准确率和召回率. 相对来说, Micro-$F1$ 更加关注数据集中规模较大的类上的分类结果, 而 Macro-$F1$ 则是对全部类别上分类结果的平均水平进行评价.

© 中国科学院软件研究所 http://www.jos.org.cn

对于链路预测任务, 我们随机抽取网络中的一部分连边作为测试样例, 使用网络的剩余部分训练学习模型, 对模型学得的节点向量表示, 通过余弦相似度计算节点间存在连边的概率. 将训练数据中存在的边排除后, 以存在概率最大的 k 条边作为链路预测结果. 与文献 [18] 的评价方式类似, 我们使用前 k 精确率 (*Precision@k*) 和平均精确率 (mean average precision, *MAP*) 对链路预测结果进行评价. 将被抽取的边集记作 E', 使用 e_{ij} 表示节点 $\mathbf{x}_i$ 和 $\mathbf{x}_j$ 之间的连边, *Precision@k* 可定义为:

$$Precision@k(i)=\frac{\left|\left\{e_{ij}|\mathbf{x}_i,\mathbf{x}_j\in\mathbf{X},index\left(e_{ij}\right)<k,\Delta_{ij}=1\right\}\right|}{k} \tag{20}$$

其中, $index\left(e_{ij}\right)$ 表示 e_{ij} 在链路预测结果中的排序序号, 若 e_{ij} 存在于被隐藏的边集中则 $\Delta_{ij}=1$, 否则 $\Delta_{ij}=0$. 由公式 (20) 可知, *Precision@k* 指标对排序靠前的预测结果较为关注. 与之不同, *MAP* 更加关注预测结果的整体水平, 其定义如下:

$$AP(i)=\frac{\sum_{\mathbf{x}_j\in\mathbf{X}}Precision@j(i)\times\Delta_{ij}}{|E'|} \tag{21}$$

$$MAP=\frac{\sum_{x_i\in\mathbf{X}_Q}AP(i)}{\left|\mathbf{X}_Q\right|} \tag{22}$$

其中, $\mathbf{X}_Q$ 为查询集合.

2.2 对比方法及数据集

在节点分类和链路预测两个任务中, 我们选取了几种典型的网络表示学习方法与 NRLIAI 模型进行比较, 这些方法总体上可分为两类: 一类是无监督方法, 包括利用深度神经网络进行节点局部邻域信息压缩的 SDNE 方法[18], 融合网络结构和节点属性进行网络表示学习的 SNEA 方法[16], 以及以将变分自编码器架构拓展到图数据处理的 VGAE 模型[22]; 另一类是半监督方法, 包括以一阶 Chebyshev 多项式作为卷积核在图上进行特征提取的 Semi-GCN 模型[21], 在 DeepWalk 基础上附加支持向量分类优化目标的 DDRW 方法[12], 将网络结构、节点属性和类别标签分别处理再进行融合的 TriDNR 方法[13], 以及通过度量学习手段融合网络结构、节点属性和类别标签并确保节点在原空间与低维隐空间中具有一致性距离的 SNEQ 方法[28]. 在对比实验中, 这些方法的参数设置都与其作者在相关文献中的最优设置保持一致, 学习率统一设置为 0.01. NRLIAI 模型中编码器的参数设置与 VGAE 模型保持一致, 参数 γ 设置为 $\gamma=0.5$. 对于所有方法, 网络向量表示的维度都设为 d=128. 对上述每种算法, 在实验中重复执行 10 次并取其均值作为最终结果.

本文使用的真实网络数据集包括: 引文数据集 DBLP, 消费者评论数据集 Epinions 以及电影评论数据集 Flixster, 这些数据集的基本统计特性记录在表 1 中.

表 1 网络数据集统计特性

数据集	节点数	连边数	类别数	节点属性	密度 (%)
DBLP	23 624	210 566	8	是	0.165 7
Epinions	18 352	290 568	5	是	0.101 6
Flixster	31 435	380 447	8	是	0.113 8

DBLP 数据集来源于一个计算机领域的英文文献集成数据库系统, 我们抽取了该系统中 10 个研究方向的 23 624 名作者 10 年间发表的期刊和会议论文信息, 以作者作为网络节点, 基于作者间的合著关系构建网络, 以每个作者发表文献最多的领域作为其所属类别, 以作者间的引用作为节点间的积极关系.

Epinions 数据集来源于一个商品评论网站, 该网站记录了用户对商品的评价信息和信任关系. 我们从该网站数据中抽取了 18 352 名用户对 5 类商品的评价记录构建网络, 将用户作为网络节点, 为对相同商品做出评价的用户间建立连边, 以每个用户做出评价最多的商品所属类别作为该用户的类别, 将用户对他人评价的信任和否定态

© 中国科学院软件研究所 http://www.jos.org.cn

度作为节点间的积极关系和消极关系.

Flixster 数据集来源于一个电影社交网站, 该网站允许用户分享电影评级, 并与具有相同观影偏好的用户建立社交关系. 我们从中抽取 31435 名用户对 8 个影片类别做出的评级信息, 以用户间的社交关系构建网络, 将用户发表评级最多的影片类别作为用户的所属类别, 将用户对影评的认同和否定作为节点间的积极关系和消极关系.

2.3 实验结果分析

对于每个网络分析任务, 我们给定 3 种先验伴随信息: 正标签、负标签以及结对约束关系. 由于实验中选用的方法对于不同类型的先验信息处理能力不同, 我们在实验中采用如下设置: 在对标签信息的处理上, 3 种无监督方法 SDNE、SNEA 和 VGAE 无法直接融合节点标签, 因此将标签信息视为节点属性进行处理; 在对结对约束关系的处理上, 4 种半监督方法 Semi-GCN、DDRW、TriDNR 和 SNEQ 将结对关系作为类别标签加以利用, 即将具有积极关系的节点视为相同类别, 将具有消极关系的节点视为不同类别; 对于无监督方法 SDNE 和 VGAE, 将结对关系作为类别标签转化为节点属性进行处理; SNEA 具备处理有向网络的能力, 在实验中将具有积极关系的节点间连边视为正连接, 将具有消极关系的节点间连边视为负连接.

2.3.1 节点分类任务

我们首先将正标签作为先验伴随信息, 设置标注比例为 10%, 在不同训练集规模下利用各种网络表示学习方法得到的向量表示进行节点分类, 分类结果如表 2 所示. 从总体上来看, 随着训练集规模的提高, 各种算法获得的分类结果都有所提升, 其中, NRLIAI 模型获得的分类结果较为稳定, 始终保持在较高水平. 而在相同条件下, 相比无监督方法, 4 种半监督方法以及 NRLIAI 模型的节点分类结果具有明显优势, 这主要是由于无监督方法对于标签信息的利用不足导致, 表明对标签信息的充分利用能提升网络表示学习对分类任务的有效性. DDRW 和 TriDNR 都使用随机游走策略获取网络的结构信息, 本质上属于浅层嵌入方法, 其网络表示学习过程以结构保持为主, 对于节点属性的保持能力比较有限. SNEQ 使用深度神经网络将表示学习任务转换为节点间距离的度量学习, 使用节点间的拓扑距离作为原空间中的节点距离度量, 忽视了属性信息对网络形成机制的重要作用, 未实现节点属性和拓扑结构的有效融合. 相比之下, Semi-GCN 利用图卷积操作能更有效地在节点邻域范围内融合网络结构和节点属性, 获得了优于上述 3 种半监督方法的分类结果. NRLIAI 同样使用了图卷积算子进行信息聚合, 但由于具有一定的生成能力, 因此在训练集规模较少时获得了优于其他方法的分类结果.

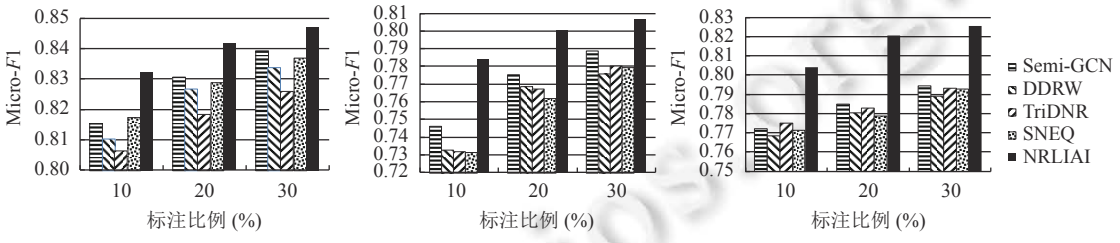
为考察算法对正标签的利用情况, 我们令训练集规模为 50%, 分别设置标注比例为 10%、20% 和 30%, 比较不同算法获得的节点分类结果. 由于无监督方法对标签信息的利用极为有限, 标签比例的变化不会对其产生显著影响, 因此只将 NRLIAI 与 4 种半监督方法进行比较, 实验结果如图 2 所示. 由分类结果可知, 随着标注数量的增大, 几种网络表示学习方法在节点分类任务中的表现都有所提升, 表明正标签是网络中的重要信息, 在学习网络表示的过程中融入更多的正标签对于保持网络结构和特性具有积极作用. 此外, 在不同先验标注比例下 NRLIAI 的表现都优于其他方法, 表明了该方法在融合正标签信息以及寻求网络结构、节点属性和类别标签三者间的一致性方面具有一定优势.

接着我们依据数据集中节点与类别间的关系随机生成一部分负标签, 与正标签共同作为先验伴随信息, 比较各种算法获得的分类结果. 考虑到无监督方法不具备区分正负标签的能力, 负标签的加入不会对其分类结果产生显著影响, 因此我们将 NRLIAI 模型与 4 种半监督方法进行比较. 设置正负标签的总体标注比例为 10%, 各算法在不同训练集规模下获得的分类结果如表 3 所示. 设定训练集规模为 50%, 在不同正负标签总体标注比例情况下, 各算法获得的节点分类结果如图 3 所示. 由上述实验结果可以发现, 当标签总体标注比例相同时, 负标签的引入导致半监督方法的节点分类表现相比只给定正标签时有所下降. 这主要是由于这些方法缺乏有效的负标签处理手段, 因而无法对负标签进行有效融合. 而 NRLIAI 模型能从正负标签中抽取判别信息并为网络表示学习过程提供指导和约束, 使节点分类的表现优于其他方法. 此外, 随着总体标注比例的提高, 正标签的数量也随之增多, 半监督方法在节点分类任务上的表现有所改善. 在给定不同总体标注比例的情况下, NRLIAI 模型都可以获得优于其他方法的节点分类结果, 并且相比只给定相同比例正标签的情况, 其节点分类表现是比较一致的, 这表明 NRLIAI 模型对于负标签的融合处理是同样有效的.

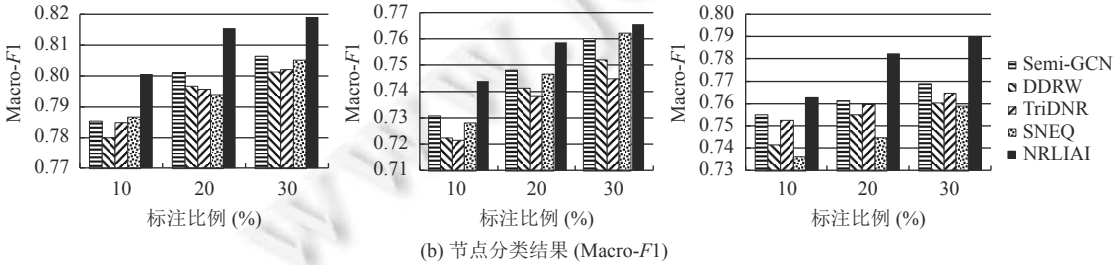
© 中国科学院软件研究所 http://www.jos.org.cn

表 2 给定 10% 正标签时不同算法的节点分类结果

指标	训练集规模 (%)	数据集	SDNE	SNEA	VGAE	Semi-GCN	DDRW	TriDNR	SNEQ	NRLIAI
Micro-*F*1	10	DBLP	0.6612	0.6603	0.6941	0.7424	0.7374	0.7458	0.7352	**0.7635**
	20		0.7175	0.7088	0.7421	0.7842	0.7814	0.7816	0.7733	**0.8161**
	30		0.7224	0.7117	0.7463	0.7937	0.7899	0.7885	0.7862	**0.8203**
	40		0.7292	0.7332	0.7531	0.8055	0.7964	0.7968	0.7943	**0.8276**
	50		0.7447	0.7555	0.7654	0.8153	0.8103	0.8062	0.8174	**0.8322**
	10	Epinions	0.6426	0.6342	0.6405	0.6664	0.6815	0.6804	0.6548	**0.7016**
	20		0.6861	0.6855	0.6878	0.7006	0.7223	0.7178	0.7101	**0.7592**
	30		0.6945	0.7020	0.7132	0.7145	0.7359	0.7318	0.7265	**0.7607**
	40		0.7004	0.7081	0.7206	0.7285	0.7394	0.7380	0.7446	**0.7748**
	50		0.7143	0.7112	0.7263	0.7462	0.7427	0.7416	0.7512	**0.7841**
	10	Flixster	0.6482	0.6377	0.6532	0.6908	0.6849	0.6917	0.6883	**0.7367**
	20		0.6835	0.6926	0.7025	0.7318	0.7291	0.7334	0.7279	**0.7735**
	30		0.6882	0.7016	0.7164	0.7441	0.7454	0.7513	0.7356	**0.7831**
	40		0.6914	0.7159	0.7200	0.7593	0.7610	0.7669	0.7675	**0.7896**
	50		0.7054	0.7208	0.7285	0.7723	0.7682	0.7750	0.7712	**0.8039**
Macro-*F*1	10	DBLP	0.6417	0.6316	0.6627	0.7122	0.7261	0.7172	0.7060	**0.7526**
	20		0.6765	0.6704	0.7003	0.7511	0.7526	0.7603	0.7436	**0.7820**
	30		0.6813	0.6791	0.7071	0.7631	0.7682	0.7754	0.7671	**0.7954**
	40		0.6918	0.6882	0.7183	0.7724	0.7760	0.7816	0.7805	**0.7972**
	50		0.7036	0.6902	0.7210	0.7855	0.7802	0.7849	0.7867	**0.8006**
	10	Epinions	0.6173	0.6120	0.6191	0.6524	0.6588	0.6541	0.6452	**0.6882**
	20		0.6702	0.6635	0.6715	0.6976	0.6895	0.6914	0.6834	**0.7268**
	30		0.6811	0.6825	0.6884	0.7146	0.7048	0.7063	0.7126	**0.7315**
	40		0.6942	0.6947	0.7019	0.7285	0.7161	0.7113	0.7248	**0.7406**
	50		0.7052	0.7004	0.7171	0.7308	0.7225	0.7216	0.7281	**0.7441**
	10	Flixster	0.6257	0.6302	0.6282	0.6781	0.6614	0.6726	0.6624	**0.7158**
	20		0.6615	0.6883	0.6742	0.7228	0.7134	0.7205	0.6996	**0.7452**
	30		0.6852	0.6997	0.6987	0.7369	0.7296	0.7366	0.7138	**0.7533**
	40		0.6904	0.7062	0.7110	0.7477	0.7340	0.7451	0.7223	**0.7597**
	50		0.6988	0.7176	0.7168	0.7552	0.7417	0.7524	0.7346	**0.7646**



(a) 节点分类结果 (Micro-*F*1)



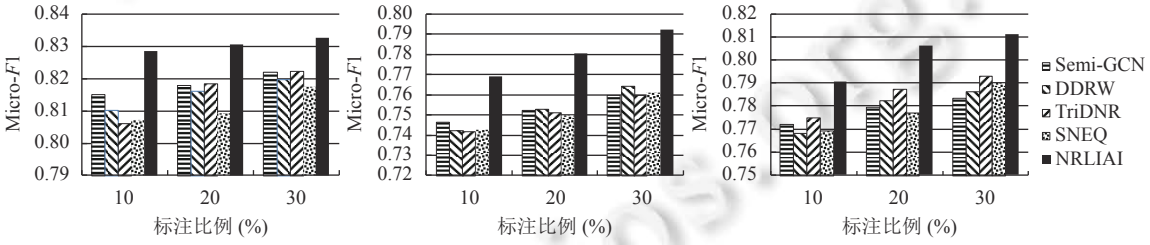
(b) 节点分类结果 (Macro-*F*1)

图 2 不同正标签标注比例下的节点分类结果

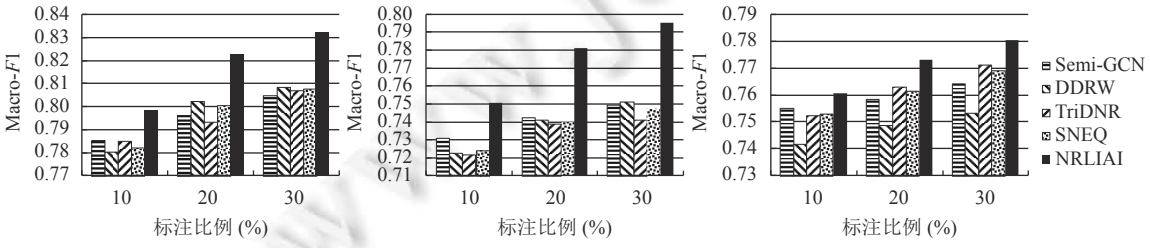
© 中国科学院软件研究所 http://www.jos.org.cn

表 3 给定 10% 正标签和负标签时不同算法的节点分类结果

指标	训练集规模 (%)	数据集	Semi-GCN	DDRW	TriDNR	SNEQ	NRLIAI
Micro-*F*1	10	DBLP	0.7424	0.7374	0.7458	0.7355	**0.7592**
	20		0.7842	0.7814	0.7816	0.7804	**0.8087**
	30		0.7937	0.7899	0.7885	0.7917	**0.8169**
	40		0.8055	0.7964	0.7968	0.7986	**0.8213**
	50		0.8153	0.8103	0.8062	0.8072	**0.8285**
	10	Epinions	0.6664	0.6815	0.6804	0.6726	**0.7046**
	20		0.7006	0.7223	0.7178	0.7257	**0.7425**
	30		0.7145	0.7359	0.7318	0.7295	**0.7518**
	40		0.7285	0.7394	0.7380	0.7332	**0.7634**
	50		0.7462	0.7427	0.7416	0.7428	**0.7689**
	10	Flixster	0.6908	0.6849	0.6917	0.6873	**0.7290**
	20		0.7318	0.7291	0.7334	0.7255	**0.7677**
	30		0.7441	0.7454	0.7513	0.7409	**0.7713**
	40		0.7593	0.7610	0.7669	0.7562	**0.7798**
	50		0.7723	0.7682	0.7750	0.7695	**0.7905**
Macro-*F*1	10	DBLP	0.7202	0.7161	0.7172	0.7061	**0.7412**
	20		0.7511	0.7526	0.7603	0.7544	**0.7769**
	30		0.7631	0.7682	0.7754	0.7620	**0.7836**
	40		0.7724	0.7760	0.7816	0.7741	**0.7925**
	50		0.7855	0.7802	0.7849	0.7823	**0.7984**
	10	Epinions	0.6524	0.6588	0.6541	0.6307	**0.6907**
	20		0.6976	0.6895	0.6914	0.6778	**0.7186**
	30		0.7146	0.7048	0.7063	0.6976	**0.7275**
	40		0.7285	0.7161	0.7113	0.7155	**0.7318**
	50		0.7308	0.7225	0.7216	0.7243	**0.7406**
	10	Flixster	0.6781	0.6614	0.6726	0.6733	**0.7071**
	20		0.7228	0.7134	0.7205	0.7194	**0.7336**
	30		0.7369	0.7296	0.7366	0.7388	**0.7503**
	40		0.7477	0.7340	0.7451	0.7441	**0.7565**
	50		0.7552	0.7417	0.7524	0.7530	**0.7608**



(a) 节点分类结果 (Micro-*F*1)



(b) 节点分类结果 (Macro-*F*1)

图 3 不同正负标签标注比例下的节点分类结果

© 中国科学院软件研究所 http://www.jos.org.cn

将节点间的积极关系和消极关系作为先验伴随信息, 我们同样在设定不同训练集规模和先验信息标注比例的情况下对各种算法获得的节点分类结果进行了比较. 当给定的结对约束关系标注比例为 10% 时, 各算法在不同训练集规模下获得的节点分类结果如表 4 所示. 设定训练集规模为 50%, 给定不同标注比例的结对约束关系, 各算法获得的节点分类结果如后文图 4 所示. 由这些实验结果可知, 对于 SDNE 和 VGAE 而言, 给定结对约束关系时获得的节点分类结果与给定节点标签时较为接近, 这是由于这些方法无法充分利用伴随信息中蕴含的关键判别信息, 难以对网络表示学习过程起到促进作用. SNEA 算法在对节点属性和网络结构进行融合时考虑了积极关系和消极关系对节点间相似性度量的不同影响, 对具有不同约束关系的节点采用不同的处理策略, 能获得有效性较高的网络表示. 4 种半监督方法的节点分类表现优于 SDNE 和 VGAE, 但相比给定节点标签时的自身表现有一定差距, 并且当结对约束关系标注比例增加时, 个别半监督方法出现分类性能变差的情况. 这主要是由于这些半监督方法将结对约束关系作为类别标签处理, 虽然能在一定程度上利用其中蕴含的辅助信息, 但对于分类任务而言, 结对约束关系与类别标签在判别性描述的有效性上存在差别. 相比其他方法, 使用 NRLIAI 模型学得的网络表示在节点分类任务中表现更优, 这表明 NRLIAI 能够高效地对网络结构、节点属性和结对约束关系进行融合并从中抽取有用信息, 进而提高网络表示的有效性.

表 4　给定 10% 结对约束关系时不同算法的节点分类结果

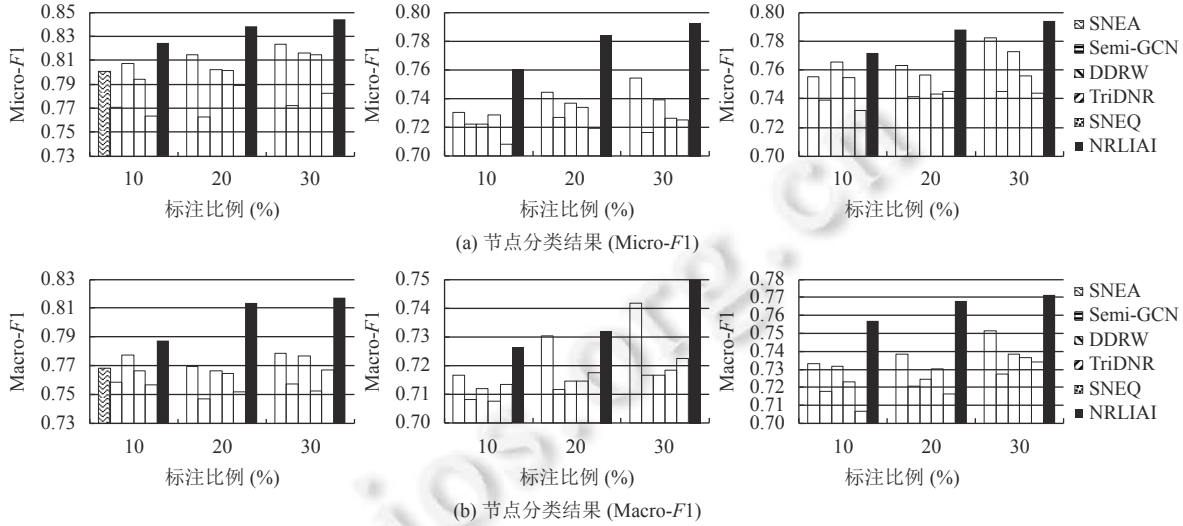
指标	训练集规模 (%)	数据集	SDNE	SNEA	VGAE	Semi-GCN	DDRW	TriDNR	SNEQ	NRLIAI
Micro-*F*1	10	DBLP	0.6601	0.7263	0.6935	0.7091	0.7245	0.7121	0.7144	**0.7515**
	20		0.7055	0.7613	0.7422	0.7472	0.7572	0.7514	0.7348	**0.7963**
	30		0.7216	0.7751	0.7454	0.7505	0.7634	0.7635	0.7556	**0.8142**
	40		0.7278	0.7913	0.7517	0.7587	0.7826	0.7814	0.7660	**0.8180**
	50		0.7410	0.8005	0.7642	0.7703	0.8073	0.7942	0.7633	**0.8246**
	10	Epinions	0.6393	0.6633	0.6388	0.6533	0.6635	0.6642	0.6628	**0.6924**
	20		0.6861	0.7041	0.6762	0.6912	0.7088	0.7002	0.6941	**0.7365**
	30		0.6924	0.7152	0.7025	0.7004	0.7120	0.7073	0.7059	**0.7400**
	40		0.6998	0.7224	0.7103	0.7173	0.7167	0.7215	0.7087	**0.7524**
	50		0.7105	0.7297	0.7158	0.7217	0.7213	0.7278	0.7076	**0.7603**
	10	Flixster	0.6412	0.6966	0.6537	0.6824	0.6825	0.6746	0.6873	**0.7135**
	20		0.6786	0.7253	0.7021	0.7180	0.7232	0.7074	0.6952	**0.7525**
	30		0.6837	0.7287	0.7154	0.7214	0.7360	0.7225	0.7008	**0.7590**
	40		0.6901	0.7386	0.7195	0.7347	0.7571	0.7317	0.7236	**0.7662**
	50		0.6995	0.7554	0.7281	0.7388	0.7652	0.7544	0.7318	**0.7717**
Macro-*F*1	10	DBLP	0.6412	0.7017	0.6630	0.7034	0.7045	0.6972	0.7042	**0.7293**
	20		0.6741	0.7408	0.6995	0.7369	0.7403	0.7337	0.7314	**0.7618**
	30		0.6783	0.7524	0.7057	0.7474	0.7566	0.7380	0.7386	**0.7730**
	40		0.6902	0.7644	0.7171	0.7539	0.7717	0.7525	0.7440	**0.7812**
	50		0.7013	0.7682	0.7201	0.7583	0.7774	0.7661	0.7563	**0.7876**
	10	Epinions	0.6108	0.6465	0.6188	0.6388	0.6362	0.6307	0.6244	**0.6865**
	20		0.6667	0.6840	0.6505	0.6641	0.6878	0.6773	0.6792	**0.7003**
	30		0.6780	0.6931	0.6678	0.6820	0.6930	0.6846	0.6748	**0.7084**
	40		0.6914	0.7107	0.6914	0.6965	0.6994	0.7014	0.7045	**0.7122**
	50		0.7013	0.7165	0.7062	0.7077	0.7118	0.7073	0.7131	**0.7263**
	10	Flixster	0.6225	0.6774	0.6277	0.6662	0.6551	0.6545	0.6533	**0.7005**
	20		0.6603	0.7001	0.6702	0.6944	0.6992	0.6892	0.6782	**0.7293**
	30		0.6824	0.7086	0.6927	0.7005	0.7127	0.7003	0.6947	**0.7364**
	40		0.6887	0.7241	0.7100	0.7137	0.7261	0.7171	0.6981	**0.7453**
	50		0.6942	0.7332	0.7143	0.7176	0.7315	0.7228	0.7069	**0.7568**

2.3.2　链路预测任务

链路预测是常用于衡量网络表示学习性能的另一个下游任务. 在该任务中, 我们首先设置先验伴随信息的标

© 中国科学院软件研究所 http://www.jos.org.cn

注比例为 10%, 由每个数据集抽取 20% 的连边作为测试样例, 利用网络剩余部分训练模型. 以 *Precision@k* 作为评价指标, 表 5 给出了 *k* 分别取 10, 100, 200, 300, 500 和 1 000 时各种算法获得的链路预测结果, *P@k* 即为 *Precision@k*. 从表中结果可以看出, NRLIAI 在大多数情况下都能获得优于其他算法的表现. 值得注意的是, 对于 NRLIAI 模型, 给定结对约束关系作为先验伴随信息时获得的链路预测结果略优于给定节点标签时的结果, 这在一定程度上反映了对于链路预测任务, 节点间的积极关系和消极关系蕴含着相比类别标签更为有效的辅助信息.



(a) 节点分类结果 (Micro-*F*1)

(b) 节点分类结果 (Macro-*F*1)

图 4 不同结对约束关系标注比例下的节点分类结果

表 5 给定 10% 先验伴随信息时不同算法的链路预测结果

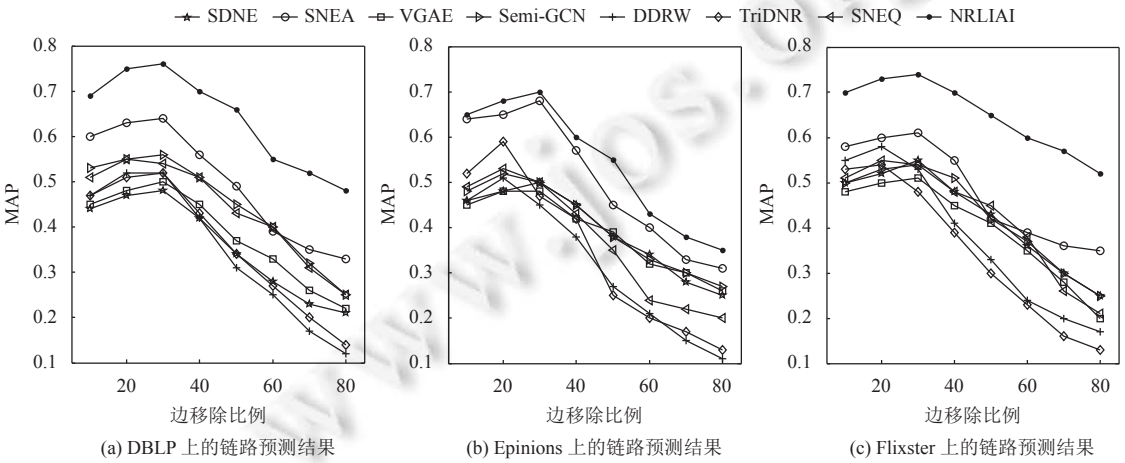
伴随信息	数据集	*P@k*	SDNE	SNEA	VGAE	Semi-GCN	DDRW	TriDNR	SNEQ	NRLIAI
正标签	DBLP	*P*@10	0.7	0.8	0.8	**1**	0.9	**1**	**1**	**1**
		P@100	0.83	0.82	0.85	0.90	0.92	0.91	0.89	**0.95**
		P@200	0.81	0.74	0.80	0.88	0.88	0.83	0.87	**0.91**
		P@300	0.72	0.70	0.74	0.82	0.79	0.77	0.80	**0.87**
		P@500	0.66	0.63	0.69	0.75	0.67	0.69	0.74	**0.80**
		P@1000	0.61	0.59	0.62	0.67	0.61	0.64	0.68	**0.76**
	Epinions	*P*@10	0.7	0.8	0.8	**1**	**1**	**1**	0.9	**1**
		P@100	0.83	0.81	0.83	0.87	0.90	0.93	0.88	**0.93**
		P@200	0.75	0.77	0.77	0.81	0.79	**0.83**	0.82	0.82
		P@300	0.68	0.71	0.72	0.72	0.70	**0.78**	0.74	0.76
		P@500	0.66	0.63	0.63	0.65	0.63	0.70	0.67	**0.73**
		P@1000	0.59	0.57	0.58	0.62	0.62	0.65	0.60	**0.71**
	Flixster	*P*@10	0.7	0.7	0.8	0.9	**1**	0.9	0.9	**1**
		P@100	0.79	0.80	0.83	0.89	0.86	0.88	0.85	**0.93**
		P@200	0.76	0.74	0.77	0.82	0.84	0.81	0.82	**0.89**
		P@300	0.67	0.66	0.67	0.77	0.78	0.74	0.76	**0.85**
		P@500	0.57	0.62	0.64	0.69	0.72	0.70	0.71	**0.81**
		P@1000	0.55	0.60	0.58	0.65	0.67	0.63	0.66	**0.76**
正标签和负标签	DBLP	*P*@10	0.7	0.8	0.8	0.9	0.8	0.8	0.9	**1**
		P@100	0.79	0.83	0.84	0.82	0.81	0.85	0.84	**0.97**
		P@200	0.74	0.75	0.79	0.77	0.77	0.78	0.80	**0.91**
		P@300	0.71	0.68	0.73	0.73	0.73	0.75	0.73	**0.85**
		P@500	0.65	0.64	0.70	0.67	0.63	0.62	0.67	**0.80**
		P@1000	0.53	0.57	0.63	0.63	0.59	0.58	0.61	**0.77**

© 中国科学院软件研究所 http://www.jos.org.cn

表 5 给定 10% 先验伴随信息时不同算法的链路预测结果 (续)

伴随信息	数据集	*P@k*	SDNE	SNEA	VGAE	Semi-GCN	DDRW	TriDNR	SNEQ	NRLIAI
	Epinions	*P*@10	0.7	0.8	0.8	0.8	0.8	0.8	0.9	**1**
		P@100	0.82	0.79	0.83	0.81	0.82	0.85	0.86	**0.91**
		P@200	0.74	0.74	0.75	0.77	0.77	0.80	**0.81**	0.80
		P@300	0.66	0.68	0.70	0.71	0.73	0.73	0.75	**0.77**
		P@500	0.63	0.61	0.62	0.67	0.65	0.68	0.68	**0.73**
		P@1000	0.55	0.55	0.57	0.62	0.62	0.60	0.63	**0.69**
	Flixster	*P*@10	0.7	0.8	0.8	0.8	0.9	0.8	0.9	**1**
		P@100	0.80	0.78	0.82	0.81	0.85	0.83	0.88	**0.90**
		P@200	0.72	0.70	0.74	0.74	0.78	0.81	0.82	**0.86**
		P@300	0.65	0.67	0.67	0.70	0.74	0.77	0.74	**0.82**
		P@500	0.58	0.58	0.63	0.63	0.67	0.69	0.70	**0.77**
		P@1000	0.53	0.52	0.55	0.60	0.62	0.62	0.63	**0.71**
结对约束关系	DBLP	*P*@10	0.7	**1**	0.8	0.8	0.7	0.8	0.8	**1**
		P@100	0.78	0.97	0.80	0.82	0.80	0.81	0.82	**1**
		P@200	0.72	0.88	0.71	0.77	0.73	0.76	0.75	**0.92**
		P@300	0.71	0.83	0.66	0.69	0.72	0.70	0.72	**0.87**
		P@500	0.64	0.79	0.58	0.65	0.68	0.67	0.66	**0.84**
		P@1000	0.52	0.75	0.56	0.61	0.65	0.60	0.62	**0.80**
	Epinions	*P*@10	0.7	0.9	0.7	0.8	0.8	0.8	0.9	**1**
		P@100	0.80	0.94	0.83	0.84	0.83	0.82	0.88	**0.98**
		P@200	0.73	0.90	0.75	0.76	0.77	0.77	0.82	**0.93**
		P@300	0.66	0.85	0.67	0.72	0.68	0.73	0.79	**0.88**
		P@500	0.62	0.78	0.62	0.66	0.62	0.68	0.74	**0.82**
		P@1000	0.55	0.72	0.53	0.62	0.57	0.63	0.70	**0.80**
	Flixster	*P*@10	0.7	**1**	0.8	0.8	0.8	0.8	0.9	**1**
		P@100	0.80	**0.96**	0.82	0.86	0.82	0.85	0.88	0.95
		P@200	0.71	0.90	0.72	0.76	0.73	0.75	0.82	**0.91**
		P@300	0.67	0.82	0.67	0.70	0.68	0.67	0.77	**0.86**
		P@500	0.55	0.77	0.64	0.63	0.66	0.65	0.74	**0.83**
		P@1000	0.52	0.71	0.58	0.62	0.61	0.61	0.68	**0.79**

进一步地, 为了考察不同算法在链路预测任务中的整体水平, 我们以节点间的结对约束关系作为先验伴随信息, 由网络中抽取不同比例的连边, 设置先验信息标注比例为 10%, 以 MAP 作为评价指标, 实验结果如图 5 所示.



(a) DBLP 上的链路预测结果 (b) Epinions 上的链路预测结果 (c) Flixster 上的链路预测结果

图 5 不同边移除比例设定下各算法的链路预测结果 (*MAP*)

© 中国科学院软件研究所 http://www.jos.org.cn

这些结果表明, 随着网络中连边抽取数量的增加, 所有方法的链路预测结果都在一定程度上受到影响, 呈现出先上升后下降的变化趋势. 这是由于, 增加连边抽取的数量使得测试集规模增大, 链路预测的命中概率也随之增大. 同时, 测试集规模增大会导致可用于训练模型的样例减少, 因此当抽取比例增加到一定水平后, 链路预测结果就会变差. 在这些算法中, NRLIAI 具有一定的生成能力, 能在一定程度上减轻训练样例减少带来的负面影响, 因而能够始终获得优于其他方法的预测结果. DDRW 和 TriDNR 受连边抽取数量变化的影响最为显著, 这主要是由于这 2 种方法都利用 DeepWalk 捕捉网络中的局部结构, 当大量连边被移除后, 网络局部结构遭到严重破坏, 导致了网络表示学习有效性的恶化.

3 结论及展望

本文提出了一种新的网络表示学习模型 NRLIAI, 能够融合网络中的先验伴随信息为网络表示学习过程引入附加辅助信息, 使网络的向量表示更加有效可靠. NRLIAI 模型以 VAE 作为信息传播架构, 建立编码器-解码器映射结构, 获得对网络结构和节点属性完整保留的网络低维表示. 同时, 利用先验伴随信息为表示学习过程提供指导, 使原网络空间与低维嵌入空间中对节点间关系的描述趋于一致. 最终, 构建一体化的优化模型, 实现上述 2 个优化目标的有机整合. 将网络表示学习结果应用于真实网络中的节点分类和链路预测任务, 通过实验结果验证了该模型的有效性. NRLIAI 模型为网络表示学习过程中伴随信息的引入和处理提供了一种可用方式, 在结构和特性保持的基础上进一步提高了表示学习的有效性.

在未来的研究中, 我们将围绕以下几个方面进一步提升该模型的性能和适应性: 第一, 解码器需要对整个网络进行重构, 导致模型的理论计算复杂度达到$O(N^2)$水平, 在大规模网络上的计算效率较低. 为此, 需要充分利用网络拓扑结构的稀疏性, 设计有效的并行计算处理机制, 对模型的优化过程加以改进和调整, 使其满足大规模网络对计算效率的需要. 第二, 本文讨论了利用几类常见伴随信息对网络表示学习过程进行辅助与指导, 如何进一步挖掘和利用类型更为丰富的伴随信息, 以及设计多种类伴随信息共存场景下的信息融合与传播机制, 是未来研究中需要着重考虑的问题. 第三, 真实网络中的伴随信息通常由用户产生, 其质量可能参差不齐, 如何削减低质量伴随信息对表示学习过程的不利影响, 并充分挖掘其与网络中其他信息间的互补性, 也是未来研究的重要方向之一.

References:

[1] Cui P, Wang X, Pei J, Zhu WW. A survey on network embedding. IEEE Trans. on Knowledge and Data Engineering, 2019, 31(5): 833–852. [doi: 10.1109/TKDE.2018.2849727]

[2] Tu CC, Yang C, Liu ZY, Sun MS. Network representation learning: An overview. SCIENTIA SINICA Informationis, 2017, 47(8): 980–996 (in Chinese with English abstract). [doi: 10.1360/N112017-00145]

[3] Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: A survey. Knowledge-based Systems, 2018, 151: 78–94. [doi: 10.1016/j.knosys.2018.03.022]

[4] Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online learning of social representations. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM, 2014. 701–710. [doi: 10.1145/2623330.2623732]

[5] Tang J, Qu M, Wang MZ, Zhang M, Yan J, Mei QZ. LINE: Large-scale information network embedding. In: Proc. of the 24th Int'l Conf. on World Wide Web. Florence: Int'l World Wide Web Conf. Steering Committee, 2015. 1067–1077. [doi: 10.1145/2736277.2741093]

[6] Berg RVD, Kipf TN, Welling M. Graph convolutional matrix completion. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM, 2018.

[7] Jin D, Wang KZ, Zhang G, Jiao PF, He DX, Fogelman-Soulié F, Huang X. Detecting communities with multiplex semantics by distinguishing background, general, and specialized topics. IEEE Trans. on Knowledge and Data Engineering, 2020, 32(11): 2144–2158. [doi: 10.1109/TKDE.2019.2937298]

[8] Hamilton WL, Ying R, Leskovec J. Representation learning on graphs: Methods and applications. IEEE Data Engineering Bulletin, 2017, 40(3): 52–74.

[9] Tang L, Liu H. Relational learning via latent social dimensions. In: Proc. of the 15th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Paris: ACM, 2009. 817–826. [doi: 10.1145/1557019.1557109]

[10] Yang C, Liu ZY, Zhao DL, Sun MS, Chang EY. Network representation learning with rich text information. In: Proc. of the 24th Int'l Conf. on Artificial Intelligence. Buenos Aires: AAAI Press, 2015. 2111–2117.

[11] Grover A, Leskovec J. Node2vec: Scalable feature learning for networks. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016. 855–864. [doi: 10.1145/2939672.2939754]

[12] Li JZ, Zhu J, Zhang B. Discriminative deep random walk for network classification. In: Proc. of the 54th Annual Meeting of the

© 中国科学院软件研究所 http://www.jos.org.cn

Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016. 1004–1013. [doi: 10.18653/v1/P16-1095]

[13] Pan SR, Wu J, Zhu XQ, Zhang CQ, Wang Y. Tri-party deep network representation. In: Proc. of the 25th Int'l Joint Conf. on Artificial Intelligence. New York: AAAI Press, 2016. 1895–1901.

[14] Chen L, Zhu PS, Qian TY, Zhu H, Zhou J. Edge sampling based network embedding model. Ruan Jian Xue Bao/Journal of Software, 2018, 29(3): 756–771 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5435.htm [doi: 10.13328/j.cnki.jos.005435]

[15] Wang HW, Wang J, Wang JL, Zhao M, Zhang WN, Zhang FZ, Xie X, Guo MY. GraphGAN: Graph representation learning with generative adversarial nets. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018. 2508–2515.

[16] Wang SH, Aggarwal C, Tang JL, Liu H. Attributed signed network embedding. In: Proc. of the 2017 ACM on Conf. on Information and Knowledge Management. Singapore: ACM, 2017. 137–146. [doi: 10.1145/3132847.3132905]

[17] Cao SS, Lu W, Xu QK. Deep neural networks for learning graph representations. In: Proc. of the 30th AAAI Conf. on Artificial Intelligence. Phoenix: AAAI, 2016. 1145–1152.

[18] Wang DX, Cui P, Zhu WW. Structural deep network embedding. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016. 1225–1234. [doi: 10.1145/2939672.2939753]

[19] Bruna J, Zaremba W, Szlam A, LeCun Y. Spectral networks and locally connected networks on graphs. In: Proc. of the 2nd Int'l Conf. on Learning Representations. Banff, 2014.

[20] Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 1025–1035.

[21] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: Proc. of the 4th Int'l Conf. on Learning Representations. 2016.

[22] Kipf TN, Welling M. Variational graph auto-encoders. In: Proc. of the 30th Conf. on Neural Information Processing Systems. New York: Curran Associates Inc., 2016. 11313–11320.

[23] Feng R, Yang Y, Hu WJ, Wu F, Zhang YT. Representation learning for scale-free networks. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018. 282–289. [doi: 10.1609/aaai.v32i1.11256]

[24] Chen HC, Perozzi B, Hu YF, Skiena S. HARP: Hierarchical representation learning for networks. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018. 2127–2134. [doi: 10.1609/aaai.v32i1.11849]

[25] Li CZ, Li ZJ, Wang SZ, Yang Y, Zhang XM, Zhou JS. Semi-supervised network embedding. In: Proc. of the 22nd Int'l Conf. on Database Systems for Advanced Applications. Suzhou: Springer, 2017. 131–147. [doi: 10.1007/978-3-319-55753-3_9]

[26] Huang X, Li JD, Hu X. Label informed attributed network embedding. In: Proc. of the 10th ACM Int'l Conf. on Web Search and Data Mining. Cambridge: ACM, 2017. 731–739. [doi: 10.1145/3018661.3018667]

[27] Yang ZL, Cohen WW, Salakhutdinov R. Revisiting semi-supervised learning with graph embeddings. In: Proc. of the 33rd Int'l Conf. on Int'l Conf. on Machine Learning. New York: JMLR.org, 2016. 40–48.

[28] He T, Gao LL, Song JK, Wang X, Huang KJ, Li YF. SNEQ: Semi-supervised attributed network embedding with attention-based quantisation. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 4091–4098. [doi: 10.1609/aaai.v34i04.5832]

[29] Kingma DP, Welling M. Auto-encoding variational Bayes. In: Proc. of the 2nd Int'l Conf. on Learning Representations. Banff, 2014.

附中文参考文献:

[2] 涂存超, 杨成, 刘知远, 孙茂松. 网络表示学习综述. 中国科学: 信息科学, 2017, 47(8): 980–996. [doi: 10.1360/N112017-00145]

[14] 陈丽, 朱裴松, 钱铁云, 朱辉, 周静. 基于边采样的网络表示学习模型. 软件学报, 2018, 29(3): 756–771. http://www.jos.org.cn/1000-9825/5435.htm [doi: 10.13328/j.cnki.jos.005435]

杜航原(1985－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为图机器学习.

白亮(1982－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为无监督学习, 社会网络挖掘.

王文剑(1968－), 女, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为机器学习, 数据挖掘.

© 中国科学院软件研究所 http://www.jos.org.cn