计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20200206
57(8): 1674-1682, 2020

基于复合关系图卷积的属性网络嵌入方法

陈亦琦　钱铁云　李万理　梁贻乐
(武汉大学计算机学院　武汉　430072)
(yiqic16@whu.edu.cn)

Exploiting Composite Relation Graph Convolution for Attributed Network Embedding

Chen Yiqi, Qian Tieyun, Li Wanli, and Liang Yile
(*School of Computer Science, Wuhan University, Wuhan* 430072)

Abstract　Network embedding aims at learning a low-dimensional dense vector for each node in the network. It has attracted much attention from researchers in recent years. Most existing studies mainly focus on modeling graph structure and neglect the attribute information. Though attributed network embedding methods take node attribute into account, the informative relations between nodes and their attributes are still under-exploited. In this paper, we propose a novel framework to employ the abundant relation information for attributed network embedding. To this end, we first present to construct the composite relations between the nodes and their attributes in attributed networks. We then develop a composite relation graph convolution network (CRGCN) to encode the composite relations in both types of networks. We conduct extensive experiments on real world datasets and results demonstrate the effectiveness of our model on various network analysis tasks.

Key words　attributed network embedding; graph convolution network; composite relation; social network analysis; basic relation

摘　要　网络嵌入的目的是学习网络中每个节点的低维稠密向量,该问题吸引了研究者的广泛关注.现有方法大多侧重于对图结构的建模,而忽略了属性信息.属性化网络嵌入方法虽然考虑了节点属性,但节点与属性之间的信息关系尚未得到充分的利用.提出了一种利用丰富的关系信息进行属性网络嵌入的新框架.为此,我们首先为属性网络构造节点及其属性之间的复合关系,随后提出一个复合关系图卷积网络(composite relation graph convolution network, CRGCN)模型对这2种网络中的复合关系进行编码.在真实世界的数据集上进行了广泛的实验,结果证明了该模型在多种社交网络分析的有效性.

关键词　属性网络嵌入;图卷积网络;复合关系;社交网络分析;基本关系

中图法分类号　TP391

信息网络,如社交网络、蛋白质网络、用户-物品评价网络等在当今社会中无处不在.网络嵌入的目标是学习网络中每个节点的低维稠密向量.网络嵌入作为网络分析任务中的一个基本问题,已经引起了研究者的广泛关注[1-7].

现有的网络嵌入方法大多侧重于对图结构的

收稿日期:2020-03-21;**修回日期**:2020-05-13
基金项目:国家自然科学基金项目(61572376,91646206);国家电网有限公司科技项目(5700-202072180A-0-00-00)
This work was supported by the National Natural Science Foundation of China (61572376, 91646206) and the State Grid Technology Project (5700-202072180A-0-00-00).
通信作者:钱铁云(qty@whu.edu.cn)

建模,而没有考虑节点属性等边信息.最近出现了面向属性网络嵌入(attributed network embedding, ANE)的方法[8-11],在网络分析任务方面展示出比传统方法更好的效果.然而,现有 ANE 方法只考虑基本的关系比如用户的属性,忽略了诸如"用户的邻居的邻居"等复合关系.

我们在图 1 中给出了属性网络中的基本关系和复合关系的一个例子.实线表示原始的基本关系,虚线表示这 2 个节点之间将有一个构造的复合关系.

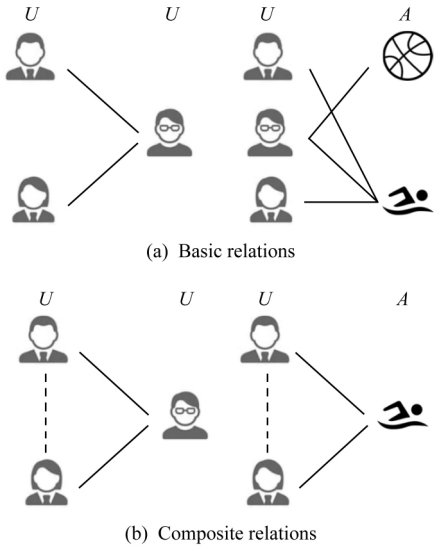


Fig. 1 An example of basic and composite relations in an attributed network

图 1 属性网络中基本关系和复合关系样例

在图 1 所示的属性网络(用户节点 U 及其属性 A)中,有 2 种类型的基本关系:

1) 用户-用户关系(2 个用户是朋友),

2) 用户-属性关系(用户的爱好是篮球或游泳).

从上述基本关系出发可以构造出同质网络的复合关系来获取网络的其他性质,我们称之为复合关系,如:

1) 用户-用户-用户(uu-uu)关系(2 个用户都有一个到共同朋友的链接),

2) 用户-属性-用户(ua-au)关系(2 个用户有相同的爱好).

显然,复合关系比基本关系传达了更多的信息.直觉上,2 个既有共同朋友又有共同爱好的用户比那些有共同朋友但没有共同爱好的用户更有可能成为朋友.虽然现有网络嵌入方法如 LINE[5] 和 SDNE[6] 利用二阶近似对 uu-uu 关系进行编码,却没有考虑属性信息,从而忽略了 ua-au 关系.

为了解决上述问题,我们提出了一个新的框架来利用节点及其属性之间的各种类型的关系.首先,在属性网络上构建复合关系.然后,构造一个复合关系的图卷积网络(composite relation graph convolution network, CRGCN)模型来编码复合关系中的信息.与现有的 ANE 方法对比,本文模型由于编码了复合关系而展示出比 ANE 方法更好的效果.

本文的主要贡献包括 3 个方面:

1) 提出了一种无监督属性网络嵌入框架,用于求解属性网络中的基本关系和复合关系;

2) 提出了一个复合关系图卷积网络来保留网络中丰富的属性信息;

3) 在真实数据集上进行了大量的实验,结果证明我们的框架对各种网络分析都非常有效.

1 相关工作

网络表示学习方法已经应用在多种分析任务上,包括链接预测[12]、节点分类[13]、社区发现[14]等.传统的方法像局部线性嵌入(LLE)[15]、Laplacian EigenMap[16]都是基于降维技术的.近期,很多基于 word2vec[17] 的方法被提出,如 DeepWalk[3],LINE[5],node2vec[18] 等;也有偏重某类网络分析任务或者结合新的神经网络架构的网络表示方法,如 SNBC[19],HOPE[20],MNMF[21],Struc2vec[4],GraphGAN[22],ANE[23] 和 DynamicTriad[24] 等.该类方法通常是从维护某种社会性质出发,通过神经网络的方式来拟合该性质,从而为每个节点学到一个更好的表示.比如:DeepWalk[3] 是首个将 word2vec[17] 思想引入网络表示中的方法,作者通过分别观察在维基文本词频和在网络节点中随机游走后节点频率的结果,发现二者都近似符合幂律分布,从而将词与词之间的上下文关系迁移到网络中来,通过随机游走"造句"来捕获节点间的潜在关系.LINE[10] 则是考虑了网络中"一阶相似性"和"二阶相似性"的性质,从网络中的邻居关系和共有邻居关系的角度进行了建模.Node2vec[18] 则是通过对 DeepWalk 的随机游走策略进行更细致的改进来学习到节点表示.HOPE[20] 通过维护有向网络中的非对称传递性来学习到节点间的高阶相似性.GraphGAN[22] 则是通过基于对抗生成的思想来对边生成的过程进行建模,从而对网络进行表示.

相比传统方法,上述网络嵌入方法通过结合社会性质和深度神经网络,取得了更好的性能.但是,

该类方法致力于建模网络的拓扑结构,而忽略了属性信息,因此它们不适合用来建模属性网络.

属性网络表示方法(attributed network embedding, ANE)同时将网络结构信息和内容信息纳入考虑.ANE 的方法可以归类为(半)监督和无监督 2 类,其中(半)监督类方法是指模型在训练时需要类别信息来进行监督指导的方法,无监督类方法是不需要类别监督信息指导的方法.经典的(半)监督方法包括 TriDNR[8],Planetoid-T[25],SEANO[26]和 LANE[27]等.例如:TriDNR 通过结合 skip-gram[17]的方法来结合结构信息,节点内容和节点类别.Planetoid-T[25]是一个结合节点内容和邻居信息的半监督图表示方法.SEANO[26]是一个探索了离群点性质的半监督属性网络表示方法.LANE[27]将属性网络和标签类别信息映射到同一个嵌入空间来学习到网络表示方法.然而,监督式的方法需要类别信息的指导,当网络中不含类别信息时,无法通过类别信息的反馈来学习表示,从而限制了其应用场景.无监督式的方法能够在无标签的网络使用,不受标签信息限制,因而具有更广泛的应用价值.比如 GAE[28]使用了自编码器的方式来捕捉拓扑结构和内容信息.VGAE[28]是一种基于变分图自编码器来结合结构和内容信息的方法.SNE[29]通过维护结构相似度和属性相似度来学到网络表示.ARGA[9]是一种基于图自编码器的对抗图表示框架,图变分自编码器 ARGVA 是它的变种.DANE[30]通过深度神经网络来捕获拓扑结构和节点属性之间的相似性.ANRL[10]使用基于属性感知的 skip-gram 方法构造了一个邻居增强的自编码器,以此来建模节点属性.其他在属性网络表示的研究方向包括:加速[31-32]或者探索其他信息的使用[27].尽管在无监督 ANE 任务上已经取得了令人瞩目的进展,但节点和属性之间的关系还没有被完全探索.

2 基于复合关系图卷积的属性网络嵌入方法

本节首先介绍属性网络中的复合关系,然后展示我们基于图卷积网络的模型.

2.1 属性网络及其关系

本节介绍属性网络及其关系.属性网络中的节点拥有其自身的属性.例如对于一个引用网络,每个节点对应一篇文章,每条边对应 2 篇文章之间的引用,属性对应文章的关键词;对于一个社交网络,每个节点对应一个用户,每条边对应一个关注关系,属性对应用户的个人信息.

属性网络的形式化定义为:$G=(U,\boldsymbol{UU},A,\boldsymbol{UA})$,其中 $U=\{u_1,u_2,\cdots,u_n\}$ 是用户集合,n 是用户数量,$\boldsymbol{UU}$ 是用户-用户关系矩阵,$A=\{a_1,a_2,\cdots,a_m\}$ 是用户的属性集合,m 是属性数量,$\boldsymbol{UA}$ 是用户-属性关系矩阵.对于同质网络 G,$u\in U$ 和 $a\in A$ 是其基本对象,uu,ua 分别是 $\boldsymbol{UU}$ 和 $\boldsymbol{UA}$ 关系矩阵的元素,代表用户和属性的基本关系.现有绝大部分 ANE 方法[9,28,33]都是建立在上述定义的同质网络 G 上.其中的关系展示在图 2(a).

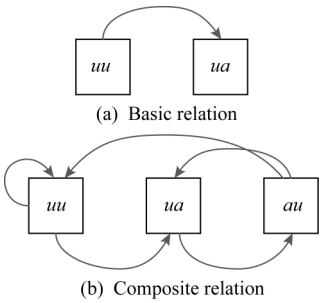


Fig. 2 Relations in an attributed network

图 2 属性网络中的关系

现有方法对于关系的利用上存在 2 方面不足:

1) 现有方法使用了 uu 关系来传递网络中的信息,却没有考虑其他基本关系,如 au(属性-用户关系的缩写),如图 2(b)所示.基本关系 au 是从属性视角获得的关系,比如对于一篇"NLP"标签(tag)的论文,可以看做在属性节点"NLP"和论文之间存在一条虚拟边,所有含有该属性的论文可以被聚合起来,进行更深入的检索.

2) 现有方法也忽略了更为复杂的关系:复合关系,如图 3 中的线条所示.我们定义复合关系为组合了至少 2 种基本关系的关系,如 uu 和 ua 组合得到的复合关系 $uuua$ 表示的是"用户和用户邻居的属性"的关系.复合关系保留了丰富的信息,如果上述关系可以被进一步挖掘,学到的表示也能保留更多的关系特性,从而改善社交网络分析任务的性能.

基于上述观察和分析,我们尝试改进关系的利用形式.首先给属性网络 G 增加基本关系矩阵 $\boldsymbol{AU}$ 的定义,用来代表 au 的关系.接着拓展 G 来包含 5 种复合关系:($uuua$;$uaau$;$uuuu$;$auua$;$auuu$),其中 $uuua$ 表示 uu 和 ua 关系的组合.基础的 au 关系和 5 种复合关系都展示在图 3 的下半部分.为了更清楚地展示,我们将复合关系分类为:

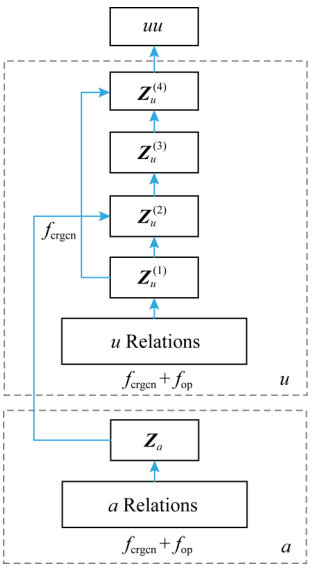


Fig. 3 The architecture of CRGCN framework

图 3 CRGCN 框架结构图

用户的复合关系:($uuua$;$uaau$;$uuuu$)

属性的复合关系:($auua$;$auuu$)

新的关系包含了比(uu;ua)更多的信息,比如用户的新关系可以显式地表达出:用户邻居的邻居($uuuu$)、用户共享的属性($uaau$)和用户的邻居的属性($uuua$)这 3 种关系;属性的新关系可以显式地表达出:共享用户的属性($auua$)和属性关联到的用户的邻居($auuu$)这 2 种关系.尽管我们可以建立像($uuuaau$)关系的更复杂的组合,但高阶的组合会增加计算复杂度,同时可能引入更多噪声,因此我们只考虑上面列出的一阶组合.

2.2 CRGCN 框架:从复合关系中学习

本节我们将介绍复合关系图卷积网络(CRGCN)框架,用于从我们提出的复合关系中学习网络嵌入.CRGCN 的整体架构如图 3 所示.

图卷积网络技术是近年来提出的一种新的已被证明有效的计算方法[9,28,33].给定 2.1 节所定义的属性网络 $G=(U,\boldsymbol{UU},A,\boldsymbol{UA})$,为了刻画图中的结构和属性信息,图卷积网络函数 f_{gcn} 的定义如下:

$$\boldsymbol{Z}^{(l+1)}=f_{\text{gcn}}(\boldsymbol{Z}^{(l)},\boldsymbol{UU}\,|\,\boldsymbol{W}^{(l)})=\sigma(g(\boldsymbol{UU})\boldsymbol{W}^{(l)}\boldsymbol{Z}^{(l)}),\tag{1}$$

其中,$\boldsymbol{Z}^{(l)}$ 是卷积的输入,$\boldsymbol{W}^{(l)}$ 是需要学习的卷积核参数矩阵,l 是层数,$\boldsymbol{Z}^{(l+1)}$ 是本层的输出.$g(\boldsymbol{UU})$ 是原始结构信息 $\boldsymbol{UU}$ 的转换.函数 g 可以通过与单位矩阵 $\boldsymbol{I}$ 相乘保证 $\boldsymbol{UU}$ 的不变,如式(2)所示,或使用拉普拉斯正则化,如式(3)所示.

$$g(\boldsymbol{UU})=I(\boldsymbol{UU}),\tag{2}$$

$$g(\boldsymbol{UU})=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{UU}\boldsymbol{D}^{\frac{1}{2}},\tag{3}$$

其中,$\boldsymbol{D}$ 表示 $\boldsymbol{UU}$ 的对角度矩阵,σ 是激活函数,计算公式为

$$sigmoid(x)=\frac{1}{1+e^{-x}},$$

$relu(x)=\max(0,x)$ 或者简单的线性变换 $linear(x,\boldsymbol{W},b)=xW+b$.

但是,一个基本的 gcn 函数只能处理像这样的简单关系($\boldsymbol{UU}$;$\boldsymbol{UA}$),卷积的结构信息仅限于 $\boldsymbol{UU}$.为了利用复合关系,我们将基本的 GCN 扩展为如下所述的复合关系 CRGCN.其公式定义为

$$\boldsymbol{Z}_{(\boldsymbol{R}_s,\boldsymbol{R}_i)}=f_{\text{crgcn}}(\boldsymbol{R}_s,\boldsymbol{R}_i\,|\,W_{(R_s,R_i)})=\sigma(g(\boldsymbol{R}_s)\boldsymbol{R}_i W_{(R_s,R_i)}),\tag{4}$$

$\boldsymbol{R}_s$ 和 $\boldsymbol{R}_i$ 是($\boldsymbol{UU}$,$\boldsymbol{UA}$,$\boldsymbol{AU}$)的 2 个关系矩阵,$\boldsymbol{Z}_{(\boldsymbol{R}_s,\boldsymbol{R}_i)}$ 是卷积的输出,$W_{(R_s,R_i)}$ 是需要学习的卷积核参数,g 是结构信息 $\boldsymbol{R}_s$ 的转换函数,σ 是激活函数或简单的线性层.更直观的解释是,$\boldsymbol{R}_s$ 可以看作 GCN 的结构信息,类似于标准 CNN 的滑动窗口;$\boldsymbol{R}_i$ 是我们需要卷积的输入,相当于 CNN 输入的图片;$W_{(R_s,R_i)}$ 则对应于 CNN 的卷积核,$\boldsymbol{Z}_{(R_s,R_i)}$ 是 CNN 的特征.

在 2.1 节中,我们构造了属性网络的 5 种复合关系.以复合关系 $uuua$ 为例,我们的 CRGCN 将使用用户-用户关系 uu 对用户属性关系进行卷积,ua 得到用户的潜在属性表示.我们将充分利用 5 种组合,而不是像基本的 GCN 那样只考虑 $uuua$ 关系.例如我们可以嵌入更多类型的关系,比如用户的潜在邻居表示($uuuu$)和属性的潜在属性表示($auua$).

通过在多种复合关系上应用 f_{crgcn} 函数,可以获得属性网络不同视角的表示:3 个用户隐变量表示($\boldsymbol{Z}_{(\boldsymbol{UU},\boldsymbol{UU})}$,$\boldsymbol{Z}_{(\boldsymbol{UU},\boldsymbol{UA})}$,$\boldsymbol{Z}_{(\boldsymbol{UA},\boldsymbol{AU})}$)这 2 个属性隐变量表示($\boldsymbol{Z}_{(\boldsymbol{AU},\boldsymbol{UU})}$,$\boldsymbol{Z}_{(\boldsymbol{AU},\boldsymbol{UA})}$).2 种关系分别使用"$a$ relations"和"u relations"表示在图 3 中.

我们融合上述隐变量得到统一表示.对于用户隐变量而言,我们获得用户的浅层表示 $\boldsymbol{Z}_u^{(1)}$:

$$\boldsymbol{Z}_u^{(1)}=f_{\text{op}}(\boldsymbol{Z}_{(\boldsymbol{UU},\boldsymbol{UU})},\boldsymbol{Z}_{(\boldsymbol{UU},\boldsymbol{UA})},\boldsymbol{Z}_{(\boldsymbol{UA},\boldsymbol{AU})}),\tag{5}$$

其中,f_{op} 是聚合函数,可以采用均值/加权/拼接操作、线性变换、神经网络或注意力网络等.这一步对应于基本 GCN 的第一层.同样地,我们可以获取属性的浅层表示 $\boldsymbol{Z}_a$:

$$\boldsymbol{Z}_a=f_{\text{op}}(\boldsymbol{Z}_{(\boldsymbol{AU},\boldsymbol{UU})},\boldsymbol{Z}_{(\boldsymbol{AU},\boldsymbol{UA})}).\tag{6}$$

与基本的多层 GCN 操作相同,我们使用多层的复合关系 CRGCN,其公式为

$$\boldsymbol{Z}_{(\boldsymbol{UA},\boldsymbol{Z}_a)}=f_{\text{crgcn}}(\boldsymbol{UA},\boldsymbol{Z}_a\,|\,W_{uaza}),\tag{7}$$

$$\boldsymbol{Z}_u^{(2)}=f_{\text{op}}(\boldsymbol{Z}_u^{(1)},\boldsymbol{Z}_{(\boldsymbol{UA},\boldsymbol{Z}_a)}),\tag{8}$$

$$\boldsymbol{Z}_u^{(3)}=f_{\text{crgcn}}(\boldsymbol{UU},\boldsymbol{Z}_u^{(2)}\,|\,W_{u3}).\tag{9}$$

在式(7)～(9)中,我们首先通过在 $\boldsymbol{UA}$ 关系上应用 CRGCN 操作,将 $\boldsymbol{Z}_a$ 转化为 $\boldsymbol{Z}_{(\boldsymbol{UA},\boldsymbol{Z}_a)}$.然后通过 f_{op} 合并 $\boldsymbol{Z}_u^{(1)}$ 和 $\boldsymbol{Z}_{(\boldsymbol{UA},\boldsymbol{Z}_a)}$ 得到 $\boldsymbol{UA}$ 的隐藏表示,即 $\boldsymbol{Z}_u^{(2)}$.接着通过结构关系 $\boldsymbol{UU}$ 卷积 $\boldsymbol{Z}_u^{(2)}$ 得到深层的表示 $\boldsymbol{Z}_u^{(3)}$.

最后,为了利用不同级别的表示信息,我们将用户基本表示向量 $\boldsymbol{Z}_{(\boldsymbol{UU},\boldsymbol{UA})}$(用户基本关系)、用户的浅层表示 $\boldsymbol{Z}_u^{(1)}$(用户复合关系)以及深层表示 $\boldsymbol{Z}_u^{(3)}$(用户和属性关系)聚合为

$$\boldsymbol{Z}_u^{(4)}=f_{\text{op}}(\boldsymbol{Z}_{(\boldsymbol{UU},\boldsymbol{UA})},\boldsymbol{Z}_u^{(1)},\boldsymbol{Z}_u^{(3)}).\tag{10}$$

将网络嵌入到潜在空间后,我们需要了解它是否有效.作为一种无监督网络嵌入学习,我们选择通过重构网络来训练表示.编码器的输出 $\boldsymbol{Z}_u^{(4)}$ 将被重新解码重构 $\boldsymbol{UU}$ 结构,解码过程如下:

$$p(\widehat{\boldsymbol{UU}}\mid\boldsymbol{UU})=\prod_{i=1}^{n}\prod_{j=1}^{n}p(\widehat{\boldsymbol{UU}}_{i,j}\mid z_i,z_j),\tag{11}$$

$$p(\widehat{UU}_{i,j}=1|z_i,z_j)=sigmoid(z_i,z_j^T).\tag{12}$$

$p(\widehat{\boldsymbol{UU}}|\boldsymbol{UU})$ 是解码器的结果,表示用户 z_i,z_j 之间是否有连边,我们不需要重构 $\boldsymbol{UA}$,因为这样可能会带来噪声.在获得了重构结果后,我们通过最小化重构误差损失函数训练模型,其公式为

$$L=\sum_{i=1}^{n}\sum_{j=1}^{n}L_{i,j},\tag{13}$$

$$L_{i,j}=-[pw\times UU_{i,j}\log p(\widehat{UU}_{i,j}=1|z_i,z_j)+(1-UU_{i,j})\log(1-p(\widehat{UU}_{i,j}=1|z_i,z_j)).\tag{14}$$

我们使用二进制交叉熵损失和 pw 来控制正样本的权重.pw 可以增强预测观测值为 1 的链接,放松对观测值为 0 链接的约束.它可以被用来测量值为 0 和 1 之间的概率,定义为

$$pw=(n\times n-nz)/nz,\tag{15}$$

其中,n 是用户的数量,nz 是 $\boldsymbol{UU}$ 中非 0 实例的个数.

我们将编码器的输出 $\boldsymbol{Z}_u^{(4)}$ 解码,用于重构 $\boldsymbol{UU}$.最后,我们采用最小化重构误差训练模型.

模型的算法描述的复杂度分析为:由于神经网络模型涉及的计算过程较复杂,并且计算工具本身存在优化的差异,为了减少该类因素的影响,我们计算复杂度时以矩阵乘法的次数为基本单位,CRGCN 模型复杂度计算为

$$T(n,m,d)=\Theta(f_1+f_2+f_3+f_4)=\Theta(2dn^2+(dn^2+dmn)+2dmn)+\Theta((dn^2+dmn)+2dmn+(dmn+d^2m))+\Theta(dn^2+d^2n)+0=\Theta(5dn^2+(7dm+d^2)n+d^2m),\tag{16}$$

其中,f_k 对应计算 $Z_u^{(k)}$ 时所需要的计算操作,n 是节点数量,m 是属性数量,d 是表示向量的维度(节点和属性表示取相同的潜在维度,$d\ll n$).该公式的复杂度主要由 n 和 m 的数值大的一方决定:当 $n\ll m$ 时,$T(n,m,d)=O(5dn^2)$;当 $m\ll n$ 时,$T(n,m,d)=O((7dn+d^2)m)$.

3 实 验

3.1 实验设置

3.1.1 数据集

我们在 3 个公开数据集 Cora,Citeseer,Pubmed 上进行了 2 种经典的分析任务:链接预测和节点聚类.数据集的统计信息如表 1 所示.上述数据集是同质属性网络,把科学论文作为节点,引用关系作为边,文档里的词作为属性[34].

Table 1 Statistics for Homogeneous Datasets

表 1 同质网络数据集统计信息

Dataset	Node	Edge	Content Words	Attributes
Cora	2 708	5 429	3 880 564	1 433
Citeseer	3 327	4 732	12 274 336	3 703
Pubmed	19 717	44 338	9 858 500	500

3.1.2 基线方法和设置

对于链接预测和节点聚类实验,我们将对比以下 7 种最新的基线方法:

1) DeepWalk[3].一个基于网络结构信息的网络表示方法.作者在观察到维基文本的词频分布与随机游走的节点频率存在相似性后,将 word2vec 的思想借鉴到网络表示中来,考虑了网络中的中心节点与上下文节点间的相关性,通过随机游走的方式来造句,得到序列后进行训练得到节点表示.

2) LINE[5].一个基于网络结构信息的网络表示方法.考虑了网络中节点间的一阶相似性和二阶相似性,通过边采样的方式来训练模型,学到节点一、二阶表示后拼接起来作为最终的特征向量,应用到相关的网络分析任务中.

3) GAE[28].一个基于自编码器框架的无监督网络表示方法,考虑了结构信息和内容信息.通过使用图卷积网络对图中的节点特征进行卷积,从而学到节点的潜在特征,再应用到相关的网络分析任务中.

4) VGAE[28].一个基于变分图自编码器的无监督网络嵌入方法,平衡了结构和内容信息.在推断模块中学习到正态分布的均值和方差参数来产生潜在表示,再在生成模块中重构出邻接关系,最终应用到相关的网络分析任务中.

5) ARGA[9].一个基于对抗约束的图自编码器的无监督网络表示算法,同时考虑了结构和属性信息.该模型在编码图信息得到节点表示后,通过一个判别器来判别一个样本是从表示中产生的还是从一个先验分布中产生的来进行约束,最终学到的表示应用到了链接预测和节点聚类任务中.

6) ARVGA[9].一个 ARGA 的变种,使用了变分图自编码器来学习嵌入.

7) ANRL[10].一个使用属性感知的 skip-gram 来捕捉网络结构信息的属性网络表示方法.该模型对节点属性编码后,分别去重构用户属性和预测图的上下文信息,从而将 2 种信息结合起来.

我们没有跟 node2vec 和 SNE 等网络表示方法进行比较,因为在 ARVGA 和 ANRL 的实验中,上述方法已经被证明性能不如我们选择的基线方法.本文的实验均在 Ubuntu16.04.5 LTS 环境下进行,使用 1.0.0 版本的 pytorch 构建网络模型和运行框架,基线方法会按照源码要求配置到对应的环境和软件版本.

对于链接预测任务,我们跟 ARVGA 方法[9]一样报告了 AUC 和 AP 指标.我们也使用了跟文献[9]相同的数据划分和测试方法:10%用于测试,5%用于校验,剩下的用于训练.对于所有的基线方法,我们使用其推荐设置,并学习得到 32 维度的节点表示来进行链接预测任务,最终报告重复 5 次实验的平均结果.我们的方法设置学习率为 0.005,最大迭代轮数 200,优化器选用 adam[35].

对于节点聚类任务,我们报告了聚类的 5 个评价指标:accuracy(Acc), $precision$, F-score($F1$), normalized mutual information(NMI)和 adjusted rand index(ARI).

对于所有的基线方法,我们使用其推荐的设置,得到 32 维度的节点表示进行节点聚类任务.我们的方法使用了和链接预测中一样的设置.由于节点聚类任务在每个方法的不同轮次上,结果波动很大,所以我们报告了每个方法最好轮次的得分作为最终结果,由于 LINE 方法做边采样没有轮次,我们调整采样边数,报告取[10^6;10^7;10^8;10^9;10^{10}]条边中效果最好的结果,对于 DeepWalk 则是调整每个点游走次数,报告在 1~10 次中最好的结果.

3.2 链接预测及其实验结果

链接预测的实验结果展示在表 2 中,方法分为网络表示方法(仅利用结构信息)、属性网络表示方法和我们的方法三大块,最好的结果用粗体表示.

Table 2 Results for Link Prediction

表 2 链接预测结果

Method	Cora		Citeseer		Pubmed	
	AUC	AP	AUC	AP	AUC	AP
LINE	0.869 2	0.898 9	0.807 4	0.858 3	0.847 3	0.880 1
DeepWalk	0.792 3	0.855 9	0.666 0	0.785 0	0.761 6	0.852 3
GAE	0.897 0	0.907 6	0.877 2	0.879 7	**0.962 0**	**0.963 0**
VGAE	0.916 2	0.929 5	0.896 0	0.907 5	0.944 4	0.946 2
ARGA	0.913 5	0.932 4	0.914 3	0.928 2	0.948 5	0.950 5
ARVGA	0.920 9	0.932 9	0.914 4	0.925 3	0.905 0	0.910 5
ANRL	0.871 9	0.866 4	0.930 5	0.930 6	0.915 7	0.909 3
CRGCN	**0.938 5**	**0.946 3**	**0.949 0**	**0.956 5**	0.954 6	0.956 5

Note: The best performance is in bold.

对于仅考虑结构信息的网络表示方法 LINE 和 DeepWalk,由于没有对属性信息进行利用,效果跟属性网络表示方法有一定的距离.

在属性网络表示方法中,CRGCN 在 Cora 和 Citeseer 数据集上取得了最好的结果,相比其他基线方法有显著性提升(成对 t 检验,满足 0.01 显著),在 Pubmed 上取得次好的效果.尽管 GAE 在 Pubmed 上取得了最好结果,这可能是因为 Pubmed 数据集上的链接情况跟属性存在相对简单的关联性,GAE 基于基础的图卷积建模,效果反而更好.但 GAE 性能并不稳定,例如在 Citeseer 数据集上其效果下降严重.

在其他基线方法中,ARGA 和 GAE 在 Cora 和 Pubmed 数据集上表现很好,原因可能是它们都是基于基础 gcn 的方法,更偏向于建模结构信息.但在有更多属性信息的 Citeseer 的数据集上,ARGA 和 GAE 就比不上能够更好地利用属性信息的 ANRL 方法.

综上所述,我们的 RGCN 通过平衡多种关系,可以在不同类型的数据集上取得稳定良好的性能.

3.3 节点聚类及其实验结果

节点聚类的结果展示在表 3~5 中,方法分为:网络表示方法(仅利用结构信息)、属性网络表示方法、我们的方法三大块,最好的结果用粗体表示.

Table 3 Clustering Results on Cora

表 3 Cora 上的聚类结果

Cora	*Acc*	*F*1	*Precision*	*NMI*	*ARI*
LINE	0.582 0	0.576 2	0.619 6	0.412 9	0.312 8
DeepWalk	0.642 2	0.631 7	0.646 5	0.428 7	0.369 7
GAE	0.541 4	0.514 9	0.564 5	0.336 6	0.231 2
VGAE	0.603 4	0.586 2	0.600 1	0.438 7	0.376 4
ARGA	0.692 8	0.679 9	0.682 9	0.499 1	0.452 1
ARVGA	0.633 3	0.632 4	0.654 8	0.466 3	0.375 6
ANRL	0.556 5	0.569 5	0.628 9	0.412 8	0.310 5
CRGCN	**0.706 8**	**0.685 9**	**0.684 6**	**0.514 3**	**0.483 9**

Note: The best performance is in bold.

Table 4 Clustering Results on Citeseer

表 4 Citeseer 上的聚类结果

Citeseer	*Acc*	*F*1	*Precision*	*NMI*	*ARI*
LINE	0.347 5	0.328 0	0.453 9	0.154 5	0.064 0
DeepWalk	0.414 2	0.397 1	0.430 4	0.166 5	0.129 8
GAE	0.392 5	0.375 4	0.394 6	0.188 3	0.133 4
VGAE	0.510 7	0.491 5	0.517 4	0.248 2	0.226 9
ARGA	0.565 1	0.551 2	0.575 7	0.291 1	0.273 7
ARVGA	0.622 8	0.587 7	0.597 9	0.350 6	0.355 1
ANRL	0.612 6	0.600 7	**0.626 2**	0.356 0	0.337 0
CRGCN	**0.651 0**	**0.606 4**	0.614 1	**0.387 8**	**0.389 7**

Note: The best performance is in bold.

Table 5 Clustering Results on Pubmed

表 5 Pubmed 上的聚类结果

Pubmed	*Acc*	*F*1	*Precision*	*NMI*	*ARI*
LINE	0.624 2	0.616 1	0.644 6	0.214 6	0.207 5
DeepWalk	0.647 6	0.629 8	0.662 2	0.243 0	0.259 3
GAE	0.648 0	0.653 8	0.649 3	0.240 2	0.236 5
VGAE	0.622 9	0.625 1	0.625 1	0.221 4	0.203 9
ARGA	0.624 5	0.624 2	0.629 2	0.214 4	0.204 1
ARVGA	0.596 4	0.598 4	0.598 5	0.194 6	0.176 5
ANRL	0.662 8	0.669 0	0.666 7	0.277 2	0.265 2
CRGCN	**0.686 1**	**0.677 4**	**0.696 9**	**0.302 5**	**0.304 4**

Note: The best performance is in bold.

从表 3～5 可以看出,通过考虑节点和它们的属性间的复合关系,我们的 CRGCN 依然取得了整体上最好的效果.

对实验结果进行详细分析.其他基线方法会在某些数据集/某些指标上取得好的结果.例如 ARGA 能在 Cora 上取得较好的效果,ANRL 可以在 Citeseer 和 Pubmed 上取得较好的效果,但只有我们的 CRGCN 能够在各个数据集和指标上能够持续保持好的性能,并在 Cora 和 Citeseer 数据集上取得显著提升.

不同于链接预测任务,节点聚类任务更困难.原因在于无监督表示学习的过程中无法学到任务相关的模式,这也是所有方法的结果都存在波动的原因.虽然增加属性对于节点聚类任务能够产生正面影响,但实际上由于无监督建模本身的特点,想要平衡属性引入的有效信息和噪声是一个挑战.我们在实验里也发现偏向于利用结构信息的方法能够在部分情况下取得相对较好的效果,比如 ARGA 和 ARGVA,它们更强调利用结构去卷积特征信息;而更偏向属性的方法如 ANRL,从节点的属性信息出发,重构了属性信息和预测邻居上下文,会在另外一部分数据集上表现良好.

为了能学到节点聚类中表现好的节点表示,需要能平衡属性和结构信息的方法,如果一个模型能够学到节点间多种类型的相关性,将会比主要偏向学习单一类型相关性的方法效果好,CRGCN 方法同时建模来自属性和结构的复合关系,因此在实验中表现出较好的性能.

3.4 参数分析

本节主要针对跟节点表示向量直接相关的维度参数进行分析,通过改变节点表示的维度,观察其对于模型性能的影响.我们以 Cora 数据集为例,分别进行链接预测和节点聚类任务,结果如图 4,5 所示:

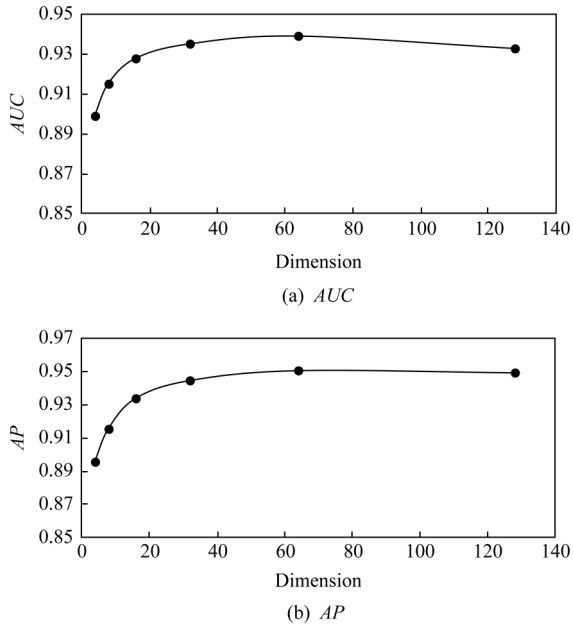


Fig. 4 Performance of link prediction with different embedding dimensions on Cora

图 4 Cora 数据集上链接预测的维度变换实验

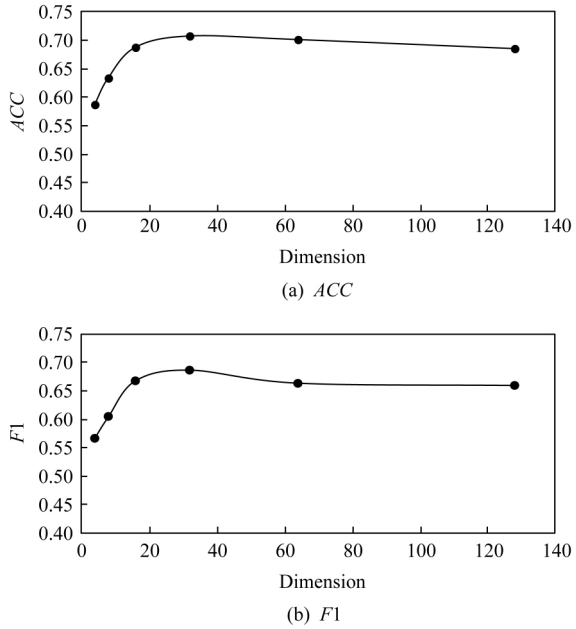


Fig. 5 Performance of node clustering with different embedding dimensions on Cora

图 5 Cora 数据集上节点聚类的维度变换实验

对于链接预测任务,观察图 4 可知,我们的模型在仅用 4 维的向量表示的时候就已经有了初步的效果,之后随着模型的维度增加,效果逐渐上升,在 64 维度左右时取得最好效果,最后趋于稳定.由此可见,初期的维度增加对于节点的表示效果能够有相对明显的改善,但维度继续增加时效果开始下降,该情况可以理解为在维护更多关系信息的同时也引入了相应的噪声,从而使得泛化性能有所下降.

对于节点聚类任务,观察图 5 可知,表示向量在 20 维左右的时候有了初步效果,在 30～40 维度之间取得最好的效果,之后趋于稳定.该任务的变化走势跟链接预测任务接近,在维度增大的同时也确实会有一定的噪声引入.

4 总 结

我们提出了一种新的用于属性网络嵌入的复合关系图卷积网络模型(CRGCN),考虑了用户和属性之间的关系,并分析了所有的一阶组合获得复合关系.接着,我们提出了一个复合关系图卷积网络来对基本关系和复合关系进行编码,把这些新的潜在表示结合在一起得到最终的嵌入.在真实世界的网络上进行广泛的实验,结果表明我们的模型优于当前最好的基线方法.

参 考 文 献

[1] Liu Zhiyuan, Sun Maosong, Lin Yankai, et al. Knowledge representation learning: A review [J]. Journal of Computer Research and Development, 2016, 53(2): 247-261 (in Chinese)
(刘知远, 孙茂松, 林衍凯, 等. 知识表示学习研究进展[J]. 计算机研究与发展, 2016, 53(2): 247-261)

[2] Ye Zhonglin, Zhao Haixing, Zhang Ke, et al. Network representation learning using the optimizations of neighboring vertices and relation model [J]. Journal of Computer Research and Development, 2019, 56(12): 2562-2577 (in Chinese)
(冶忠林, 赵海兴, 张科, 等. 基于邻节点和关系模型优化的网络表示学习[J]. 计算机研究与发展, 2019, 56(12): 2562-2577)

[3] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations [C] //Proc of the 20th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 701-710

[4] Ribeiro L F R, Saverese P H P, Figueiredo D R. Struc2vec: Learning node representations from structural identity [C] //Proc of the 23rd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2017: 385-394

[5] Tang Jian, Qu Meng, Wang Mingzhe, et al. Line: Large-scale information network embedding [C] //Proc of the 24th Int Conf on World Wide Web. Cambridge: Cambridge University Press, 2015: 1067-1077

[6] Wang Daixin, Cui Peng, Zhu Wenwu. Structural deep network embedding [C] //Proc of the 22nd ACM SIGKDD Int Conv on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1225-1234

[7] Yang Cheng, Sun Maosong, Liu Zhiyuan, et al. Fast network embedding enhancement via high order proximity approximation [C] //Proc of the 26th Int Joint Conf on Artificial Intelligence. Menlo Park: AAAI, 2017: 3894-3900

[8] Pan Shirui, Wu Jia, Zhu Xingquan, et al. Tri-party deep network representation [C] //Proc of the 25th Int Joint Conf on Artificial Intelligence. Phoenix: IJCAI, 2016: 1895-1901

[9] Pan Shirui, Hu Ruiqi, Long Guodong, et al. Adversarially regularized graph autoencoder for graph embedding [C] //Proc of the 27th Int Joint Conf on Artificial Intelligence. Menlo Park: AAAI, 2018: 2609-2615

[10] Zhang Zhen, Yang Hongxia, Bu Jiajun, et al. ANRL: Attributed network representation learning via deep neural networks [C] //Proc of the 27th Int Joint Conf on Artificial Intelligence. Menlo Park: AAAI, 2018: 3155-3161

[11] Sun Guolei, Zhang Xiangliang. A novel framework for node/edge attributed graph embedding [C] //Proc of the Pacific-Asia Conf on Knowledge Discovery and Data Mining. Berlin: Springer, 2019: 169-182

[12] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks [J]. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019-1031

[13] Bhagat S, Cormode G, Muthukrishnan S. Node classification in social networks [M] //Social Network Data Analytics. Berlin: Springer, 2011: 115-148

[14] Papadopoulos S, Kompatsiaris Y, Vakali A, et al. Community detection in social media [J]. Data Mining and Knowledge Discovery, 2012, 24(3): 515-554

[15] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000, 290(5500): 2323-2326

[16] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering [C] //Proc of the Conf and Workshop on Neural Information Processing Systems. Cambridge: MIT Press, 2002: 585-591

[17] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [C] //Proc of the Conf and Workshop on Neural Information Processing Systems. Cambridge: MIT Press, 2013: 3111-3119

[18] Grover A, Leskovec J. Node2vec: Scalable feature learning for networks [C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 855-864

[19] Nandanwar S, Murty M N. Structural neighborhood based classification of nodes in a network [C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1085-1094

[20] Ou Mingdong, Cui Peng, Pei Jian, et al. Asymmetric transitivity preserving graph embedding [C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1105-1114

[21] Wang Xiao, Cui Peng, Wang Jing, et al. Community preserving network embedding [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Menlo Park: AAAI, 2017: 203-209

[22] Wang Hongwei, Wang Jia, Wang Jialin, et al. Graphgan: Graph representation learning with generative adversarial nets [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park: AAAI, 2018: 2508-2515

[23] Dai Quanyu, Li Qiang, Tang Jian, et al. Adversarial network embedding [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park: AAAI, 2018: 2167-2174

[24] Zhou Lekui, Yang Yang, Ren Xiang, et al. Dynamic network embedding by modeling triadic closure process [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park: AAAI, 2018: 571-578

[25] Yang Zhilin, Cohen W, Salakhudinov R. Revisiting semi-supervised learning with graph embeddings [C] //Proc of Int Conf on Machine Learning. New York: ACM, 2016: 40-48

[26] Liang Jiongqian, Jacobs P, Sun Jiankai, et al. Semi-supervised embedding in attributed networks with outliers [C] //Proc of the 2018 SIAM Int Conf on Data Mining. Philadelphia: SIAM, 2018: 153-161

[27] Huang Xiao, Li Jundong, Hu Xia. Label informed attributed network embedding [C] //Proc of the 10th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2017: 731-739

[28] Kipf T N, Welling M. Variational graph auto-encoders [J]. arXiv preprint arXiv: 1611.07308, 2016

[29] Liao Lizi, He Xiangnan, Zhang Hanwang, et al. Attributed social network embedding [J]. arXiv preprint arXiv:1705.04969, 2017

[30] Gao Hongchang, Huang Heng. Deep attributed network embedding [C] //Proc of the 27th Int Joint Conf on Artificial Intelligence. Menlo Park: AAAI, 2018: 3364-3370

[31] Huang Xiao, Li Jundong, Hu Xia. Accelerated attributed network embedding [C] //Proc of the 2017 SIAM Int Conf on Data Mining. Philadelphia: SIAM, 2017: 633-641

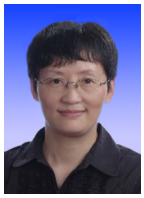
[32] Wu Wei, Li Bin, Chen Ling, et al. Efficient attributed network embedding via recursive randomized hashing [C] //Proc of the 27th Int Joint Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 2861-2867

[33] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks [C] //Proc of the 5th Int Conf on Learning Representations, 2017 [2020-03-20]. http://iclr.net/tag/iclr~2017/

[34] Van Der Maaten L. Accelerating t-SNE using tree-based algorithms [J]. The Journal of Machine Learning Research, 2014, 15(1): 3221-3245

[35] Kingma D P, Ba J. Adam: A method for stochastic optimization [J]. arXiv preprint arXiv:1412.6980, 2014

Chen Yiqi, born in 1994. Master. Received his BSc degree in computer science from Nanjing University of Aeronautics and Astronautics, China in 2016. His main research interests include network representation learning and Web mining.



Qian Tieyun, born in 1970. PhD, professor. Member of CCF. Her main research interests include text mining, Web mining, and natural language processing.

Li Wanli, born in 1993. PhD candidate. Received his master degree in circuits and systems from South China Normal University, China in 2019. His main research interests include natural language processing and Web mining.

Liang Yile, born in 1996. Master. Received his BSc degree in computer science from Hunan University, China in 2018. His main research interests include recommendation systems and Web mining.