DOI:10.7544/issn1000-1239.2017.20170172
54(8):1795-1803,2017

基于聚类和决策树的链路预测方法

杨妮亚[1] 彭 涛[1,2] 刘 露[1]
[1](吉林大学计算机科学与技术学院 长春 130012)
[2](符号计算与知识工程教育部重点实验室(吉林大学) 长春 130012)
(yangny15@mails.jlu.edu.cn)

Link Prediction Method Based on Clustering and Decision Tree

Yang Niya[1], Peng Tao[1,2], and Liu Lu[1]
[1](*College of Computer Science and Technology, Jilin University, Changchun* 130012)
[2](*Key Laboratory of Symbol Computation and Knowledge Engineering (Jilin University), Ministry of Education, Changchun* 130012)

Abstract Link prediction is one of the primal problems in data mining. Due to the network complexity and the data diversity, the problem of link prediction for different types of data in heterogeneous networks has become more and more complicated. Aiming at link prediction in bi-typed heterogeneous information network, this paper proposes a link prediction method based on clustering and decision tree, called CDTLinks. One kind of objects is considered as the features of the other kind of objects. Then, they are clustered separately. Three heuristic rules are proposed to construct decision trees for bi-typed heterogeneous networks. The branch of the tree with the highest information gain is selected. Finally, we can judge whether there is a link between two nodes through the clustering result and the decision tree model. In addition, we define the concept of potential link nodes and introduce the number of layers, which can reduce the running time and improve the accuracy. The proposed CDTlinks method is validated on DBLP and AMiner datasets. The experimental results show that the CDTlinks model can be used to conduct link prediction effectively in bi-typed heterogeneous networks.

Key words link prediction; clustering; decision tree; heterogeneous information network; heuristic rules

摘 要 链路预测是数据挖掘研究的主要问题之一.由于网络的复杂性、数据的多样性,根据网络结构及已有信息对异质网络中的不同类型的数据进行链路预测的问题也变得更加复杂.针对双类型异质信息网络,提出了一种基于聚类和决策树的链路预测方法CDTLinks.通过将网络中2种类型对象互为特征的方法得到对象的特征表示,并分别进行聚类.对于双类型异质网络提出了3种启发式规则来构建决策树,根据信息增益来选择树中不同分支.最后,根据聚簇分布结果以及决策树模型来判断任意2个不同

收稿日期:2017-03-18;**修回日期**:2017-05-11
基金项目:国家自然科学基金项目(60903098);吉林省发改委产业技术研究与开发专项(2015Y055);吉林省科技厅重点科技攻关项目(20150204040GX)
This work was supported by the National Natural Science Foundation of China (60903098), the Industry Technology Research and Development Projects of Development and Reform Commission of Jilin Province (2015Y055), and the Key Scientific Research Project of Department of Science of Jilin Province (20150204040GX).
通信作者:刘露(liulu12@mails.jlu.edu.cn)

类型节点之间是否存在链接.另外,定义了潜在链接节点并引入层数的概念,在降低算法运行时间的同时提高了准确率.在 DBLP 和 AMiner 数据集上验证了提出的 CDTlinks 方法,结果表明:在双类型异质网络中,CDTlinks 模型能够有效地进行链路预测.

关键词 链路预测;聚类;决策树;异质信息网络;启发式规则

中图法分类号 TP391

异质信息网络挖掘是数据挖掘的一类重要问题.预测网络中节点之间的链接关系具有重要的研究价值.链路预测旨在根据网络结构和已有信息发现并且还原网络中缺失的信息,或者预测未来节点之间可能存在的关系.节点间不同的链接关系对网络分析有重要的现实意义[1].链路预测也有广泛的应用,比如通过预测网络中节点之间的关系,在社交网络中进行朋友推荐[2]、识别隐藏和虚假的链接对网络进行重构[3]以及社团划分[4]等.

在异质信息网络中,不同类型的节点和链接包含着丰富的语义关系,使得链路预测变得更加复杂.例如文献信息网络包含会议/期刊、作者、论文等多种类型的节点以及多种链接关系.传统的异质信息网络链路预测是通过分析整个网络对象之间的链接关系找到虚假的链接,或者对未来的链接进行预测.这使得异质信息网络中链路预测变得十分复杂并且效率较低.

受到以上方法的启发,本文提出一种双类型异质信息网络链路预测的方法.该方法主要解决 4 个问题:1)如何在简化异质信息网络的同时保留主要语义信息;2)如何构建决策树模型;3)如何选取合适的属性作为决策树的决策节点;4)如何结合不同类型节点聚簇分布情况以及决策树进行链路预测.

为了解决上述问题,我们提出了一种双类型异质信息网络中基于聚类和决策树的链路预测方法.该方法提取异质网络中 2 种关键类型的对象,根据 2 种类型节点之间存在的链接关系构造邻接矩阵,采用 2 种类型对象互为特征的方法,对 2 种类型的对象分别进行聚类.这样保留主要语义关系的同时对节点间的链接进行分析.定义了 3 个启发式规则作为构建决策树的候选属性,选取信息增益最大的规则作为决策树的决策节点来构造决策树模型.最后,根据不同类型节点间聚类分布情况以及决策树模型对 2 种不同类型节点进行链路预测.以文献信息网络为例,我们通过分析网络中作者与期刊/会议之间的链接关系,对作者和期刊/会议这 2 种类型的节点进行聚类,根据 3 个启发式规则构建决策树来预测未来作者和期刊/会议之间可能出现的链接.

本文的主要贡献有 5 个方面:1)提出了一种双类型异质网络的链路预测方法,对 2 种类型节点之间的关系进行预测;2)采用 2 种类型对象互为特征的方法,对 2 种类型的对象分别进行聚类;3)定义了 3 种启发式规则作为构建决策树的候选属性;4)分析不同类型节点的聚类结果和节点之间的链接关系,得到决策树的属性,构造决策树模型;5)在不同数据集中进行实验.结果表明:本文提出的双类型异质信息网络链路预测方法可以有效地预测网络中节点之间的链接关系.

1 相关工作

链路预测是数据挖掘中一个关键的问题,在同质网络中,研究者们做了很多深入的研究. Bliss 等人[5]通过应用协方差矩阵自适应演化策略(CMA-ES)来优化在 16 个邻域和节点相似性索引的线性组合中使用的权重,提出了一种预测未来链路的方法;Scellato 等人[6]描述了一个监督学习框架,通过分析链路节点之间的关系来预测新的朋友和朋友的朋友之间的联系;Zhu[7]提出了一种非参数潜在特征关系模型的最大余量学习方法,该方法可以扩展到具有数百万实体和数千万个正链接的大规模实际网络中;Schifanella 等人[8]引入一个空模型,保留用户的活动,同时消除当地的相关性.实验结果证明由语义相似性构建的社会网络能够更准确地捕捉到实际的链路关系.

由于网络中充斥着许多不同类型的对象和链接关系,异质信息网络中的链路预测问题逐渐引起了研究者们的关注. Zeng 等人[9]构建基于元路径的异构网络的投资行为预测模型,该模型考虑与特定投资者的投资行为相关联的多个实体和关系类型,投资行为预测模型为元路径提供了一个有效的相似性度量函数;Huang 等人[10]使用元路径描述不同类型的节点和关系的不同语义,提出了一种基于元路径异构信息网络链路预测模型;Lakshmi 等人[11]为提

高效率,提出了异质信息网络中链路预测的并行方法,利用现有的多关系链接预测以及社区发现算法,在每个社区计算多关系链接预测分数;Aggarwal 等人[12]提出了一个有效的两级方案,为了将网络动态性和时间敏感度相结合,使用了宏观和微观决策;Lee 等人[13]对存在的文献网络进行修改并且突出其中重要的关系,将随机游走算法应用到修改后的网络链路预测问题.

研究者们在同质网络和多类型异质网络上做了很多的链路预测方面的研究,但关于双类型异质网络链路预测方法的研究还很少.在很多情况下,与多类型网络相比较来说,双类型网络在模型表示和计算的过程中相对简单.针对上述情况,本文针对双类型异质信息网络提出了链路预测方法.

2 问题定义

在本节中,我们先介绍链路预测方法需要用到的一些基本概念,再给出双类型异质信息网络中链路预测的形式化定义.

定义 1. 异质信息网络[14].给定一个有向图$G=(V,E)$,V 是节点集,E 是边集.存在一个节点类型映射函数 $\tau:V\rightarrow A$,一个边类型映射函数 $\varphi:E\rightarrow R$,其中,每个节点 $v\in V$ 都属于一个特定的对象类 $\tau(v)\in A$,每条边 $e\in E$ 都属于一个特定的关系类 $\varphi(e)\in E$.若网络中节点类型数 $|A|>1$ 或边类型数量 $|R|>1$,则该网络被称为异质信息网络,反之,为同质信息网络.

定义 2. 双类型异质信息网络[14].给定一个异质信息网络 $G=(X\cup Y,\boldsymbol{W})$.$X$ 和 Y 分别代表 2 种不同类型的对象集合,$\boldsymbol{W}$ 代表对象之间的链接关系矩阵,其中,$W_{XY}(i,j)=p_{ij}\ (i=1,2,\cdots,m;j=1,2,\cdots,n)$.对于任意 $x_i\in X,y_j\in Y$,p_{ij} 为节点 x_i 与节点 y_j 之间链接的数量,因此,存在 $\boldsymbol{x}_i=(p_{i1},p_{i2},\cdots,p_{in})$,$\boldsymbol{y}_j=(p_{j1},p_{j2},\cdots,p_{jm})$.

定义 3. 潜在链接节点.给定一个双类型异质网络 $G=(X\cup Y,\boldsymbol{W})$,且存在节点 $y_j,y_k\in Y$,若存在节点 $x_i\in X$,使得 x_i 与 y_j,x_i 与 y_k 之间均存在链接关系.则 y_j 与 y_k 互为直接链接节点,表示为 $y_j\in LinkNode(y_k)$.若存在 y_i 与 y_j 互为直接链接节点,y_j 与 y_k 互为直接链接节点,且 y_i 与 y_k 不存在直接链接关系;那么,y_i 与 y_k 互为潜在链接节点(latent link node),y_i 与 y_k 的关系表示为 $y_i\in LatentLinkNode(y_k)$.

此时,节点 y_i 与节点 y_k 之间的链接关系可表示为 $y_i-y_j-y_k$,y_i 到 y_k 的链接个数为 2.y_i 与 y_k 也互为 2 层潜在链接节点.随着链接关系中链接数量的增加,节点间的相近关系也随之变化,因此,我们引入 n 层潜在链接节点的概念来分析异质网络中存在链路的可能性,即如果 y_i 与 y_k 之间存在 n 个链接,那么它们互为 n 层潜在链接节点.下面我们给出双类型异质网络中链路预测问题的形式化定义.

问题 1. 双类型异质网络中的链路预测.给定一个双类型异质网络 $G=(X\cup Y,\boldsymbol{W})$,对于网络中任意 2 个节点 $x_i\in X,y_j\in Y$,根据网络中节点的聚类结果以及决策树的分析结果,预测 x_i 与 y_j 之间是否存在链接.

3 双类型异质网络聚类过程

在用决策树进行链路预测之前,根据双类型网络中的信息对 X 类型节点以及 Y 类型节点进行聚类.双类型异质信息网络是一种特殊的异质信息网络,它仅包含 2 种类型的对象,在对 X 类型节点进行聚类时,以 Y 类型节点作为对应的 X 类型节点的特征.同样,对 Y 类型节点进行聚类时,以 X 类型节点作为对应的 Y 类型节点的特征.采用 2 种对象互为特征的方法来得到每个节点的特征表示.以文献信息网络为例,2 种节点类型分别为会议和作者,将会议作为目标对象.因此,一个会议可以表示为向量 $\boldsymbol{x}_i=(p_{i1},p_{i2},\cdots,p_{in})$,其中,$p_{ij}$ 表示在会议 i 上作者 j 发表的论文数.如图 1 所示,会议 x_1 表示为 $\boldsymbol{x}_1=(2,2,0,0)$,会议 x_2 表示为 $\boldsymbol{x}_2=(0,1,2,0)$,会议 x_3 表示为 $\boldsymbol{x}_3=(0,0,3,1)$.将每个作者在这个会议上发表论文的数作为特征值,利用余弦相似度量计算 2 个会议的相似程度.则 x_1 和 x_2 的相似程度为

$$sim(\boldsymbol{x}_1,\boldsymbol{x}_2)=\frac{2\times0+2\times1+0\times2+0\times0}{\sqrt{2^2+0^2}+\sqrt{2^2+1^2}+\sqrt{2^2+0^2}+\sqrt{0^2+0^2}}=0.321.$$

得到 X 类型对象和 Y 类型对象的特征表示后,分别进行聚类.以 Y 类型对象为特征,Y 和 X 两种类型对象之间的链接权值作为特征值对 X 类型对象进行聚类.通过计算聚类中心与每个对象的距离反复对聚类进行调整,使得聚类内部的误差平方和[15]最小,从而得到 X 类型对象的 K_1 个聚类.同样,以 X 类型对象为特征,X 和 Y 两种类型对象之间的链接权值作为特征值对 Y 类型对象进行聚类.

通过计算聚类中心与每个对象的距离反复对聚类进行调整,得到 Y 类型对象的 K_2 个聚类.

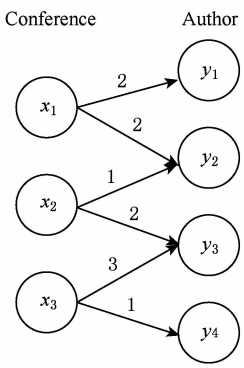

Fig. 1 An example of relationship between the conference and the author

图 1 会议和作者的关系实例

4 决策树的构建过程

在第 3 节中,我们得到了 2 种类型对象的聚类结果.本节我们介绍如何构建决策树来预测网络中节点之间的链接关系.

给定一个双类型异质信息网络 $G=(X\cup Y, \boldsymbol{W})$,对于网络中任意 2 个节点 $x_i\in X, y_j\in Y$,根据以下 3 条规则来判断 x_i 与 y_j 之间是否存在链接.若 x_i 与 y_j 存在链接,则函数 $linkpredict(x_i,y_j)=1$;反之,$linkpredict(x_i,y_j)=0$.

规则 1. x_i 所在聚簇为 C_k,若节点 y_j 与 x_i 同聚簇的节点之间存在链接,那么,x_i 与 y_j 之间可能存在链接.该规则可表示为

$$\exists x,x_i\in X,\ \exists y_j\in Y,x,x_i\in C_k,$$
$$(x,y_j)\in E\Rightarrow linkpredict(x_i,y_j)=1. \quad (1)$$

规则 2. 若节点 x_i 的直接链接节点或潜在链接节点与 y_j 存在链接,那么,x_i 与 y_j 之间可能存在链接.该规则可表示为

$$\exists x,x_i\in X,\ \exists y_j\in Y,$$
$$x\in LinkNode(x_i)\cup LatentLinkNode(x_i),$$
$$(x,y_j)\in E\Rightarrow linkpredict(x_i,y_j)=1. \quad (2)$$

规则 3. y_j 所在聚簇为 C_l,若节点 x_i 与 y_j 同聚簇的节点之间存在链接,那么,x_i 与 y_j 之间可能存在链接.该规则可表示为

$$\exists y,y_j\in Y,\ \exists x_i\in X,y,y_j\in C_l,$$
$$(x_i,y)\in E\Rightarrow linkpredict(x_i,y_j)=1. \quad (3)$$

我们应用以上 3 条规则来预测双类型异质网络中的链接,以文献信息网络为例,我们将会议名作为 X 类型节点,将作者名作为 Y 类型节点,并将它们之间的链接关系存储在邻接矩阵 $\boldsymbol{W}$ 中.如果一个作者 y_j 与会议 x_i 同领域的会议间存在链接,那么,该作者 y_j 与会议 x_i 可能存在链接(规则 1).如果一个作者与 y_j 与会议 x_i 的直接链接节点或潜在链接节点之间存在链接,那么,该作者 y_j 与会议 x_i 可能存在链接(规则 2).如果一个会议 x_i 与作者 y_j 同领域的作者间存在链接,那么,该会议 x_i 与作者 y_j 可能存在链接(规则 3).

本文把 3 条规则作为决策树的候选属性集 Pro.在构造决策树的过程中,将信息增益[16]作为选择候选属性的度量指标来衡量 3 条规则哪一条可以更好地进行链路预测.设 D 是数据集,根据数据的类别对 D 进行划分,其中类别由存在链接和不存在链接 2 部分组成.存在链接的类别数据记为 D_1,不存在链接的类别数据记为 D_2,则 D 的熵表示为

$$info(D)=-\sum_{i=1}^{2}p_i\,\mathrm{lb}(p_i), \quad (4)$$

$$p_i=\frac{D_i}{D},\ i=1,2, \quad (5)$$

其中,p_i 表示第 i 个类别在整个训练元组中出现的概率.假设将训练元组 D 按属性 $attr$ 进行划分,其中 $attr\in Pro$,则 $attr$ 对 D 划分的期望信息[16]为

$$info_{attr}(D)=\sum_{j=1}^{2}\frac{D_j}{D}info(D_j). \quad (6)$$

则信息增益[16]为两者之间的差值:

$$gain(attr)=info(D)-info_{attr}(D). \quad (7)$$

每次分裂时,计算候选属性 $attr$ 中每个属性的增益值,选择增益值最大的属性作为决策树的分支,构造决策树.构造决策树并使用决策树进行链路预测的示意图,如图 2 所示.在构造决策树的过程中,计算每个属性的信息增益,选择信息增益值较大的属性作为决策树的分支.在链路预测的过程中,判断对象 x 与对象 y 之间是否存在链接,若规则 1 对应的信息增益最大,则用图 2 左侧分支进行链路预测.算法 1 描述了决策树的构造过程.

算法 1. 决策树构造算法

输入:双类型网络 $G=\{(X\cup Y),\boldsymbol{W}\}$、一种类型对象 $X=\{x_1,x_2,\cdots,x_m\}$、另一种类型对象 $Y=\{y_1,y_2,\cdots,y_n\}$、X 与 Y 之间的链接关系矩阵 $\boldsymbol{W}$,其中,$W_{XY}(i,j)=p_{ij}(i=1,2,\cdots,m;j=1,2,\cdots,n)$,加入 $\boldsymbol{WN}$(由 G 中节点构造的不存在的链接矩阵,占总体链接的 10%),进行链路预测的 X 类型对象 x 和 Y 类型对象 y;

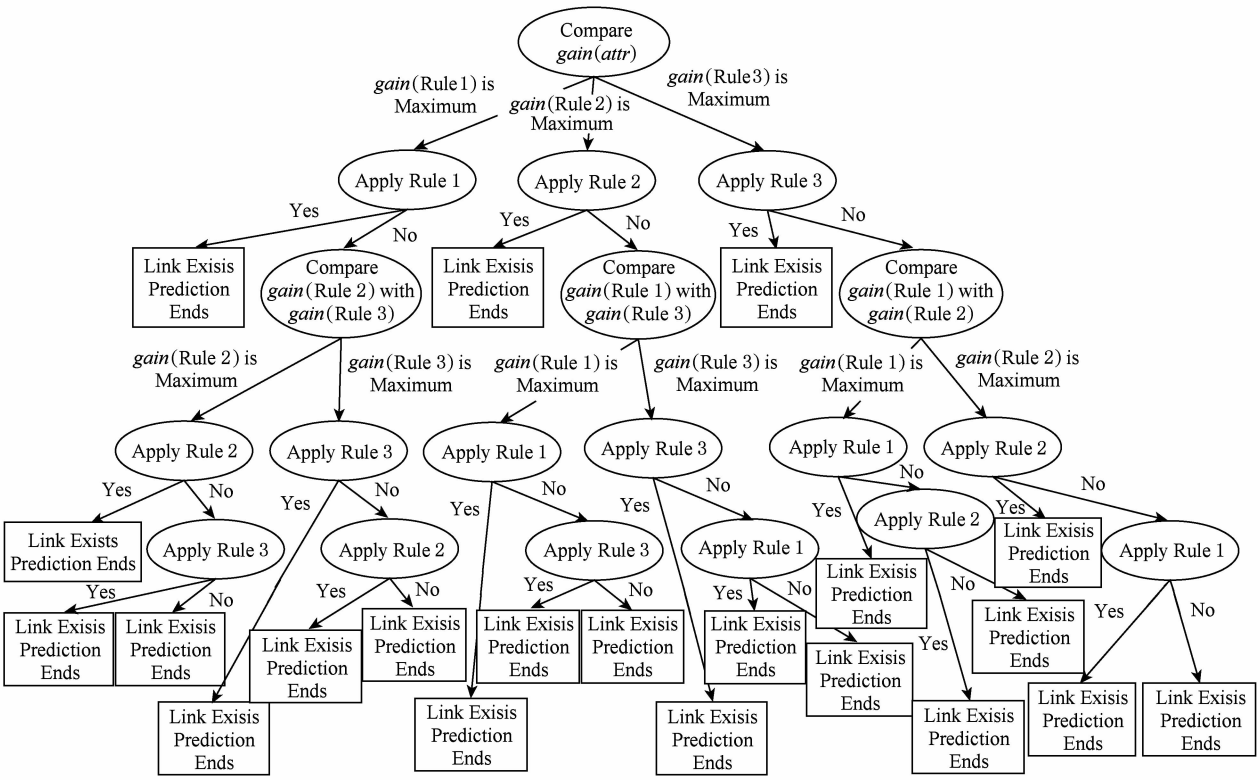

Fig. 2 Illustration of constructing decision tree

图 2 构造决策树示意图

输出:决策树.

① for each $p\in(\boldsymbol{W}+\boldsymbol{WN})$ ($i=1,2,\cdots,m$; $j=1,2,\cdots,n$)

② 链接 p 两边节点记作 m 和 n,m 和 n 按照 3 个启发式规则得到每一个训练数据在各个属性的对应值(值为 Yes/No);

③ end for

④ for each $attr\in Pro$(Pro 为决策树属性的集合)

⑤ 计算 $attr$ 的信息增益;

⑥ if $info_{attr}(D)\geqslant info_{attr_1}(D)(attr_1\in Pro)$

⑦ $attr$ 作为决策树的分支;

⑧ $Pro=Pro-attr$.

⑨ end if

⑩ end for

5 基于聚类和决策树的链路预测方法

通过网络结构以及网络中已有的信息来预测节点未来的链接关系是链路预测的主要任务.网络中存在着多种类型的节点以及多种类型的链接关系,根据不同节点间的关联关系和语义关系来预测链接能够帮助研究者更好地分析网络中的数据.因此,本节中提出了一种双类型异质网络中基于聚类和决策树的链路预测方法.

首先,我们提取异质网络中 2 种不同类型的对象以及对象间的链接关系,通过对象间互为特征的方法得到双类型对象的特征表示,并将 2 种类型对象进行聚类.得到不同类型对象的聚簇分布后,定义了 3 个规则作为构建决策树的候选属性,选取信息增益较大的规则来构建决策树.最后,将需要判断的节点输入到决策树中,根据决策树以及聚簇分布情况来判断任意 2 个不同类型节点之间是否存在链接.

下面我们给出一个双类型文献信息网络进行链路预测的实例,如图 3 所示.2 种类型节点分为作者和会议名,图 3 中已知对象间已有的链接以及对象的聚类分布情况,目标为预测下一时刻 Andy 与 SIGMOD 间是否存在链接.Andy 与 Bob 与为合作作者,Bob 与 Cindy 互为合作作者,那么,Bob 与 Andy 互为二层潜在链接作者.Cindy 和 David 在同一聚类 C_1 中,那么,David 与 Andy 的合作作者 Cindy 在同一聚类中.根据规则 2 和规则 1,如果 David 与 SIGMOD 之间存在链接,那么 Andy 与 SIGMOD 之间很可能存在链接.会议 SIGMOD 与 VLDB 在同一聚类中,根据规则 1,如果 Andy 与

VLDB之间存在链接,那么 Andy 与 SIGMOD 之间很可能存在链接.在图 3 中,David 与 VLDB 之间存在链接.所以,根据规则 3,David 与 SIGMOD 之间很可能存在链接.根据规则 1,Cindy 与 SIGMOD 之间很可能存在链接.根据规则 2,Andy 与 SIGMOD 之间很可能存在链接.

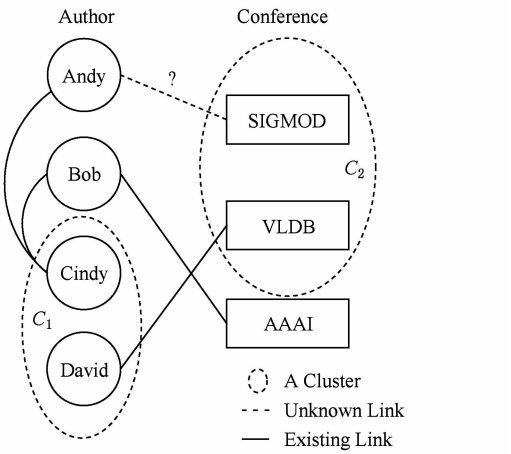

Fig. 3 An example of link prediction for bibliographic information network

图 3 文献信息网络链路预测实例

6 实验与结果

在本节中,我们使用本文提出的聚类和决策树方法构建一个面向双类型异质网络的链路预测模型,并在 2 个数据集上进行测试了我们的方法.

6.1 度量标准

为了评估链路预测模型 CDTLinks 的效果,我们使用准确率 *Accuracy*[17]、精确率 *Precision*[18]、召回率 *Recall*[19] 和 *F-Measure*[20] 4 种度量标准来检测链路预测模型的性能.精确率和召回率分别反映了模型找对和找全链接的能力.我们将链路预测的结果分为有链接和无链接 2 种情况.针对 2 种情况,分别给出相应的精确率和召回率.对于有链接的数据,精确率是被正确识别的链接数占被识别为有链接的数据的比例.召回率,也称查全率,是被正确检测出有链接的数据占实际存在链接数据的比例.有链接情况精确度和召回率的计算为

$$Precision_{\mathrm{E}}=\frac{|C_{\mathrm{E}}\cap T_{\mathrm{E}}|}{|C_{\mathrm{E}}|}\times 100\%, \tag{8}$$

$$Recall_{\mathrm{E}}=\frac{|C_{\mathrm{E}}\cap T_{\mathrm{E}}|}{|T_{\mathrm{E}}|}\times 100\%, \tag{9}$$

其中,C_{E} 表示被链路模型判断为有链接的数据集合,T_{E} 表示测试集中存在的有链接数据的集合.对于无链接的数据,精确率指被正确识别为无链接的对象数量占被识别为无链接的数据的比例.召回率指被正确识别为无链接的对象数量占实际为无链接数据的比例.无链接情况精确度和召回率的计算为

$$Precision_{\mathrm{N}}=\frac{|C_{\mathrm{N}}\cap T_{\mathrm{N}}|}{|C_{\mathrm{N}}|}\times 100\%, \tag{10}$$

$$Recall_{\mathrm{N}}=\frac{|C_{\mathrm{N}}\cap T_{\mathrm{N}}|}{|T_{\mathrm{N}}|}\times 100\%, \tag{11}$$

同样,C_{N} 表示被链路模型判断为无链接的数据集合,T_{N} 表示测试集中存在的无链接数据的集合.*Precision* 和 *Recall* 不成正相关,我们采用这 2 个指标的调和平均值 *F-Measure* 来作为评估标准,*F-Measure* 的计算为

$$F\text{-}Measure=\frac{(\beta^2+1)Precision\times Recall}{\beta^2\times Precision+Recall}. \tag{12}$$

β 是一个反应精确度和召回率相对重要程度的权值,若 $\beta>1$,则召回率的重要性大于准确率,反之亦然.在本文中将设 $\beta=1$.

Accuracy 作为度量标准来衡量所有数据是否和真实类别一致,即衡量模型作出正确决定的能力.*Accuracy* 的公式定义为

$$Accuracy=\frac{TP+TN}{|T_{\mathrm{E}}|+|T_{\mathrm{N}}|}\times 100\%, \tag{13}$$

其中,TP 表示有链接数据集中被正确识别为有链接数据的数量,TN 表示无链接数据集中被正确识别为无链接数据的数量,$|T_{\mathrm{E}}|$ 和 $|T_{\mathrm{N}}|$ 之和为所有数据总数.

6.2 数据集

在本文中,我们使用 2 个真实的信息网络:DBLP① 网络和 AMiner[21] 网络.

1) DBLP 数据集是异质网络中常用的实验数据,其中的数据类型包括会议/期刊、作者、发表时间等等.提取期刊、作者以及二者之间的链接信息作为实验数据.将出现在实验中的期刊人工标记它所属的领域.在实验中,读取其中的 20 000 条数据,出现 22 个期刊、31 181 个作者、892 983 条期刊与作者的链接关系,人工标记期刊所属的领域.

2) AMiner 数据集是社会网络挖掘常用的数据集,其中包括作者、会议、主题等信息.本文使用 aminer-topic-data-pubs.xml 文件中的数据.对 AMiner 数据集进行了预处理,提取出 6 类会议信息.以提取

① http://dblp.org/

出的会议,作者以及会议与作者之间的链接信息作为实验数据,其中包括 33 个期刊、21 381 个作者、836 524 条期刊与作者的链接关系.

6.3 参数分析及结果

本节中,我们通过实验来验证 CDTLinks 链路模型.首先,分析潜在链接节点的层数 μ 对链路模型的影响.对 DBLP 数据集和 AMiner 数据集,判断作者和期刊之间是否存在链接.其中,待预测作者的合作作者信息对链路预测具有很重要的意义.如果待预测作者的直接链接作者与待预测期刊之间有链接关系,那么待预测的作者和期刊之间存在链接关系的可能性很高.如果待预测作者的二层潜在链接合作作者与待预测期刊之间有链接关系,那么待预测的作者和期刊之间存在链接关系的可能性相对较低.同时,搜索二层潜在作者信息所花费的时间相比于搜索直接链接作者信息所花费的时间要多.理论上,异质信息网络中的潜在链接节点层数可以无限大.但是,随着搜索潜在节点信息的层数的增加,链路预测结果的准确率降低,花费的时间增加.

图 4 和图 5 分别给出了随着潜在链接节点层数 μ 的增加,在 DBLP 和 AMiner 数据集上 *Accuracy*, *Precision*, *Recall* 和 *F-Measure* 的变化曲线.从图 4 和图 5 中可以看出,在 2 个数据集中,*Accuracy*, *Precision*, *Recall* 和 *F-Measure* 在 $\mu=3$ 时取得峰值.当 $\mu<3$ 时,随着 μ 的增加,链路预测的 *Accuracy*, *Precision*, *Recall* 和 *F-Measure* 增加.当 $\mu>3$ 时,*Accuracy*, *Precision*, *Recall* 和 *F-Measure* 不再增加.这是因为随着潜在链接搜索层数的增加,从网络链接中获得的信息越多,CDTLinks 表现出的性能也随着增加.当增加到一个峰值时,随着搜索层数的增加,导致错误或冗余的信息过多,CDTLinks 表现出的性能反而下降.

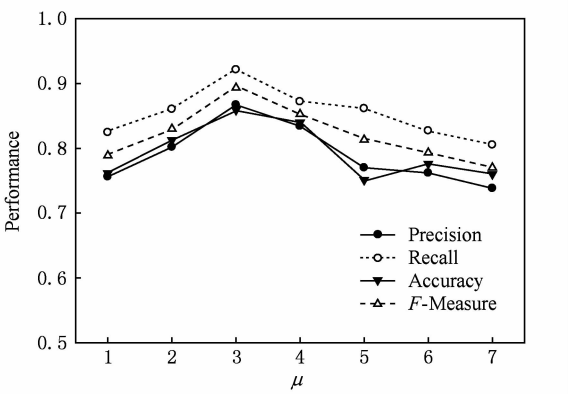

Fig. 4 The performance of CDLinks in DBLP with different μ

图 4 CDLinks 在 DBLP 数据集中 μ 变化时的性能

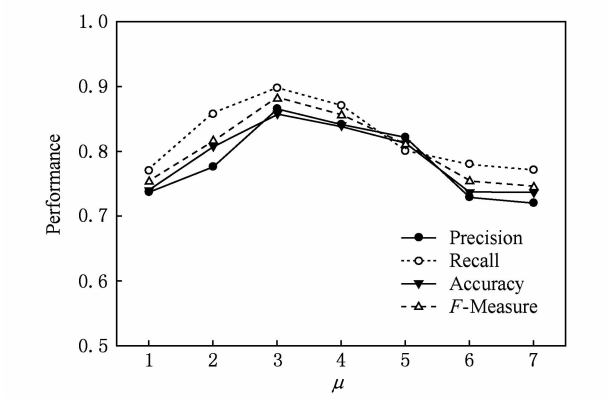

Fig. 5 The performance of CDLinks in AMiner with different μ

图 5 CDLinks 在 AMiner 数据集中 μ 变化时的性能

图 6 给出了随着潜在链接节点层数 μ 的增加,CDTLinks 运行时间的变化.从图 6 中可以看出,当 $\mu>3$ 时,算法的迭代时间显著增加.在实验中,网络中节点的数量为 n,链接的数量为 m,聚类个数为 k,决策树特征的数量为 p.因此,聚类部分计算的时间复杂度为 $O(nk)$;链接分析部分计算的时间复杂度为 $O(n^{\mu})$;决策树计算部分的时间复杂度为 $O(pm)$.随着潜在链接节点层数的增加,算法的运行时间增加.综合考虑算法的性能和时间上的开销,我们将潜在链接节点层数 μ 设为 3.

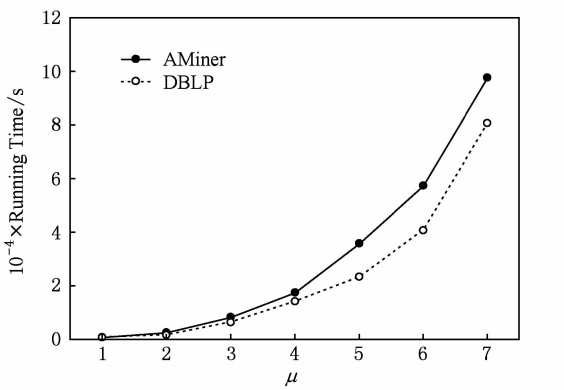

Fig. 6 Running time of μ on DBLP and AMiner datasets

图 6 不同的 μ 在 DBLP 和 AMiner 数据集中的运行时间

我们将提出的 CDTLinks 链路预测方法与 2 个基线算法进行比较.其中,CDTLinks 方法中的潜在链接节点层数 $\mu=3$.如图 7 和图 8 所示,通过 *Accuracy*, *Precision*, *Recall* 和 *F-Measure* 值,在 2 个数据集 DBLP 和 AMiner 上对我们提出的 CDTLinks 链路预测方法进行测试.计算 CDTLinks 方法的 *Accuracy*,

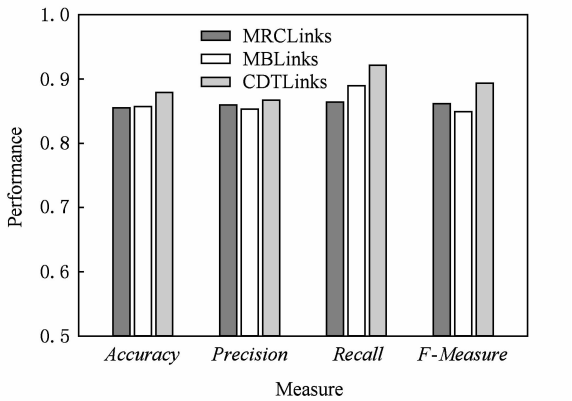

Fig. 7 Performance comparision of three link prediction methods on DBLP dataset

图 7 3 种链路预测方法在 DBLP 数据集上性能的比较

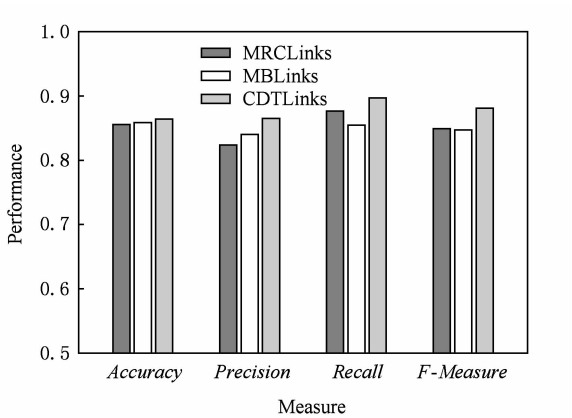

Fig. 8 Performance comparision of three link prediction methods on AMiner dataset

图 8 3 种链路预测方法在 AMiner 数据集中的性能比较

Precision, *Recall* 和 *F-Measure* 值, 我们将 *Precision* 设置为 $(Precision_E + Precision_N)/2$, 将 *Recall* 设置为 $(Recall_E + Recall_N)/2$. 将我们提出的方法 CDTLinks 与 MRCLinks[11], MBLinks[13] 相比. 在 DBLP 和 AMiner 两个数据集进行链路预测的实验, 实验结果表明, 我们给出的 CDTLinks 方法优于 MRCLinks[11], MBLinks[13] 方法. 该方法提取异质网络中 2 种关键类型的对象, 根据 2 种类型节点之间存在的链接关系构造邻接矩阵. 采用 2 种类型对象互为特征的方法, 对 2 种类型的对象分别进行聚类. 这样保留主要语义关系的同时对节点间的链接进行了分析, 较为准确地对 2 种类型对象进行领域划分. 同时, 定义了 3 个启发式规则作为构建决策树的候选属性, 选取信息增益最大的规则作为决策树的决策节点来构造决策树模型, 从而保证了将决策树模型运用到链路预测中的效率. 根据节点间的聚类分布情况以及网络中的链接信息使用决策树模型对 2 种不同类型节点进行链路预测, 经过实验可以证明, 我们提出的 CDTLinks 链路预测算法在链路预测的过程中得到了很好的效果.

7 结 论

本文针对双类型异质信息网络提出了一种基于聚类和决策树的链路预测方法. 该方法结合了聚类和决策树, 在双类型的异质网络中具有良好的预测效果. 2 种对象互为特征的表示方法充分利用了网络中节点之间的信息, 得到 2 种对象的聚类分布情况. 3 种启发式规则根据网络中的节点以及聚类分布更好地分析了网络中节点和链接的关系. 利用 3 种启发式规则计算信息增益, 将信息增益最大的规则作为决策节点能够更快地构建决策树, 进而快速有效地进行链路预测. 通过实验证明了本文提出的 CDTLinks 算法的正确性和有效性.

参 考 文 献

[1] Lü Linyuan. Link prediction on complex networks [J]. Journal of University of Electronic Science & Technology of China, 2010, 39(5): 651-661 (in Chinese)
(吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39(5): 651-661)

[2] Xu Bin, Chin A, Wang Hao, et al. Using physical context in a mobile social networking application for improving friend recommendations [C] //Proc of the 4th Int Conf on Internet of Things. Piscataway, NJ: IEEE, 2011: 602-609

[3] Zhang Peng, Zeng An, Fan Ying. Identifying missing and spurious connections via the bi-directional diffusion on bipartite networks [J]. Physics Letters A, 2014, 378(32-33): 2350-2354

[4] Yang Liu, Cao Jinxin, Liu Bo, et al. Community division algorithm based on feedback of unbiased Q value [J]. Journal of Southeast University, 2011, 41(1): 31-36

[5] Bliss C A, Frank M R, Danforth C M, et al. An evolutionary algorithm approach to link prediction in dynamic social networks [J]. Journal of Computational Science, 2014, 5(5): 750-764

[6] Scellato S, Noulas A, Mascolo C. Exploiting place features in link prediction on location-based social networks [C] // Proc of the 17th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2011: 1046-1054

[7] Zhu Jun. Max-margin nonparametric latent feature models for link prediction [C] //Proc of the 29th Int Conf on Machine Learning. New York: ACM, 2012: 719-726

[8] Schifanella R, Barrat A, Cattuto C, et al. Folks in folksonomies: Social link prediction from shared metadata [C] //Proc of the 3rd ACM Int Conf on Web Search & Data Mining. New York: ACM, 2010: 271-280

[9] Zeng Xiangxiang, Li You, Leung S C H, et al. Investment behavior prediction in heterogeneous information network [J]. Neurocomputing, 2016, 217: 125-132

[10] Huang Liwei, Li Deyi, Ma Yutao, et al. A meta path-based link prediction model for heterogeneous information networks [J]. Chinese Journal of Computers, 2014, 37(4): 848-858 (in Chinese)
(黄立威, 李德毅, 马于涛, 等. 一种基于元路径的异质信息网络链路预测模型[J]. 计算机学报, 2014, 37(4): 848-858)

[11] Lakshmi T J, Bhavani S D. Heterogeneous link prediction based on multi relational community information [C] //Proc of the 6th Int Conf on Communication Systems and Networks. Piscataway, NJ: IEEE, 2014: 1-4

[12] Aggarwal C C, Xie Yan, Yu P S. A framework for dynamic link prediction in heterogeneous networks [J]. Statistical Analysis & Data Mining, 2014, 7(1): 14-33

[13] Lee J B, Adorna H. Link prediction in a modified heterogeneous bibliographic network [C] //Proc of Int Conf on Advances in Social Networks Analysis and Mining. Los Alamitos, CA: IEEE Computer Society, 2012: 442-449

[14] Sun Yizhou, Han Jiawei, Zhao Peixiang, et al. RankClus: Integrating clustering with ranking for heterogeneous information network analysis [C] //Proc of the 12th Int Conf on Extending Database Technology: Advances in Database Technology. New York: ACM, 2009: 565-576

[15] Han J W, Kamber M, Pei J. Data Mining Concepts and Techniques Third Edition [M]. San Francisco, CA: Morgan Kaufmann, 2012: 102-120

[16] Sivatha S S S, Geetha S, Kannan A. Decision tree based light weight intrusion detection using a wrapper approach [J]. Expert Systems with Applications, 2012, 39(1): 129-141

[17] Khoshelham K. Accuracy analysis of kinect depth data [J]. ISPRS-Int Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2012, 3812(5): 133-138

[18] Zhang Ke, Hutter M, Jin Huidong. A new local distance-based outlier detection approach for scattered real-world data [C] //Proc of the 13th Conf on Knowledge Discovery and Data Mining. Berlin: Springer, 2009: 813-822

[19] Tzeng J Y, Byerley W, Devlin B, et al. Outlier detection and false discovery rates for whole-genome DNA matching [J]. Journal of the American Statistical Association, 2003, 98 (461): 236-246

[20] Croft W B, Metzler D, Strohman T. Search Engines: Information Retrieval in Practice [M]. Reading, MA: Addison-Wesley, 2010: 23-37

[21] Tang Jie, Zhang Jing, Yao Limin, et al. Arnetminer: Extraction and mining of academic social networks [C] //Proc of the 14th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2008: 990-998

Yang Niya, born in 1992. Master. Her main research interests include Web mining and machine learning.

Peng Tao, born in 1977. PhD, professor. Member of CCF. His main research interests include Web mining, information retrieval and machine learning.

Liu Lu, born in 1989. PhD. Her main research interests include Web mining, information retrieval and machine learning.