基于霍克斯点过程的动态网络表示学习方法

尹 赢,张建朋,吉立新,李治成
(战略支援部队信息工程大学信息技术研究所,河南郑州 450003)

摘 要: 网络表示学习是将网络中的节点映射到低维空间形成低维稠密特征向量的分布式学习方法. 本文在现有网络表示学习研究的基础上,提出一种基于霍克斯点过程的动态网络表示学习方法. 该方法基于霍克斯点过程有效结合了网络历史连边信息和网络演化中的三元闭包特性对当前节点产生连边的影响,解决了现有方法难以有效捕捉网络历史信息和演化特性的问题. 在多种数据集的实验结果表明, 本文提出的方法较其它方法在节点分类、链路预测和可视化等实验中的性能均有较大的提高,实验中的 F1 分数值和 AUC 值分别提高了 3.72% ~6.41% 和 2.22% ~4.69%.

关键词: 网络表示学习; 动态网络; 霍克斯点过程; 三元闭包理论
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2020)11-2154-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.11.009

Dynamic Network Representation Learning Based on Hawkes Point Process

YIN Ying, ZHANG Jian-peng, JI Li-xin, LI Zhi-cheng
(*Institute of Information Technology, Information Engineering University, Zhengzhou, Henan* 450003, *China*)

Abstract: Network representation learning is a distributed learning method that maps nodes in a network to low-dimensional spaces to form low-dimensional dense vectors. Based on the existing network representation learning research, this paper proposed a dynamic network representation learning method based on Hawkes point process, which effectively combines the network historical edges and the ternary closure characteristics in the network evolution to generate the new edges of the current nodes. It solves the problem that the existing methods are difficult to effectively capture the network historical information and evolution characteristics of dynamic networks. Extensive experiments demonstrated that the embeddings learned from the proposed MHNE (Multivariate Hawkes process Network Embedding) model can achieve better performance than the state-of-the-art methods in downstream tasks, such as node classification and link prediction. The F1 score and AUC value in the experiments increased by 3.72% ~6.41% and 2.22% ~4.69%, respectively.

Key words: network representation learning; dynamic network; Hawkes point process; ternary closure theory

1 引言

现实生活中常见的社交系统、文献信息系统和生物医学系统等都可以抽象为复杂网络的形式,针对复杂网络的研究可有效帮助我们分析系统的特性,以有效处理复杂网络中节点分类、社团发现和链路预测等网络应用任务. 网络表示学习作为连接复杂网络原始数据和应用任务之间的桥梁,为网络中的节点学习到可作为机器学习模型输入的低维稠密向量,使得机器学习方法可以高效地处理网络应用任务. 现有的网络表示学习方法大多针对节点和连边不具备动态属性的静态网络,主要包括基于矩阵特征分解的方法[1~3],基于浅层神经网络的方法[4,5]和基于深度学习的方法[6~8].

近年来,针对静态网络的表示学习逐渐发展成熟. 但是,随着网络的不断演化,现实生活中的复杂网络呈现出明显的动态特性. 现有的动态网络表示学习方法大多是从静态网络表示学习模型中衍生而来:受基于矩阵特征值分解的静态网络表示学习方法的启发,LI 等人提出 DANE 算法[9]利用网络快照的邻接矩阵和属性矩阵的变化量,在前一时刻节点表示的基础上更新当前时刻的节点表示. CUI 等人提出 DHPE 算法[10]基于广义奇异值分解和矩阵摄动理论动态更新节点的向量表示. ZHANG 等人提出的 TIMERS 模型[11]改进了奇

收稿日期:2019-06-18;修回日期:2019-11-29;责任编辑:孙瑶
基金项目:国家自然科学基金创新研究群体资助项目(No.61521003)

异值分解在动态网络表示中的应用. 除了基于矩阵分解的动态网络表示学习方法外,还有将 LINE 模型[12]扩展于动态网络表示学习的 DNE 算法[13]和基于 SDNE 模型[6]的 DynGEM 算法[14]等. 这些基于时间快照的动态网络表示学习方法往往忽略了网络的动态变化过程,仅对动态网络在时间步长上的信息进行处理的表示学习方式较为粗糙. 例如,在社交网络中,网络中的连边随着用户间的交互而不断变化,而时间步长上的网络快照只是粗粒度地反映了用户在时间步长内的累计信息,无法有效捕捉连边的形成过程. Zuo 等人提出的 HTNE 算法[15]利用霍克斯点过程建立节点的邻居形成过程,提供了一种新的动态网络表示学习思路. 但是,该方法仅考虑了历史邻居节点对当前邻居节点形成的影响,而忽略了网络演化机制对当前邻居节点形成的影响. 因此,针对现有方法存在的不足,本文提出一种综合网络历史连边信息和网络演化特性,基于霍克斯时间点过程对连边形成过程进行建模的网络表示学习方法,本文的主要贡献如下:

(1)本文提出一种针对动态网络进行表示学习的新模型,该模型基于霍克斯点过程建立节点的历史结构与当前连边的关系,从而保留了动态网络中历史信息对当前连边产生强度的影响;

(2)本文所提霍克斯点过程模型综合考虑了网络的历史连边信息和网络演化特性,更全面地捕捉了多种网络信息对当前连边产生强度的影响;

(3)应用不同真实网络数据集的实验结果表明,本文提出的动态网络表示学习方法在节点分类和链路预测等数据挖掘任务上精度具有一定的提升,实验中的 F1 分数值和 AUC 值分别提高了 3.06% ~7.18% 和 2.22% ~4.69%.

2　相关定义

定义 1(动态网络)　定义网络在 $t(0<t<T)$ 时刻的时间快照为 $G^t=(V_t,E_t)$,V_t 表示 t 时刻网络中的节点集合,E_t 表示 t 时刻网络中的连边集合. 时间 T 内的动态网络 G 可表示为包含一系列时间快照的集合 $G=\{G^1,G^2,\cdots,G^T\}$.

定义 2(动态网络表示学习)　动态网络表示学习的目标是在任意快照时刻 t 学习映射函数 $\Gamma^t:v_i\rightarrow\mathbb{R}^d(0<t<T)$,使得网络中的节点在快照时刻 t 映射成低维稠密向量$\mathbb{R}^d$. 其中,d 表示向量的维度,且满足 $d\ll V_t$. 通常情况下,映射函数的目标是保留节点在原始网络结构上的内在相似性和时间上的平滑性.

定义 3(三元闭包理论[16])　三元闭包理论通俗来讲指:在社会关系中,拥有共同朋友的个体更倾向于成为朋友. 三元闭包效应影响着网络局部结构的形成,是一种反映网络演化机制的重要特性.

定义 4(霍克斯过程[17])　霍克斯点过程是一种特殊的线性自激点过程. 在霍克斯点过程中,新事件的发生不仅受事件内部特征的影响,发生于时刻 $t_s<t$ 的历史事件也会以一定的强度影响新事件的产生. 可定义新事件产生的强度函数为:

$$\lambda(t)=\gamma_t+\sum_{t_s<t}\psi(t-t_s)\tag{1}$$

其中,γ_t 表示事件产生的基强度,$\psi(t-t_s)$ 表示历史事件对新事件的影响,其影响强度随时间而不断衰减. $\psi(t-t_s)$ 由两部分组成:一是历史事件对当前事件的激励强度,二是历史事件自身的时间衰减系数,通常可表示为 $\psi(t-t_s)=\alpha\gamma(t,t_s)$.

3　基于霍克斯点过程的动态网络表示

本文提出的基于霍克斯点过程的动态网络表示学习方法的模型框架如图 1 所示. 首先,将动态网络中新连边的产生过程建模为与之相关的两个时间序列 L_1 和 L_2. 然后基于时间序列,建立霍克斯点过程以捕捉动态网络的历史信息和演化机制对节点表示的影响.

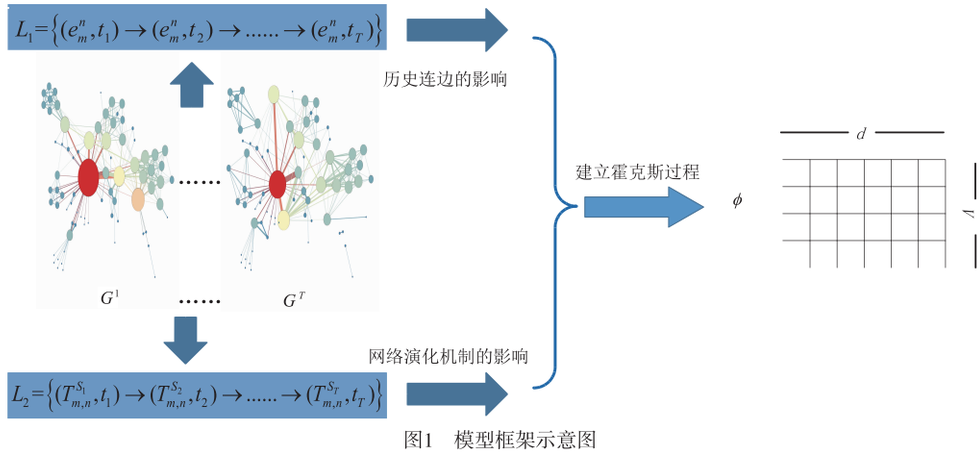

图1　模型框架示意图

3.1 模型的建立

模型的建立共分为三个部分:首先建模时间序列,以建立新连边与网络历史信息的关系;然后建立基于霍克斯点过程建模连边产生强度的模型;最后对模型进行优化和求解.

3.1.1 建模时间序列

网络的形成过程是一个连边不断产生和消失的过程,而连边的建立过程可看作一个时间序列.我们可将连边的产生定义为这一序列中的事件,其产生方式有两种:一是连边存在于历史时刻且在当前时刻也得以保留,二是连边不存在于历史时刻而在演化过程中于当前时刻形成.这两种连边产生方式分别与以下两个时间序列相关.

(1)历史连边序列 动态网络中,若 t 时刻节点 m 和 n 之间存在连边,则该连边可表示为带有时间戳的二元组 (e_m^n,t).若该连边存在于历史时刻 t_i,则可表示为 (e_m^n,t_i).那么,由节点 m 和 n 组成的连边集合可建模为时间序列 $L_1=\{(e_m^n,t_1)\rightarrow(e_m^n,t_2)\rightarrow\cdots\rightarrow(e_m^n,t_T)\}$.直观上考虑,历史上交互越多的节点在当前时刻更倾向于连接.因此,t 时刻包含节点 m 的连边的产生强度受历史连边 (e_m^n,t_i) 的影响.

(2)开三角序列 动态网络中,若 t 时刻节点 m 和 n 之间存在共同邻居 k,则组成的开三角可表示为三元组 (e_m^k,e_k^n,t).在时刻 t 由节点 m、n 及 S 个共同邻居组成的所有开三角可表示为集合 $(T_{m,n}^S,t)$,其中 $T_{m,n}^S=\{(e_m^{k_1},e_{k_1}^n),(e_m^{k_2},e_{k_2}^n),\cdots,(e_m^{k_S},e_{k_S}^n)\}$.那么,由节点 m、n 及共同邻居组成的开三角也可建模为时间序列 $L_2=\{(T_{m,n}^{S_1},t_1)\rightarrow(T_{m,n}^{S_2},t_2)\rightarrow\cdots\rightarrow(T_{m,n}^{S_T},t_T)\}$.当节点间在历史上不存在连边时,由三元闭包理论可知,若两节点具有共同邻居,那么在网络演化过程中就趋于连接.

3.1.2 基于霍克斯点过程建模连边的产生强度

霍克斯点过程的物理含义为:当前事件的发生概率,不仅受事件内部关系的影响,还受相关的历史事件影响.因此,我们可以通过建立事件内部关系和历史事件对当前事件的影响强度,来预测当前事件发生的概率.针对本文研究的动态网络,连边的产生过程可理解为霍克斯点过程,连边产生的概率即为事件发生的强度.该事件的发生强度不仅受网络中已存在的历史连边信息的影响,还受网络演化机制的影响(即开三角序列的影响).我们在这两种影响下分别基于霍克斯点过程建模连边的产生强度.

(1)基于网络的历史连边信息,当节点 m 和 n 在历史上存在连边,可利用霍克斯点过程建立如下模型:

$$\lambda_{e_m^n}^1(t)=\gamma_{m,n}+\sum_{t_s<t}\alpha_{e_y^x}\gamma(t,t_s) \tag{2}$$

其中,$\lambda_{e_m^n}^1(t)$ 表示 t 时刻节点 m 和 n 产生连边 (e_m^n,t) 的概率,$\gamma_{m,n}$ 表示 m 和 n 产生连边的基强度.累加项中 $\alpha_{e_y^x}$ 表示历史连边 (e_y^x,t_s) 对产生连边 (e_m^n,t) 的影响强度,$\gamma(t,t_s)$ 是时间衰减函数,通常可表示为指数形式 $\gamma(t,t_s)=\exp(-\theta(t-t_s))$.

基强度 $\gamma_{m,n}$ 反应了节点之间最本质最直接的联系,我们将映射空间中节点表示向量 $\boldsymbol{v}_m$ 和 $\boldsymbol{v}_n$ 之间的负欧式距离作为节点 m 和 n 的基强度,即 $\gamma_{m,n}=g(\boldsymbol{v}_m,\boldsymbol{v}_n)=-\|\boldsymbol{v}_m-\boldsymbol{v}_n\|^2$.将当前节点 m 和 n 对应的历史节点分别表示为 x 和 y,网络中包含历史节点 x 和 y 的局部结构越稳定,当前节点 m 和 n 就越趋于连接.我们可用 x 和 y 的集聚系数来表征网络在时变过程中局部结构的稳定性,局部集聚系数指网络中包含节点 r 的闭三角数与包含节点 r 的三元组数之间的比值,可公式化表示为:

$$C_r=\frac{2E_r}{m_r(m_r-1)} \tag{3}$$

其中,m_r 表示与节点 r 相连的连边数,E_r 表示与 r 相连的节点之间存在的连边数.当 m 和 n 的历史节点 x 和 y 的集聚系数越大,当前节点 m 和 n 之间产生连边 (e_m^n,t) 的概率就越大.因此,可令 $\alpha_{e_y^x}=C_xC_yg(\boldsymbol{v}_x,\boldsymbol{v}_y)$,式(2)可表示为:

$$\lambda_{e_m^n}^1(t)=g(\boldsymbol{v}_m,\boldsymbol{v}_n)+\sum_{t_s<t}C_xC_yg(\boldsymbol{v}_x,\boldsymbol{v}_y)\gamma(t,t_s) \tag{4}$$

(2)除了历史上出现的连边会以一定概率在当前时刻出现以外,在网络演化过程中,由三元闭包理论可知,历史上具有共同邻居的节点也趋于连接.即当存在历史事件 $(e_m^k,e_k^n,t_s)\in\xi$,则新事件 (e_m^n,t) 产生的概率就会提高.同样,我们也可使用节点的局部集聚系数来衡量这种影响强度:节点 k 的集聚系数越大,其附近的三元闭包过程就越强,那么 (e_m^n,t) 产生的强度就越大.同时,节点拥有的共同邻居数越多连接的可能性就越大.历史事件 $(e_m^k,e_k^n,t_s)\in\xi$ 距离当前时刻越近,连边 (e_m^n,t) 产生的概率强度越大.

因此,基于网络演化的三元闭包特性,可将式(4)更新为:

$$\lambda_{e_m^n}(t)=g(\boldsymbol{v}_m,\boldsymbol{v}_n)+\sum_{t_s<t}\Big(C_xC_yg(v_x,v_y)\gamma(t,t_s)+\sum_{(e_m^k,e_k^n,t_s)\in\xi}g(v_{k'},v_k)C_k\gamma(t,t_s)\Big) \tag{5}$$

其中,$g(v_{k'},v_k)=-\|v_{k'}-v_k\|^2$ 表示向量空间中,当前共同邻居和历史共同邻居的负欧氏距离.由式(5)可知,$\lambda_{e_m^n}(t)$ 可能是一个负值,而连边产生的概率强度应为正值.因此,我们取 $\lambda_{e_m^n}(t)$ 的指数值作为连边产生的概率强度,即 $\tilde{\lambda}_{e_m^n}(t)=\exp(\lambda_{e_m^n}(t))$.基于霍克斯点过程,当给定与 m 相关的历史连边序列 L_1 和历史开三角序列 L_2,我们可以得到节点 m 和 n 在 t 时刻产生连边 (e_m^n,t) 的概率强度为:

$$P(n \mid m, H(L_1, L_2)) = \frac{\tilde{\lambda}_{e_m^n}(t)}{\sum_{n^* \in V且e_m^{n^*} \in E} \tilde{\lambda}_{e_m^{n^*}}(t)} \tag{6}$$

3.1.3 模型优化与求解

对于网络中的所有节点,节 3.2.2 中模型的似然函数可表示为:

$$\log \ell_1 = \sum_{m \in V} \sum_{n \in V且e_m^n \in E} \log P(n \mid m, H(L_1, L_2)) \tag{7}$$

为降低算法的复杂度,我们采用负采样法优化算法[18]. 节点被选作负样本的概率与其在序列中出现的频次有关,因此我们根据节点的度分布来进行采样. 根据文献[18]采样概率可表示为:

$$P(v_i) = \frac{f(v_i)^{\frac{3}{4}}}{\sum_{j=1}^{K} f(v_j)^{\frac{3}{4}}} \tag{8}$$

基于历史连边信息和三元闭包特性,新连边(e_m^n, t)产生的目标函数可表示为:

$$O(X) = \log\sigma(\tilde{\lambda}_{e_m^n}(t)) + \sum_{i=1}^{K} \mathrm{E}_{v_i \sim P(v)}[-\log\sigma(\tilde{\lambda}_{e_m^{v_i}}(t)] \tag{9}$$

其中,K 是负样本数,$\sigma(x)$是 sigmoid 函数.

在以上基于霍克斯点过程的模型中,我们使用梯度下降法[19]优化更新目标函数,最终可得到包含丰富历史信息的节点表示. MHNE 模型训练算法流程如算法 1 所示.

算法 1　基于霍克斯点过程的动态网络表示学习

输入:动态网络 $G = \{G^1, G^2, \cdots, G^T\}$,向量维度 d,模型参数 θ,时间步长 h.

输出:潜在的节点表示矩阵 $\boldsymbol{X} \in \mathbf{R}^{V \times d}$.

```
初始化节点表示 X ∈ R^{V×d}.
for epoch in len(epochs).
    for batch(L1, L2) do.
        根据式(5)计算历史信息及网络演化机制对产生当前连边的影响.
        根据式(8)选取负样本 Neg_{E^n}.
        根据式(9)计算目标函数 O(X).
        使用梯度下降法进行梯度更新 X = X − η · ∂O(X)/O(X).
    end
end
```

4　实验与结果

对于本文提出的动态网络表示学习方法,我们将其得到的表示向量应用于节点分类、可视化和链路预测中以验证其可行性和有效性,并通过参数敏感性分析验证算法的鲁棒性.

4.1　度量标准

节点分类度量指标　常见的节点分类度量指标有精度(precision)、召回率(recall)和 F1 分数. 精度(precision)反映了被预测为 i 类且真正属于第 i 类的样本占所有被判定为 i 类样本的比重;召回率(recall)则反映了被判定为 i 类且真正属于第 i 类的样本占 i 类所有样本的比重;F1 分数指标是基于精度和召回率的调和平均,F1 分数的定义公式如下:

$$\mathrm{F1}(i) = \frac{2 \times \mathrm{precision}(i) \times \mathrm{recall}(i)}{\mathrm{precision}(i) + \mathrm{recall}(i)} \tag{10}$$

Micro-F1 和 Macro-F1 的定义如下:

$$\text{Micro-F1} = \frac{2 \times (\sum_{i=1}^{M} \mathrm{precision}(i)) \times (\sum_{i=1}^{M} \mathrm{recall}(i))}{\sum_{i=1}^{M} \mathrm{precision}(i) + \sum_{i=1}^{M} \mathrm{recall}(i)} \tag{11}$$

$$\text{Marco-F1} = \frac{\sum_{i=1}^{M} \mathrm{F1}(i)}{M} \tag{12}$$

其中,M 表示所有样本的类别数. Micro-F1 和 Macro-F1 是综合考察多分类问题的评价指标,其值越大则表示分类效果越好.

链路预测评价指标　AUC 指标可从整体上衡量链路预测的精确度,该指标可解释为从测试集中随机选取一条存在的连边的分数值$\mathrm{Score}_{a,b}$比随机选择一条不存在的连边的分数值Score_{a^*,b^*}高的概率. 连边的分数值由预测算法得到,若$\mathrm{Score}_{a,b} > \mathrm{Score}_{a^*,b^*}$,则加一分;若$\mathrm{Score}_{a,b} = \mathrm{Score}_{a^*,b^*}$,则加 0.5 分;若$\mathrm{Score}_{a,b} < \mathrm{Score}_{a^*,b^*}$,则加 0 分. 随机比较 n 次,有 n_1 次加一分,有 n_2 次加 0.5 分,AUC 值的计算公式可表示为:

$$\mathrm{AUC} = \frac{n_1 + 0.5 n_2}{n} \tag{13}$$

4.2　数据集及对比算法

本文在多个真实数据集上测试本文提出的方法,数据集的统计信息见表 1.

表 1　实验数据集

数据集	DBLP	Epinions
节点数	28085	21575
连边数	236894	2590798
节点类型数	10	5
网络平均集聚系数	0.715648	0.153612

DBLP 数据集[20]:DBLP 数据集是一个包含大量计算机科学出版物信息的数据集. 我们通过 DBLP 数据集中十个研究领域中 28085 位作者在十年内的合著关系来构建网络. 作者的所属类别为其发表文献最多的研究领域.

Epinions 数据集[21]:Epinions 数据集由用户对商品的评论信息、用户 ID、商品 ID 和时间戳等信息构成,我们从 Epinions 数据集的子集中抽取出十年内属于五个类别的 21575 位用户作为网络节点,在对同一商品有评论行为的用户间建立连边,且用户的类别由其评论最多的商品的类别决定.

将 MHNE 算法和以下三种基线算法进行对比:

Avg Deepwalk 算法. 通过在不同时间快照上应用 Deepwalk 算法得到节点在不同时刻的向量表示.

STWalk 算法[22]. 通过构造动态网络的时空扩展图,从时间维度和空间维度对网络进行表示学习.

HTNE 算法[15]. 该算法基于节点在时序网络中的邻居形成序列,利用霍克斯点过程捕捉历史邻居节点对当前邻居序列的影响,从而学习到包含历史邻居信息的节点表示.

4.3 实验与结果分析

在本节中,设计节点分类、可视化和链路预测等网络应用任务来验证提出的动态网络表示方法的可行性和有效性,并分析了算法的参数敏感性. 实验默认参数设置如下:向量维度为 128,负样本数为 5,梯度下降的学习率为 0.01. 应用到随机游走和 skip-gram 模型的算法中,设置随机游走长度为 10,游走次数为 50 次,窗口大小为 10.

4.3.1 节点分类

在 DBLP 和 Epinions 数据集上,用 SVM 分类器分别对算法的结果进行分类,取 10 次实验的均值作为最终的分类结果,实验结果如表 2 和表 3 所示.

表 2 基于 DBLP 数据集的节点分类实验结果

分类指标	算法	20%	30%	40%	50%	60%	70%	80%
Macro-F1	Avg Deepwalk	0.6127	0.6180	0.6253	0.6285	0.6310	0.6334	0.6371
	STWalk	0.6235	0.6270	0.6336	0.6413	0.6540	0.6594	0.6586
	HTNE	0.6354	0.6402	0.6521	0.6559	0.6594	0.6603	0.6559
	MHNE	**0.6452**	**0.6685**	**0.6724**	**0.6792**	**0.6890**	**0.6975**	**0.6856**
Micro-F1	Avg Deepwalk	0.6171	0.6193	0.6297	0.6302	0.6390	0.6389	0.6323
	STWalk	0.6291	0.6302	0.6382	0.6476	0.6550	0.6598	0.6612
	HTNE	0.6389	0.6427	0.6584	0.6627	0.6583	0.6613	0.6594
	MHNE	**0.6497**	**0.6711**	**0.6793**	**0.6801**	**0.6850**	**0.6983**	**0.6926**

表 3 基于 Epinions 数据集的节点分类实验结果

分类指标	算法	20%	30%	40%	50%	60%	70%	80%
Macro-F1	Avg Deepwalk	0.5221	0.5279	0.5302	0.5386	0.5321	0.5391	0.5302
	STWalk	0.5235	0.5271	0.5327	0.5367	0.5409	0.5465	0.5497
	HTNE	0.5664	0.5689	0.5786	0.5796	0.5803	0.5851	0.5867
	MHNE	**0.5821**	**0.5964**	**0.5970**	**0.6054**	**0.6127**	**0.6089**	**0.6103**
Micro-F1	Avg Deepwalk	0.5234	0.5321	0.5343	0.5392	0.5441	0.5486	0.5467
	STWalk	0.5324	0.5386	0.5401	0.5427	0.5489	0.5504	0.5526
	HTNE	0.5689	0.5703	0.5794	0.5828	0.5864	0.5893	0.5907
	MHNE	**0.5901**	**0.5989**	**0.6054**	**0.6086**	**0.6154**	**0.6121**	**0.6128**

我们设置训练集大小从 20% 增加到 80%,从各个算法的节点分类实验结果可知:在 DBLP 和 Epinions 数据集上,本文提出的 MHNE 算法较以上几种动态网络表示学习算法在节点分类上都具有更好的实验性能,其 F1 分数值最高. 在 DBLP 数据集上,当训练集占比为 70% 时,MHNE 算法的 Macro-F1 和 Micro-F1 分数值最高,较给出的对比算法分别高出 3.72% ~6.41%、3.70% ~5.94%;在 Epinions 数据集上,当训练集占比为 60% 时,MHNE 算法的 Macro-F1 和 Micro-F1 分数值最高,较给出的对比算法分别高出 3.24% ~8.06%、2.90% ~7.13%. 从实验结果可以看出,将网络历史信息融于当前节点的表示学习有利于提高节点的表示质量. 当我们利用网络的历史连边信息和网络演化特性辅助节点进行表示学习时,得到的节点表示向量的分类精度更高.

4.3.2 链路预测

通过不同的网络表示学习算法,可得到节点 m 和 n 的表示向量 $\boldsymbol{v}_m$ 和 $\boldsymbol{v}_n$. 实验中,随机在数据集中选取 5000 条存在的连边作为正样本集. 同时,随机构造 5000 条不

存在的连边作为负样本集. 然后随机从正负样本集中选取连边计算预测的精度(precision)、F1 分数值和 AUC 值(计算 AUC 值时,本文用 $\boldsymbol{v}_m$ 和 $\boldsymbol{v}_n$ 的内积$\boldsymbol{v}_m \cdot \boldsymbol{v}_n$ 表示连边 e_m^n 的分数值),为保证预测结果的准确性,分别在 DBLP 和 Epinions 数据集上独立重复实验 10 次取均值作为最终的预测结果,预测结果如表 4 和表 5 所示.

表 4　DBLP 数据集上的链路预测结果

Method	Precision	F1	AUC
Avg Deepwalk	0. 8524	0. 8502	0. 8654
STWalk	0. 8673	0. 8621	0. 8694
HTNE	0. 8789	0. 8753	0. 8901
MHNE	**0. 8969**	**0. 8957**	**0. 9123**

表 5　Epinions 数据集上的链路预测结果

Method	Precision	F1	AUC
Avg Deepwalk	0. 8321	0. 8296	0. 8421
STWalk	0. 8326	0. 8357	0. 8393
HTNE	0. 8474	0. 8523	0. 8548
MHNE	**0. 8698**	**0. 8637**	**0. 8721**

从表 4 和表 5 可以看出:MHNE 算法在 Epinions 数据集上的链路预测精度、F1 值和 AUC 值都较其它三种算法高,且都低于在 DBLP 数据集上的预测值. 在 DBLP 数据集上,MHNE 算法得到的链路预测精度、F1 值和 AUC 值较其它三种算法分别高出 1. 80% ~4. 45%、2. 04% ~4. 55% 和 2. 22% ~4. 69%. 这主要是因为本文提出的 MHNE 算法结合了多种历史信息对连边产生强度的影响且根据历史信息距当前时刻的时间远近对影响强度进行了一定的调整,使得不同时刻的历史信息在不同程度上对连边的预测起到了一定的辅助作用.

4. 3. 3　网络可视化

本文取 DBLP 数据集中属于数据挖掘、人工智能、信息检索和计算机视觉四个领域的 2663 位作者的表示向量进行可视化,并利用 t-SNE 降维算法将 128 维的表示向量将至 2 维. 实验中,使用不同颜色的点来表示不同研究领域的作者的二维可视化结果:橙色的点表示信息检索领域的作者,绿色的点表示计算机视觉研究领域的作者,紫色的点表示数据挖掘领域的作者,蓝色的点表示人工智能领域的作者.

图 2 是对不同算法得到的表示向量进行可视化的结果. 从图中可以看出,Avg Deepwalk 算法只能将信息检索领域的作者映射到一个独立的社区,其他三个领域的作者混淆在一起被映射到相同的社区;STWalk 算法将数据挖掘领域的作者映射到相对分散的位置,没能保留该类节点的属性,而信息检索和计算机视觉领域的作者被映射到同一社区难以区分;HTNE 算法能将人工智能、信息检索和计算机视觉领域的作者映射到不同的社区,但将人工智能领域的部分作者映射到了数据挖掘领域中;相较于其它算法,本文提出的 MHNE 算法能将属于不同领域的作者映射到不同社区,其可视化效果最好.

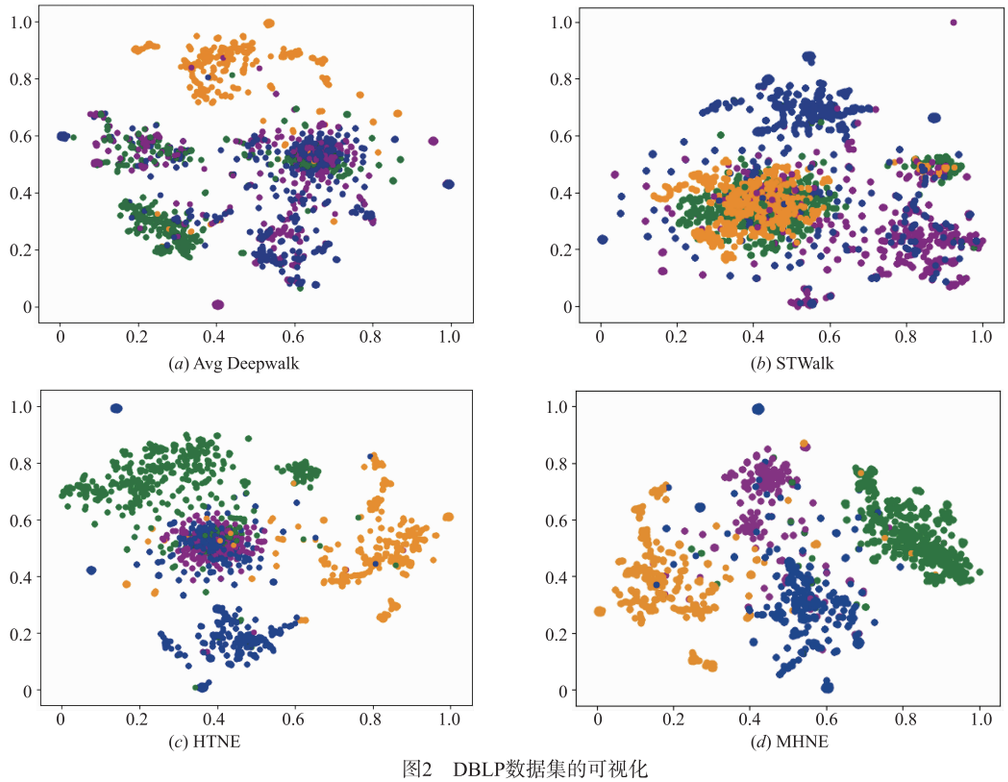

图2　DBLP数据集的可视化

4.3.4 参数 θ 的敏感性分析

时间衰减函数可表示为指数形式:$\gamma(t,t_s)=\exp(-\theta(t-t_s))$,其中 θ 表示历史事件激励的时间衰减系数. 我们通过改变 θ 的大小来观察 Micro-F1 指标和 AUC 指标的变化情况,以对算法的敏感性进行分析. 分别在 DBLP 和 Epinions 数据集上进行实验,设置训练集占比为 70%,θ 的变化范围为 0.01 ~ 1. 由式(2)可知:θ 越大,历史事件对当前事件的影响越小.

实验结果如图 3 和图 4 所示. 从图中可以看出,参数 θ 的最佳取值在两个数据集上具有一定的差异. 在 DBLP 数据集上,当 θ 取 0.15 时,得到的分类精度和链路预测精度最高,当 $0.15<\theta<0.4$ 时,分类精度和链路预测精度基本保持不变,当 $\theta>0.4$ 时,分类精度和链路预测精度随 θ 的增大略有减小;在 Epinions 数据集上,当 θ 取 0.3 时,节点分类精度最高,θ 取 0.2 时,链路预测精度最高,$\theta>0.4$ 时,分类精度和链路预测精度随着 θ 的增大而大幅减小.

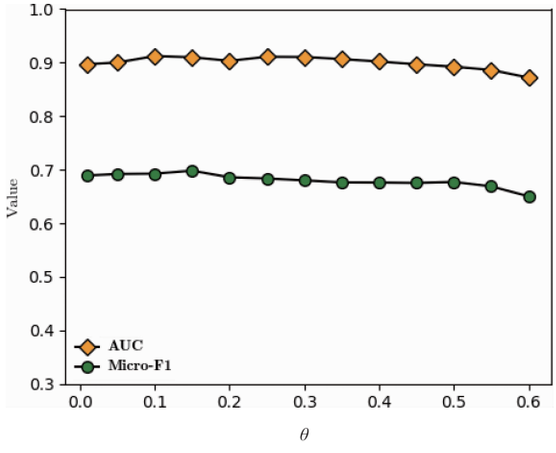

图3 参数θ在DBLP数据集上的敏感性分析

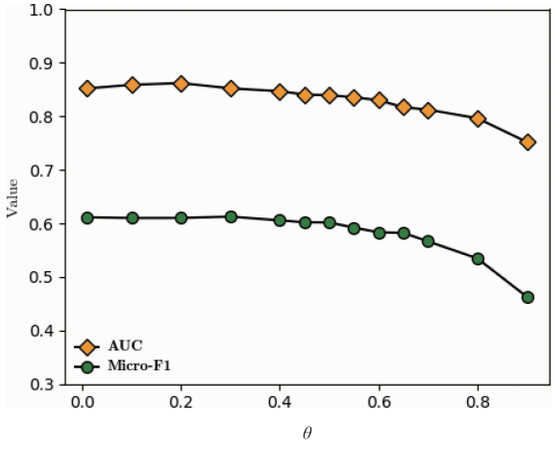

图4 参数θ在Epinions数据集上的敏感性分析

造成两种数据集出现不同实验结果的主要原因是:DBLP 数据集由作者及其合著关系构成,作者间的历史合著关系在短时间内不会产生太大变化. 因此,DBLP 数据集的历史信息的时间衰减程度较小,θ 的取值应较小. 而 Epinions 数据集通过用户及用户的评论行为构成社交网络,其历史信息的时间衰减程度较大,θ 的取值应较大.

5 结束语

为结合网络的动态特性对网络进行表示学习,本文提出一种基于霍克斯点过程的动态网络表示学习方法. 通过建模网络的历史结构信息和演化机制对连边产生强度的影响来训练节点的表示向量,该模型提高了表示向量的表示质量,并有效应用于节点分类和链路预测中. 目前,针对网络动态特性的研究仍处于起步阶段,本文仅考虑了节点和连边都只有一种类型的同质网络的动态特性. 而现实网络兼具异质性和动态性,如何将丰富的异质信息考虑进动态网络的表示学习是下一个研究重点.

参考文献

[1] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323 – 2326.

[2] 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.

[3] CAO S, LU W, XU Q. GraRep: Learning graph representations with global structural information[A]. ACM International on Conference on Information and Knowledge Management[C]. USA: ACM, 2015. 891 – 900.

[4] PEROZZI B, A1-RFOU R, SKIENA S. Deep Walk: online learning of social representations[A]. ACM Sigkdd International Conference on Knowledge Discovery & Data Mining[C]. USA: ACM, 2014. 701 – 710.

[5] GROVER A, LESKOVEC J. Node2vec: Scalable feature learning for networks[A]. ACM Sigkdd International Conference on Knowledge Discovery & Data Mining[C]. USA: ACM, 2016. 855 – 864.

[6] WANG D, CUI P, ZHU W. Structural deep network embedding[A]. ACM Sigkdd International Conference on Knowledge Discovery & Data Mining[C]. USA: ACM, 2016. 1225 – 1234.

[7] WANG Hongwei, WANG Jia, WANG Jialin, et al. GraphGAN: graph representation learning with generative adversarial nets[A]. Thirty-Second AAAI Conference on Artificial Intelligence[C]. USA: AAAI, 2018. 2508 – 2515.

[8] DAI Q, LI Q, TANG J, et al. Adversarial network embed-

ding[A]. Thirty-Second AAAI Conference on Artificial Intelligence[C]. USA:AAAI,2018. 2167 – 2174.
[9] LI J,DANI H,HU X,et al. Attributed network embedding for learning in a dynamic environment[A]. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management[C]. USA:ACM,2017. 387 – 396.
[10] ZHU D,CUI P,ZHANG Z,et al. High-order proximity preserved embedding for dynamic networks[J]. IEEE Transactions on Knowledge and Data Engineering,2018,30(11):2134 – 2144.
[11] ZHANG Z,CUI P,PEI J,et al. Timers:Error-bounded SVD restart on dynamic networks[A]. Thirty-Second AAAI Conference on Artificial Intelligence[C]. USA:AAAI,2018. 224 – 231.
[12] TANG J,QU M,WANG M,et al. LINE:Large-scale information network embedding[A]. Proceedings of the 24th International Conference on World Wide Web[C]. USA:ACM,2015. 1067 – 1077.
[13] DU L,WANG Y,SONG G,et al. Dynamic network embedding:An extended approach for skip-gram based network embedding[A]. International Joint Conferences on Artificial Intelligence Organization[C]. IJCAI,2018. 2086 – 2092.
[14] GOYAL P,KAMRA N,HE X,et al. Dyngem:Deep embedding method for dynamic graphs[A]. International Joint Conference on Artificial Intelligence (International Workshop on Representation Learning for Graphs)[C]. IJCAI,2017. arXiv:1805. 11273.
[15] ZUO Y,LIU G,LIN H,et al. Embedding temporal network via neighborhood formation[A]. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining[C]. USA:ACM,2018. 2857 – 2866.
[16] ZHOU L,YANG Y,REN X,et al. Dynamic network embedding by modeling triadic closure process[A]. Thirty-Second AAAI Conference on Artificial Intelligence[C]. USA:AAAI,2018. 571 – 578.
[17] HAWKES A G. Spectra of some self-exciting and mutually exciting point processes[J]. Biometrika,1971,58(1):83 – 90.
[18] MIKOLOV T,SUTSKEVER I,CHEN K,et al. Distributed representations of words and phrases and their compositionality[A]. Advances in Neural Information Processing Systems[C]. MIT,2013. 3111 – 3119.
[19] BOTTOU,L. Stochastic gradient learning in neural networks[J]. Proceedings of Neuro Nimes,1991,91(8):12.
[20] MOREIRA C,CALADO P,MARTINS B. Learning to rank academic experts in the DBLP dataset[J]. Expert Systems,2015,32(4):477 – 493.
[21] TANG J,GAO H,LIU H. mTrust:discerning multi-faceted trust in a connected world[A]. Proceedings of the Fifth ACM International Conference on Web Search and Data Mining[C]. USA:ACM,2012. 93 – 102.
[22] PANDHRE S,MITTAL H,GUPTA M,et al. STwalk:learning trajectory representations in temporal graphs[A]. Proceedings of the ACM India Joint International Conference on Data Science and Management of Data[C]. USA:ACM,2018. 210 – 219.

作者简介

尹　赢　女,1994 年出生,四川绵竹人. 2017 年毕业于西安交通大学电信学院,2017 年进入国家数字交换系统工程技术研究中心. 现为硕士研究生,主要从事网络表示学习的有关研究.
E-mail:15883880517@163. com

张建朋(通信作者)　男,1988 年出生,河北廊坊人. 助理研究员. 2018 年获得埃因霍温理工大学博士学位. 主要的研究方向包括数据挖掘、大数据分析以及社会网络分析.
E-mail:zjp@ ndsc. com. cn

吉立新　男,1969 年出生,江苏淮安人. 研究员. 现为国家数字交换系统工程技术研究中心总工程师,主要研究方向为电信网分析.
E-mail:jlxndsc@139. com

李治成　男,1996 年出生,云南昌宁人. 2018 年毕业于四川大学计算机学院,2018 年进入国家数字交换系统工程技术研究中心. 现为硕士研究生,主要从事复杂网络相关研究.
E-mail:lizc520a@ gmail. com