计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210300030

一种基于监督学习的异构网链路预测模型

黄寿孟
三亚学院信息与智能工程学院　海南 三亚 572022

摘　要　传统的异构网链路预测研究有基于元路径监督学习的 PathPredict 算法与 MPBP 算法,但它们并不能充分利用异构网提供的丰富信息来进行链路预测。在原有传统监督学习算法的基础上,首先为了增加链路熵和时间动态信息而设计了 HLE-T 算法,然后通过链路强弱关系的数值分段构建多分类问题的监督学习算法 MSLP 链路预测模型,最后在 4 个稠密度不同的数据集下完成了实验测试。实验结果表明,MSLP 链路预测模型一定程度上提升了异构网中的链路预测性能,对未来链路预测研究具有一定的借鉴意义。

关键词:监督学习;异构信息;链路预测;预测模型
中图法分类号　TP309

Heterogeneous Network Link Prediction Model Based on Supervised Learning

HUANG Shou-meng
College of Information and Intelligence Engineering,Sanya University,Sanya,Hainan 572022,China

Abstract　The research on traditional heterogeneous network link prediction has path-predicted algorithm and MPBP(meta-path feature-based backpPropagation neural network model) algorithm based on the metapath supervised learning. However, they can't make full use of the rich information provided by heterogeneous network to make link prediction. Based on the traditional supervised learning algorithm,this paper first designs the HLE-T(heterogeneous link entropy with time) algorithm in order to increase the link entropy and time dynamic information. Moreover,it constructs the MSLP(modified supervised link prediction) model of the Supervised learning algorithm with the multi-classification problem by the numerical segment of the link strength and weak relationship,and finally completes the experimental test under four data sets with different density. The experimental results show that the MSLP model improves the link prediction performance in heterogeneous network to some extent,and has some reference significance for the future link prediction research.

Keywords　Supervised learning,Heterogeneous information,Link prediction,Predictive model

1　引言

信息网络中的链路预测(Link Prediction)是指利用现有的网络节点关系、网络结构要素等已知信息,通过算法分析来预测可能出现的下一个网络链接节点[1]。目前链路预测的应用场景广泛,有构建复杂网络演化模型[2-3]、刻画网络中节点的相似性[4]、生物网络中研究蛋白质相互作用网络和新陈代谢网络[5]、在线社交网络中的朋友推荐,还有通过已知部分节点类型的网络中预测尚无标签节点的类型,例如判断一篇学术论文的类型,判断一个手机用户是否产生了切换运营商的念头,以及预测缺失链接,预测网络中的错误链接,有利于网络重组和结构功能优化等方面。

在异构信息网中,可根据研究任务的不同性质(如节点相似性、链路预测、节点推荐等)对其分类研究,比如依据异构信息节点相似性来构建链路预测算法,并广泛应用于各类推荐系统中。传统的单一类型节点和连边的网络建模方式过于简单,容易忽略不同类型实体间本有的相关信息,于是国内外的学者们开始转向对包含了多类型节点和连边的异构信息网络进行研究。早期,Sun 等[6]在研究合著关系时提出了一种基于监督学习的 PathPredict 算法,该算法通过监督学习的训练,对影响合著关系的元路径加权来进行链路预测。此方法并没有考虑节点属性信息,利用节点拓扑信息或节点自身额外信息计算它们之间的相似度,得分越高的节点对被认为越容易在未来形成链接,但是网络节点拓扑信息容易得到而节点自身的属性有时难以得到[7]。为此,Zhang 等[8]提出 MACP 算法,加入节点属性信息及元路径的传递相似性,在综合考虑各种相似性指标以及节点自身属性、网络隐藏属性等信息的基础上,将链路预测问题考虑建模为一个基于监督学习的分类模型,实验证明该方法比 PathPredict 算法的预测效果好。为了更好地进行监督学习下的异构网络链路预测,Liang 等[9]提出通用框架,采用监督排序的算法实现异构网络节点信息加权的预测。基于学习的算法的优势在于,它可以充分融合网络提供给我们的各种各样的信息,将其组合成特征向量,并且通过训练,自动获得对这些特征的最优加权,从而提高预测性能。之后,Li 等[10]把这种二分类监督学习模型改进为三层神经网络的全新集成链路预测模型 MPBP,通

基金项目:海南省高等学校科学研究一般项目(Hnky2011-51)
This work was supported by the Education Department of Hainan Province(Hnky2021-51).
通信作者:黄寿孟(huang123888@126.com)

过监督训练得到神经网络的参数，进而对可能存在节点进行预测，这样势必会造成训练过程中正负样例不均衡的现象，从而影响数据预测结果。

传统的基于学习的链路预测算法将网络中所有的节点对进行训练，这样也会造成训练数据集的庞大[11]。针对这一点，我们可以通过一定策略选取网络中的部分节点对来构造训练集，这样既能减少训练模型所需的时间，又不会降低预测性能。另一方面，上述研究内容围绕异构网的元路径相似性来预测未来的目标节点，但是随着元路径相似性的集合越来越庞大，链路预测的性能反而会降低，那么如何改变元路径相似性度的影响呢？本文基于异构网丰富信息的节点相似度影响信息，比如时间信息、链路熵，重新提取节点对之间的标签进行监督训练，从而提高链路预测的性能。

2 相关工作

在异构网的链路预测研究中，影响两个形成链路节点的因素有多种[12]。以预测合著关系的学术网为例，不同的元路径节点对节点间形成目标元路径的影响不同，这就要采用一种监督学习的方式来训练得到一组与链路预测任务最合适的优化模型。

2.1 PC 算法

PC(Path Count)算法[13]是预测异构网中基于元路径相似的链路技术，该算法认为元路径数量越多，形成目标路径的实例可能性越大，说明路径相似性就越高。PC 算法可以表示为：

$$PC(x,y)=|\{p_{x\rightsquigarrow y}:p_{x\rightsquigarrow y}\in\mathscr{P}\}| \tag{1}$$

其中，x 和 y 表示两个预测节点，$p_{x\rightsquigarrow y}$ 表示从节点 x 到节点 y 之间的路径实例，$|\mathscr{P}|$ 表示元路径的数量。

2.2 PathSim 算法

Sun 等人提出了异构网中相邻节点相似度 PathSim 算法[6]，该算法是在 PC 算法基础上改进而来，相邻节点之间的相似度不仅与元路径数量有关，还与该节点度值有关，度值大的节点，其相似度高的可能性大。该算法的计算方法如下：

$$PathSim(x,y)=\frac{2\times PC(x,y)}{PC(x,x)+PC(y,y)} \tag{2}$$

2.3 MMI 算法

MMI 算法[10]是改进优化的 PathSim 算法，计算节点相似度时添加节点额外信息来反映节点全局重要性，比如合著网络中在作者论文节点上添加被引数量的额外信息，有利于提升预测准确度。算法的相似度计算如下：

$$MMI(x,y)=\frac{PC(x,y)\times sim_T(x,y)}{ext(T,x)+ext(T,y)} \tag{3}$$

其中，$ext(T,x)$表示异构网中节点 x 在 T 类型上的一种额外信息。

$$sim_T(x,y)=\frac{\min\{ext(T,x),ext(T,y)\}}{\max\{ext(T,x),ext(T,y)\}} \tag{4}$$

2.4 PathPredict 算法

PathPredict 算法是传统的基于监督学习的元路径预测算法[14]。对于学术合著关系，首先利用 PC 算法和 PathSim 算法计算学术网络中合著关系节点对的元路径集合作为训练集，并将训练集中的节点对组成特征向量进行分类[15]，特征向量值标记为 0 或 1，“0”代表没有合著关系，“1”代表有合著关系，最后采用监督学习 LR 分类算法作为链路预测模型。

2.5 MPBP 算法

MPBP 算法再次优化改进 PathPredict 模型，可预测多个链路关系[16]。它采用 BP 神经网络捕捉异构网中隐藏的复杂信息作为三层神经网络的预测框架，提取出异构网中节点对的元路特征数量作为元路径特征值并将此值作为输入，分类结果作为输出，从而训练出 BP 神经网络的链路预测模型，完成多个链路预测任务。

3 链路预测模型

3.1 问题描述

异构网 $G=(V,E)$中，历史时间段 $T_0=[t_0,t_1]$内，节点 A_1与节点 A_{i+1}之间存在一条目标元路径 T_p，预测在未来时间段 $T_1=(t_1,t_2]$内节点 A_1与节点 A_{i+1}之间是否形成关于目标元路径 T_p的路径实例。对于图 1 所示的学术异构网，根据时间段 T_0的合著关系数据预测时间段 T_1的链路关系，比如若想预测作者 A_2与 A_4是否产生合著关系，在图 1 中可找到 T_0时段内存在一条路径 $A_2\rightarrow P_2\rightarrow P_3\leftarrow P_4\leftarrow A_4$，为了方便，可省略一条元路径中不同节点类型之间的关系，即表示成 $A\rightarrow P\leftarrow A$。

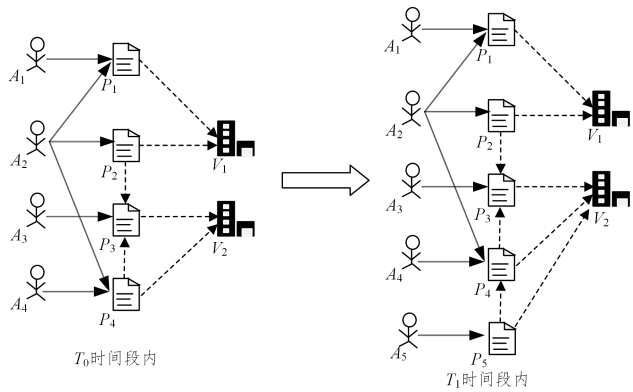

图 1 学术网络链路预测图

Fig. 1 Academic network link prediction

3.2 链路预测模型框架

异构网链路预测目标链路形成时会受到不同路径实例的节点时间标签以及节点对链路形成的贡献(或者说是权重)所影响，传统的做法是凭借先验知识人工提取。本文采用遍历网络模式自动生成与目标元路径相关联的路径集合 T_P，然后从历史时间段内计算已有相关元路径节点的链路熵与时间标签结合起来的元路径特征空间，考虑到网络结构因素，产生目标链路的节点对之间的距离一般不超过 S_P 集合中最长元路径的长度，即采用长度受限元路径生成算法(简称 MLMG 算法)获取与目标元路径相关集合，通过改进后的监督学习算法预测训练集为最大长度的元路径节点链接强弱信息。其链路预测模型的整体框架如图 2 所示。

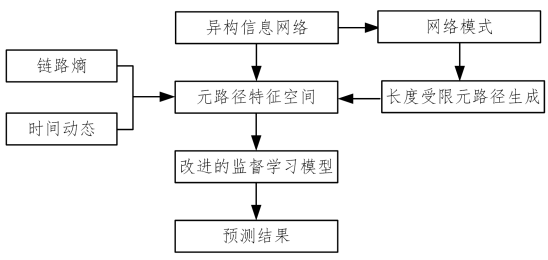

图 2 监督学习的异构网链路预测模型框架图

Fig. 2 Monitor the heterogeneous network link prediction model framework for learning

3.3 链路预测模型伪代码

3.3.1 MLMG 算法

根据异构网模式的存储结构,从目标元路径的首节点 s 开始,采用深度优先方式进行搜索,找到一组最大长度为 K 的相关元路径集合 S_P。MLMG 算法伪代码如算法 1 所示。

算法 1 MLMG

Input:网络模式 T_G,受限最大长度 k,目标元路径首节点类型 s,尾结点类型 t,临时路径的栈 stack

Output:相关元路径集合 S_P

1. 初始化 S_P←Ø,stack←Ø
2. 栈中元素个数 sl←0
3. while 当前的受限长度 l 小于 k do
4. 把节点 s 加入栈 stack 中,sl←sl+1
5. if sl==l 且 s==t then
6. 把 stack 中结果作为一相关条元路径加入到 S_P
7. return 到 MLMG 算法的上一层
8. if sl>l then
9. return 到 MLMG 算法的上一层
10. for j∈节点 s 的相邻节点 do
11. 把节点 j 加入到栈中
12. s←j,sl←sl+1
13. 递归进入到 MLMG 算法的下一层
14. 把节点 j 退栈,sl←sl−1
15. end for
16. end while
17. return S_P

3.3.2 链路熵与时间信息结合的 HLE-T 算法

异构网的路径实例节点的链路镝与时间动态信息是预测目标链路节点对之间的元路径相似度的关键因素,现有的元路径相似度算法中并不计算出不同类型节点的链路熵与时间信息的衰减统计特征。本文引入链路熵,能有效区分不同节点的影响,结合时间衰减函数统计节点时间动态信息,得到异构网的目标链路相似度。HLE-T 算法伪代码如算法 2 所示。

算法 2 HLE-T

Input:异构网 G(V,E)中从节点 x 至节点 y 的相关元路径 $p=A_i\to\cdots\to A_j$,节点 $x\in A_i$,节点 $y\in A_j$,时间衰减因子 λ,数据集中的最新时间 t_c

Output:节点 x 和节点 y 在元路径 p 下的相似度 $\mathrm{Sim}_{xy}^{\mathcal{P}}$

1. 初始化 $\mathrm{Sim}_{xy}^{\mathcal{P}}$←0,sumEntropy←0
2. 初始化 λ←0.8,tc←数据集中的最新时间
3. if x in G then
4. for A_{i+1} in 节点 x 的相邻节点 do
5. pathInstance←[ ]
6. tempEntropy←0
7. for A_{i+2} in 节点 A_{i+1} 的相邻节点 do
8. if y in 节点 A_{i+2} 的相邻节点 do
9. pathInstance.append([x,A_{i+1},A_{i+2},y])
10. t_p←当前 pathInstance 建立时间
11. 计算时间系数 $f(t_p)$
12. 计算 tempEntropy
13. sumEntropy← sumEntropy+tempEntropy
14. end if
15. end for
16. end for
17. end if
18. 计算 $I(L_{x,y}^1)$
19. $\mathrm{Sim}_{xy}^{\mathcal{P}}$←sumEntropy−$I(L_{x,y}^1)$
20. return $\mathrm{Sim}_{xy}^{\mathcal{P}}$

第 11 行代码计算时间系数表示为时间衰减函数 $f(t_p)=\lambda^{t_c-t_p}$,第 12 行代码计算 $tempEntropy=f(t_p)\cdot\sum_{i=1}^{l-1}I(L_{a_i,a_{i+1}}^1)$,其中节点 x 和 y 在未来时间段内形成目标路径实例的事件定义为 $L_{x,y}^1$,若没有形成目标路径则定义为 $L_{x,y}^0$,而 $I(L_{x,y}^1)$ 则表示节点 x 和 y 形成目标路径实例的先验概率。

$$I(L_{x,y}^1)=1-I(L_{x,y}^0)=1-\frac{C_{M_{ij}-k_x^i}^{k_y^i}}{C_{M_{ij}}^{k_y^i}} \tag{4}$$

其中,k_y^i 表示节点 y 的 A_i 类型的邻居的数量,k_x^i 表示节点 x 的 A_j 类型的邻居的数量,M_{ij} 表示类型为 A_i 的节点与 A_j 类型节点之间链接的数量。

3.3.3 监督学习优化模型 MLSP 算法

异构网中两节点对形成目标元路径实例的数量越多,它俩的链接强度就越强,同时两节点间的公共邻居个数越多则它俩的链接强度也越强。采用传统的邻接矩阵 $\boldsymbol{A}$ 来表示两节点间的公共邻居,那么两节点的链接强弱矩阵 $\boldsymbol{LS}$ 表示为 $\boldsymbol{LS}=\boldsymbol{A}+\alpha\boldsymbol{A}^2$,其中 α 为路径惩罚参数,用于调节预测节点的公共邻居对形成链路的影响的强弱,α 越大说明节点公共邻居对未来链路形成的影响越强,其取值范围为 $0<\alpha<1$。

对预测节点对[x,y]来说,LS_{xy} 表示节点 x 和 y 之间的链接强弱关系值,按值可划分成 3 种 $label_{xy}$ 取值,如下所示:

$$label_{xy}=\begin{cases}0, & LS_{xy}<1\\ 1, & 1\leqslant LS_{xy}<2\\ 2, & LS_{xy}\geqslant 2\end{cases} \tag{5}$$

这样就将原来的监督学习二分类问题转化了三分类问题,$label_{xy}=0$ 代表节点对[x,y]没有链接,$label_{xy}=1$ 代表节点对[x,y]会产生弱链接,$label_{xy}=2$ 代表节点对[x,y]产生强链接。将这 3 个类别值作为监督学习的数据训练集,得到预测模型的参数,然后在测试阶段来检测模型的优劣。

当训练集中有 n 个节点对{($x_{[1]}$,$y_{[1]}$),($x_{[2]}$,$y_{[2]}$),…,($x_{[n]}$,$y_{[n]}$)}时,按照 $label_{xy}$ 取值分类,对每个节点对 $x_{[i]}$ 表示一个预测节点对,$y_{[i]}$ 表示 $x_{[i]}$ 节点对的标签,$y_{[i]}\in\{0,1,2\}$,则可用下面的假设函数表示 $x_{[i]}$ 节点对在不同类别下的概率。

$$\begin{aligned}h_\theta(x_{[i]})&=\begin{bmatrix}P(y_{[i]}=0\mid x_{[i]};\theta)\\ P(y_{[i]}=1\mid x_{[i]};\theta)\\ P(y_{[i]}=2\mid x_{[i]};\theta)\end{bmatrix}\\ &=\frac{1}{\sum_{j=0}^{2}e^{\theta_j^{\mathrm{T}}x_{[i]}}}\begin{bmatrix}e^{\theta_0^{\mathrm{T}}x_{[i]}}\\ e^{\theta_1^{\mathrm{T}}x_{[i]}}\\ e^{\theta_2^{\mathrm{T}}x_{[i]}}\end{bmatrix}\end{aligned} \tag{6}$$

在式(6)中,参数 θ 是整个函数的核心,可定义下面的代价函数得出参数 θ。

$$J(\theta)=-\frac{1}{m}\sum_{i=1}^{m}\sum_{j=0}^{1}1\{y_{[i]}=j\}\log\frac{e^{\theta_2^{\mathrm{T}}x_{[i]}}}{\sum_{l=1}^{2}e^{\theta_l^{\mathrm{T}}x_{[i]}}}+\frac{\lambda}{2}\sum_{i=0}^{2}\sum_{j=0}^{n}\theta_{ij}^2 \tag{7}$$

采用梯度下降算法求得代价函数式(7)的极小值,再用式(6)假设函数在测试集上进行预测,计算出节点对 $x_{[i]}$ 对应类型的概率,从而得出 $y_{[i]}$ 可能产生链路的概率值。

传统监督学习方法都是将训练数据转化成二分类问题,即将节点信息训练标签设置为 1 或 0(有链接即为 1,否则为

0)，进行训练学习，容易造成部分网络结构信息的丢失，因为它并没有充分考虑到节点之间的结构关系对监督训练的影响。而 MLSP 算法改进传统二分类模型的不足，充分考虑到节点链路的不同强弱关系，采用链接强弱关系矩阵对节点路径加权分类，即用节点对之间的标签作为分类标签，不再是简单的 0 或 1 二分类标签，它已经将节点对之间的强弱关系细化成多个类别，即转化成一个多分类问题，这样的算法模型充分考虑到了异构网节点的丰富信息，有利于提高预测算法性能。

4 实验分析

4.1 测试数据集

本文实验数据集采用目前学术界普遍认可的 Aminer 数据集[17]，其来自 DBLP 网络数据库，由清华大学唐杰等提供。由于 Aminer 数据集量太大，本文选取下面的子集来测试。

训练集按年份取 $T_0=[2001,2006]$，$T_1=[2007,2010]$，测试集取 $T_0'=[2005,2010]$ 和 $T_1'=[2011,2014]$。将出版论文与作者信息构建成一个子网络，按作者发表论文的情况从 Aminer 数据集中取 4 个数据子集 D_2，D_4，D_6，D_8，详细数据如表 1 所列。

表 1 T_0时间段内的节点信息

Table 1 Node information for the T_0-period

数据集	作者数	论文数	$A\rightarrow P$	$P\rightarrow P$	$P\rightarrow V$
D_2	5056	11983	16765	7621	11983
D_4	1366	5490	7362	3218	5490
D_6	625	3928	4821	2522	3928
D_8	355	2819	3826	1827	2819

其中，数据集 D_n 表示在数据集中作者至少发表 n 篇论文，如 D_2 代表在 T_0 时间段内该数据集中每位作者至少发表 2 篇论文。显然 n 值越大，节点数量越少，节点边数越多，网络越稠密。

4.2 评测指标

目前对链路预测效能的评价指标有 Precision(精确度)、Accuracy(准确率)、Recall(召回率)、ROC 曲线、AUC 等，这里采用最常用的两个性能评价指标：AUC 和 Precision@L。

4.2.1 AUC 指标

AUC 指标是指在链路预测模型中，预测算法对测试未来链路给出的形成分值或链接概率。AUC 指标就是比较测试集中的边的相似值与不存在的边的相似性的大小。一般可通过抽样比较方式得到 AUC 指标值，即从测试集中随机抽取样本中真实存在的链路和不存在链路为一组测试，将每组测试结果记录下来。VT_i 表示真实存在节点的预测值，VF_i 表示真实不存在节点的预测值，则每组 AUC 指标的计算方法如下：

$$AUC=\begin{cases}1, & VT_i>VF_i\\ 0.5, & VT_i=VF_i\\ 0, & VT_i<VF_i\end{cases} \tag{8}$$

经过 n 组抽样比较后，假设 $AUC=1$ 的次数为 n_1，$AUC=0.5$ 的次数为 n_2，$AUC=0$ 的次数为 n_3，那么 AUC 值综合得分如下：

$$AUC=\frac{1\times n_1+0.5\times n_2+0\times n_3}{n}=\frac{n_1+0.5n_2}{n} \tag{9}$$

4.2.2 Precision@L 指标

Precision@L 指标只考虑前 L 位是否预测正确，该指标用来检测链路的正确预测连边数量，表示在异构网中预测得分排名前 L 个链接中真实存在的数量为多少。假设异构网中某预测算法得分排名前 L 个预测点中有 m 个预测准确，那么 Precision@L 指标的值为 m，其算法的精确度可定义为 m/L。当然 Precision@L 的大小与 L 的值有关，但是对于两个 AUC 值相近的算法，Precision@L 值大的算法其预测精确度高。

4.3 结果分析

4.3.1 不同元路径相似度预测分析

本次实验设置元路径长度 $k=4$，时间衰减因子 $\lambda=0.8$，数据集用 D_n 来测试算法 PC，PathSim，MMI，HLE-T 不同元路径相似度的性能情况。实验结果如图 3 所示，HLE-T 算法的 AUC 指标值反映了链路预测结果的整体水平，在 D_n 4 个数据集中的表现均好于其他 3 个算法，说明在不同异构信息网络中 HLE-T 算法对于元路径预测效果极佳。这是因为与其他 3 个算法相比，HLE-T 算法考虑到不同元路径节点对预测结果的影响，同时引入节点时间标签作为额外的路径信息，提取到更多的元路径特征，从而有效提高了预测的整体性能。另外，随着异构网络中节点对越来越多，HLE-T 算法的 AUC 预测结果越来越精确。从实验结果图 3 中得到，PathSim 算法的 AUC 预测效果较差，这是因为 PathSim 算法只是为了寻求地位对等的节点，忽略了节点的其他可见信息，同时说明异构网中地位对等的节点在链路形成的贡献中不一定是最佳的。

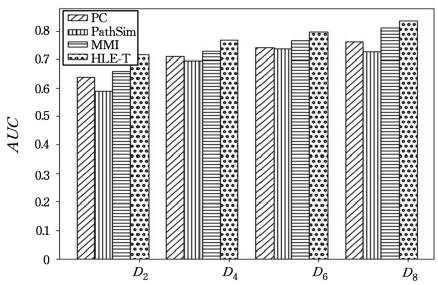

图 3 AUC 指标的结果对比图

Fig. 3 Comparison of results for the AUC indicator

从表 2 所列的 Precision@20 指标可以看出，HLE-T 算法稍微好于其他 3 类算法；但从表 3 所列 Precision@100 指标来看，HLE-T 算法明显优于其他 3 类算法。这是因为 Precision@L 指标反映预测结果有多少个被正确预测，所以在 Precision@20 指标上，由于测试集真实存在链接节点对较多，因此 4 种算法检测出的前 20 的节点对中大多数会存在链接；但是 $L=100$ 的节点对中，由于 HLE-T 算法引入时间标签作为元路径的额外信息，更加合理地利用到了异构网中的信息，这时 HLE-T 算法优势明显体现出来，能更好地寻找最有可能存在链接的节点对，从而验证了 HLE-T 算法在结合节点链路熵与时间信息下的有效性。

表 2 Precision@20 指标的对比

Table 2 Comparison of Precision@20

数据集	D_2	D_4	D_6	D_8
PC	18	17	17	15
PathSim	12	11	10	11
MMI	18	17	17	17
HLE-T	19	18	18	19

表 3 Precision@100 指标对比

Table 3 Comparison of Precision@100

数据集	D_2	D_4	D_6	D_8
PC	64	59	58	56
PathSim	35	41	42	38
MMI	71	65	66	60
HLE-T	76	76	70	64

同时,我们发现从数据集 D_2-D_8 的不同稠密程度来说,每种算法在稠密数据集的 *AUC* 值都呈现了上升趋势,而 Precision@L 则呈现下降趋势。这是因为 *AUC* 值在网络相对稠密的节点连边数量上升,而正负样例的失衡比例下降,这对于监督学习模型的学习过程是有利的,所以所有算法的 *AUC* 值上升。但是随着网络节点越来越多,网络中节点对的真实绝对连边的数量会越来越少,所以 Precision@L 指标也会越来越低,而这种趋势不影响更加合理利用节点信息的 HLE-T 算法与其他算法的对比。

4.3.2 不同监督学习模型的预测分析

实验参数设置如下:MLMG 算法中自动生成长度 $k=4$,MLSP 算法中 $\lambda=0.8$,$\alpha=0.01$,而 PathPredict 模型采用二分类算法,MPBP 算法按文献[10-11]中的参数设置:隐藏层神经元的数量在范围[6,18]中取最优值。实验结果如图 4 所示,MLSP 算法在 4 个数据测试集中的 *AUC* 值都高于其他算法,这是因为 MLSP 算法更好地学习到异构网更多的合理语义信息,在进行监督学习时能够比其他二分类模型更好地整合节点对的链接概率。MLSP 算法合理扩充了异构网训练集中的标签信息,通过节点的链接强弱信息,将标签标记转化为节点强弱关系矩阵,为训练集节点对分配链接强弱值,采集到更丰富的标记信息,从而转化为多分类问题,再根据不同类别的概率来预测计算节点链接概率。

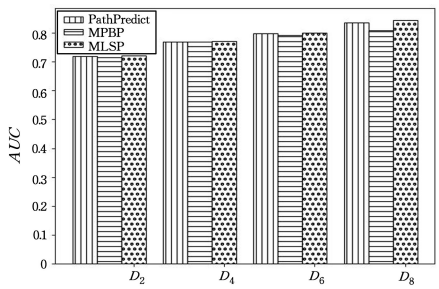

图 4 不同监督学习算法的 AUC 性能对比

Fig. 4 Comparison of AUC performance between different supervised learning algorithms

如表 4 所列,MLSP 算法在 Precision@20 指标上的表现也是最佳状态;而在表 5 所列的 Precision@100 指标上中,MLSP 算法只显示在前 3 个数据集上表现最佳,这是因为随着网络稠密度增加,MLSP 算法模型中获取节点对的链路信息更细节,各个监督学习模型进行分类时链接强弱关系的概率值会变小,所以预测精确度反而降低。MLSP 算法中的分类器能够学习到更多的节点强弱信息,而这些信息正好与节点链接预测概率相关,在对所有预测概率排序后,预测节点关系越强的,排名越靠前,所以 MLSP 算法在 Precision@L 指标上更加充分利用了丰富网络信息,它的预测效果必然比其他的基准算法好。

表 4 不同监督学习算法的 Precision@20 指标对比

Table 4 Precision@20 comparison of different learning algorithms

数据集	D_2	D_4	D_6	D_8
PathPredict	19	18	18	19
MPBP	17	18	19	19
MLSP	19	19	19	19

表 5 不同监督学习算法的 Precision@100 指标对比表

Table 5 Precision@100 comparison of different learning algorithms

数据集	D_2	D_4	D_6	D_8
PathPredict	76	76	70	64
MPBP	69	76	72	61
MLSP	77	76	73	62

结束语 异构网链路预测问题是学术界研究的热点问题之一,本文链路预测任务是目标元路径且为对称元路径,并在预测过程中假设目标元路径实例形成的贡献是单一的,它并没有存在与另一类型的节点之间有某种隐藏的特性,下一步计划对非对称元路径与不同类型节点的相关性建模进行研究。

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] HU W,LI J,CHENG J,et al. Security Monitoring of Heterogeneous Networks for Big Data Based on Distributed Association Algorithm[J]. Computer Communications,2020,152:206-214.

[3] KOVÁCS I A,LUCK K,SPIROHN K,et al. Network-based Prediction of Protein Interactions[J]. Nature Communications,2019,10(1):1-8.

[4] DAUD A,AHMAD M,MALIK M S I,et al. Using Machine Learning Techniques for Rising Star Prediction in Co-author Network[J]. Scientometrics,2015,102(2):1687-1711.

[5] SHI C,LI Y,ZHANG J,et al. A Survey of Heterogeneous Information Network Analysis[J]. IEEE Transactions on Knowledge and Data Engineering,2016,29(1):17-37.

[6] 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.

[7] JIANG L,YANG C C. User Recommendation in Healthcare Social Media by Assessing User imilarity in Heterogeneous Network[J]. Artificial Intelligence in Medicine,2017,81(9):63-77.

[8] ZHANG F,WANG M,XI J,et al. A Novel Heterogeneous Network-based Method for Drug Response Prediction in Cancer Cell Lines[J]. Scientific Reports,2018,8(1):355-367.

[9] LIANG W,LI X,HE X,et al. Supervised Ranking Framework for Relationship Prediction in Heterogeneous Information Networks[J]. Applied Intelligence,2018,48(5):1111-1127.

[10] LI J,ZHAO D,GE B F,et al. A Link Prediction Method for Heterogeneous Networks Based on BP Neural Network[J]. Physica A-Statistical Mechanics and Its Applications,2018,495(1):1-16.

[11] PENG Y C. Research on Link Prediction in Heterogeneous Information Networks[D]. Harbin:Harbin Institute of Technology,2020.

[12] LAI J,SHENG H L. Research on Link Prediction Performance of Complex Networks Based on Clustering Analysis[J]. Compu-

ting Technology and Automation, 2019(4): 144-150.
[13] WANG H, LE Z C, GONG X, et al. Link Prediction of Complex Networks is Analyzed from the Perspective of Informatics[J]. Journal of Chinese Computer Systems, 2020, 41(2): 316-326.
[14] BAI H, MA Y L, BI Y, et al. A Complicated Network Link Prediction Algorithm Based on Local Similarity of Nodes[J]. Computer Applications and Software, 2020, 37(5): 298-301.
[15] LIU S X, LI X, CHEN H C, et al. Link prediction method based on matching degree of resource transmission for complex network[J]. Journal on Communications, 2020, 41(6): 70-79.
[16] QI F P, WANG T, FU Z Q. Link prediction in complex networks based on mutual information[J]. Journal of University of Science and Technology of China, 2020, 50(1): 57-63.
[17] REVELLE M, DOMENICONI C, SWEENEY M, et al. Finding Community Topics and Membership in Graphs[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2015: 625-640.

HUANG Shou-meng, born in 1975, master, associate professor. His main research interests inclue information technology and information security.

(上接第 87 页)

[10] XIE J Y, GIRSHICK R, FARHADI A. Unsupervised deep embedding for clustering analysis[C]// ICML. 2016: 478-487.
[11] GUO X F, GAO L, LIU X W, et al. Improved deep embedded clustering with local structure preservation[C]// IJCAI. 2017: 1753-1759.
[12] PENG X, XIAO S J, FENG J S, et al. Deep subspace clustering with sparsity prior[C]// IJCAI. 2016: 101-115.
[13] JIANG Z X, ZHENG Y, TAN H C, et al. Variational deep embedding: An unsupervised and generative approach to clustering [C]// IJCAI. 2017: 4305-4324.
[14] NKIPF T, WELLING M. Semi-supervised classification with graph convolutional networks[C]// ICLR. 2017: 1-14.
[15] NKIPF T, WELLING M. Variational graph auto-encoders[J]. NIPS, 2016, 21(11): 1-3.
[16] WANG C, PAN S R, HU R Q, et al. Attributed Graph Clustering: A Deep Attentional Embedding Approach[C]// IJCAI. Marina del Rey CA USA: Association for the Advancement of Artificial Intelligence (AAAI), 2019: 3670-3676.
[17] LI X L, ZHANG H Y, ZHANG R. Embedding Graph Auto-Encoder with Joint Clustering via Adjacency Sharing[C]// WWW. 2020: 1-11.
[18] WANG C, PAN S R, LONG G D, et al. MGAE: Marginalized Graph Autoencoder for Graph Clustering[C]// ACM on Conference on Information and Knowledge Management. 2017: 889-898.
[19] ZHANG X T, LIU H, LI Q M, et al. Attributed Graph Clustering via Adaptive Graph Convolution[C]// IJCAI. 2019: 4327-4333.
[20] BO D Y, WANG X, SHI C, et al. Structural Deep Clustering Network[C]// WWW. 2020: 1-11.
[21] SUN J G, LIU J, ZHAO L Y. Research on clustering algorithm [J]. Journal of Software, 2008, 19(1): 48-61.
[22] JAIN A K, DUBES R C. Algorithms for clustering data [J]. Technometrics, 1988, 32(2): 227-229.
[23] REYNOLDS D A. Gaussian mixture models[C]// Encyclopedia of Biometrics. 2015: 1-23.
[24] JOHNSON S C. Hierarchical clustering schemes[J]. Psychometrika, 1967, 32(3): 241-254.
[25] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: Analysis and an algorithm[C]// Advances in Neural Information Processing Systems. 2002: 849-856.
[26] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensional ity of data with neural networks [J]. Science, 2006, 7(28): 504-507.
[27] HINTON G E. Learning multiple layers of representation[J]. Science, 2007, 7(4): 428-434.
[28] RAMACHANDRAN P, ZOPH B, LE Q V. Searching For Activation Functions[C]// ICLR. 2018: 1-13.
[29] CASANOVA A, ROMERO A, LIO P, et al. Graph Attention Networks[C]// IJCAI. 2018: 1-12.
[30] CHEPURI S P, LEUS G. Subsampling For Graph Power Spectrum Estimation[C]// IEEE SAM. 2016: 1250-1263.
[31] VAN DER MAATEN L, HINTON G. Visualizing data using t-sne[J]. Journal of Machine Learning Research, 2008, 9(Nov): 2579-2605.
[32] DENKER J, GARDNER W R, GRAF H, et al. Neural Network Recognizer for Hand-Written Zip Code Digits[C]// NIPS. 1988: 323-331.
[33] STISEN A, BLUNCK H, BHATTACHARYA S, et al. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition [C] // SenSys. ACM, 2015: 127-140.

KANG Yan, born in 1972, Ph.D, associate professor. Her main research interests include transfer learning, deep learning and integrated learning.

LI Hao, born in 1970, Ph.D, professor. His main research interests include distributed computing, grid and cloud computing.