计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20170961
56(3):623-634,2019

基于层次信息粒表示的属性图链接预测模型

罗 晟[1,2] 苗夺谦[1,2] 张志飞[1,3] 张远健[1,2] 胡声丹[1,2]
[1](同济大学计算机科学与技术系 上海 201804)
[2](嵌入式系统与服务计算教育部重点实验室(同济大学) 上海 201804)
[3](计算机软件新技术国家重点实验室(南京大学) 南京 210023)
(tjluosheng@gmail.com)

A Link Prediction Model Based on Hierarchical Information Granular Representation for Attributed Graphs

Luo Sheng[1,2], Miao Duoqian[1,2], Zhang Zhifei[1,3], Zhang Yuanjian[1,2], and Hu Shengdan[1,2]
[1](*Department of Computer Science and Technology, Tongji University, Shanghai* 201804)
[2](*Key Laboratory of Embedded System and Service Computing (Tongji University), Ministry of Education, Shanghai* 201804)
[3](*State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing* 210023)

Abstract With the accumulation of the network graph data coupled with node attributes, the relations between node attributes and node linkages become more and more complex, which brings a lot of challenges to the task of the link prediction in complex network. The main reason is the inconsistency existing in the different source data, that is, the relations between the latent linkages which are implied by the node attributes and the observed linkages from network topological structure, respectively. This phenomenon directly affects the correctness and accuracy of link predictions. In order to effectively deal with multi-source data inconsistency and fuse the heterogeneous data, with the idea of granular computing and data multi-layer granular representation, we model the original data at different levels of granular representation. According to the data granular representation, we ultimately eliminate data inherent inconsistencies by finding the optimal granular structure. In this paper, we firstly define the data granular representation and the relation between different level granular; Then, we construct a log-likelihood model of the data, and place a lot of constraints decided by the granular relations to regularize the model; At last, we use the trained model to perform the link probability between nodes. Experiments show that, multi-source data can ultimately reduce the inconsistency by granular representation, and the statistic model regulated by these granular relations outperforms the state-of-the-art methods, and effectively improves the accuracy of the link prediction in the attributed graph.

Key words granular representation learning; granular computing; attributed graph; link prediction; data fusion

收稿日期:2017-12-14;**修回日期**:2018-08-14

基金项目:国家自然科学基金项目(61673301,61502259);南京大学计算机软件新技术国家重点实验室开放课题(KFKT2017B22)
This work was supported by the National Natural Science Foundation of China (61673301, 61573255) and the Open Research Funds of State Key Laboratory for Novel Software Technology (Nanjing University) (KFKT2017B22).

通信作者:苗夺谦(dqmiao@tongji.edu.cn)

摘 要 随着具有结点属性信息的网络图数据的增加,结点属性及结点链接关系越来越复杂,这对复杂网络的链接预测任务带来了一系列的挑战.这些不同来源的原始数据之间存在着不一致性,即结点的属性诱导的潜在链接关系与网络拓扑结构观测到的链接边之间存在着不一致的情况,这一现象将直接影响结点对之间的链接预测准确性与精确性.为了有效处理多源数据的不一致性,融合异构数据的差异,借助粒计算思想,通过对原始数据的多粒度表示,将原始数据在不同层次的粒度进行信息表示建模.最终依据这些数据的粒度表示,寻找最优的粒层结构,并最大化地消除数据内在的不一致性.首先,定义了数据的粒度不同层次表示及粒层关系;其次,对所观测到的链接数据,构建对数似然统计模型,并综合不同粒度层数据特点对模型进行修正;最后,使用多源数据训练统计模型,将学习好的模型用于预测结点对之间的链接概率.实验表明:与现有链接预测模型相比,多源数据经过粒度表示极大地平衡了多源数据的不一致性,有效提升了链接预测任务的准确性.

关键词 粒度表示学习;粒计算;属性图;链接预测;数据融合

中图法分类号 TP18

随着信息技术的快速发展,越来越多的网络应用将人们紧密地联系在一起,形成了人与人之间的一个链接网络.现有大量的复杂网络分析相关的研究工作发表在统计物理、统计学、计算机科学和应用数学等领域[1-2].复杂网络分析早已成为国内外学者研究的热点问题.链接预测正是复杂网络分析的一个基础性工作,具有重要的地位.早期的链接预测[3]分析方法的出发点是建立在数据的网络拓扑图之上,这类方法较为直观,易于解释.除了网络拓扑结构数据之外,结点还具有属性信息,这类复杂网络一般称之为属性图[4].一个常见的链接网络如图1所示:

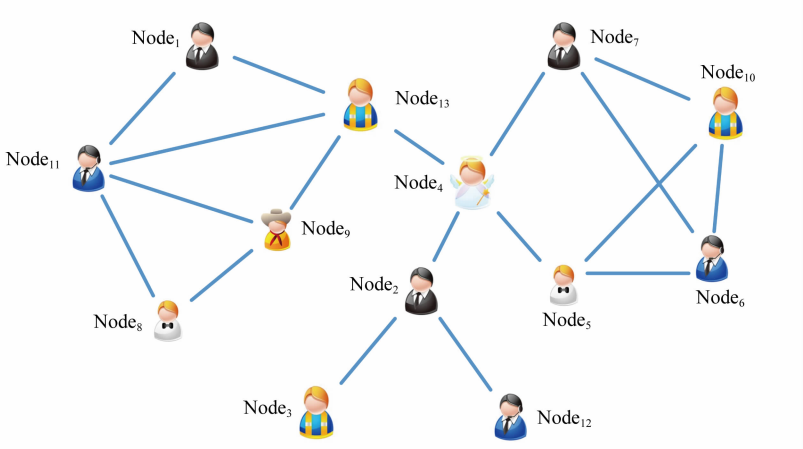

Fig. 1 A toy example of the link network

图1 一个示例链接网络

然而,伴随着计算机的存储能力和计算能力的快速增长,以及进入Web 2.0时代之后,人们参与数据发布的积极性高涨,以至于现在的复杂网络数据的规模越来越庞大.同时,网络结点属性的丰富程度也越来越高.复杂网络数据的来源及质量得到了巨大的提升.数据的快速增长导致人们获取的数据亦呈现大数据特点,即体量巨大(volume)、类型多样(variety)、产生速度快(velocity)、易变性(variability)以及真实性(veracity)① 等特征.这些数据特性对现有智能学习系统带来了巨大的挑战,特别是数据的易变性,也就是数据的不一致性,对现有数据挖掘与分析工作提出了新的要求.

本文从数据的不一致性出发,分析现有多源异构的复杂网络不同数据,以及由此引发的信息过载所带来的数据差异性问题;同时采用粒计算范式,克服原有计算模式的单一粒度视角,建立异构多源数据的层次粒度表示.不一致问题带来的挑战,其背后的主要原因是异构数据之间存在的信息间隔(infor-

① https://en.wikipedia.org/wiki/Big_data

mation gap). 为此,本文提出数据的层次粒度表示用于处理信息间隔问题. 数据的层次粒度表示的主要工作是设计一个基于拓扑结构图的统计模型;同时,以结点属性表为基础构建统计模型的先验知识. 最后将这些异构的浅层模型表示的数据提升至更为抽象的高层信息粒,以期望在高层信息粒层,加以一定的约束条件,多源数据能够达到数据一致. 该链接预测模型的动机在于,试图在多层粒度空间上寻找链接预测问题的最优解.

1　相关工作

本节主要介绍与链接预测主题相关的研究背景以及研究现状.

现有大量链接预测相关的研究,其中最简单和直观的方法以结点的相似性为基础. 这类方法通过计算结点之间的相似性的评分来构造结点之间存在链接关系的可能性. 常见的相似性计算方法有公共邻居(common neighbors, CN)[5]、杰卡得系数(Jaccard index, JI)[6]、索尔顿系数(Salton index, SI)[7]、资源分配系数(resource allocation index, RAI)[8]和adamic-adar系数(adamic-adar index, AAI)[9]等. 这类方法主要使用拓扑结构上的邻居或路径等特征来计算结点对之间的相似程度. 很明显,这类方法的主要缺陷是缺乏考虑结点属性以及扩展性不足.

另一类方法聚焦在最大化观测结构的似然,建构图的生成模型. 然后,使用数据学习后的最优模型预测未知结点对之间的链接概率. Clauset等人[10]提出了一种从拓扑网络结构推断潜在层次组织结构的生成技术,并将该模型用于缺失链接的预测. 也有一些研究工作,使用概率相关模型(probabilstic relational models, PRM)[11]描述关系数据集(relational dataset, RD)的属性联合分布、优化分布,并将其用于结点对的链接关系预测. 相似的工作还有Relational Markov Networks[12], Relational Dependency Networks[13], Local Naïve Bayes Model[14]等. 还有一类链接预测方法则是基于机器学习技术[15-18].

此外,对于链接预测问题,每个社区对于结点的链接关系建立过程同样具有重要作用[19]. 例如对于一个中国的社交网络链接预测问题来说,大部分结点都属于中国这个社区;同时这些结点属于或不属于网球、游泳、足球等社区. 很明显,中国社区对于建立结点链接关系问题而言,影响程度小于网球社区、游泳社区、足球社区等社区. 也就是说社区在结点对建立链接关系的过程中所处的地位及作用是不同的. 相关的工作还有:王鑫等人[20]研究了交互意见和地位理论与链接关系的强相关性,提出了一种基于符号网络的链接预测模型;刘冶等人[21]研究了主数据源与附加数据源的特性,并提出了一种基于低秩和稀疏分解的多源融合链接预测算法;张泽华等人[22]将粗糙集理论引入图挖掘领域,提出了网络社区的领域粗糙化扩张方法等等. 值得注意的是,现有的这些方法,要么忽略了结点属性,要么拓扑网络数据与结点属性之间存在的潜在交互性,要么忽略了社区在结点对建立链接关系过程中的不同作用.

为了处理以上问题,本文首先对于各个来源的数据进行粒度表示学习. 具体来说,对于拓扑结点图数据,使用一种概率生成模型对图数据进行抽象表示(作用相当于提升数据信息粒层);对于结点属性表,使用聚类方法对数据进行抽象表示(提升数据信息粒层). 同时,对数据抽象后产生的高层数据信息粒加以一定条件的约束,以期在粒度表示的条件下,达到数据的一致,从而优化层次粒度表示模型,最终优化结点对链接关系预测的效果.

本文的贡献可以概括为3点:

1) 提出了一种关于拓扑结构图数据的概率生成模型. 这个模型充分考虑潜在社区贡献度因子,又考虑结点与社区之间的结点-隶属关系.

2) 提出了一种基于数据层次信息粒表示的问题求解方法. 该方法将原始多源异构数据抽象为不同层次结构下的信息粒,并考虑将不一致问题消除在这种层次信息粒表示的数据结构中.

3) 提出了基于粒度视角的链接预测方法,根据粒度计算范式,学习最优的层次粒表示模型,并将此模型用于表示缺失及观测链接关系的生成概率. 实验表明这一方法相较现有方法,有较为显著的性能提升.

2　基本知识

本节主要介绍数据的粒度表示以及信息粒在复杂网络链接预测模型中降低数据不一致性的重要作用. 粒计算是人们处理日常事务的一般性思维模式. 人们在计算现实世界问题时,通常是从多个角度,多个层次的观点看待问题,而不会局限于某一些局部特征,这一方法论也被称之为粒计算[23-24].

2.1　基本定义

在处理复杂网络数据时,根据粒计算理论,本文

将网络结构中观测到的网络拓扑结构数据与结点的属性数据(特征、标签等)归结为原始信息粒. 在原始信息粒的基础上,又可以构造当前信息粒的一种抽象表示(如图像处理过程中边缘是像素的一种抽象表示),形成高层信息粒,假如当前信息粒层不适合问题求解,可以在此基础上,继续构造上一层信息粒,直至当前信息粒有利于问题求得最优解. 由此,便形成了数据的层次粒度结构表示.

下面给出相关的定义.

定义 1. 属性图. 任意给定一个网络拓扑图 $G(V,E)$,其中网络拓扑图结点集 $V=\{v_i\}$, $i=1,2,\cdots,N$, N 为网络拓扑图的结点总数,链接边集 $E=\{e_{ij}\mid \forall i,j\leqslant N\}$ 为 V 上的一个二元关系,且有

$$e_{ij}=\begin{cases}1, & v_i,v_j \text{ 存在链接},\\ 0, & \text{其他}.\end{cases} \tag{1}$$

一般地,若网络拓扑图 $G(V,E)$ 的结点具有属性信息,则这一类型的网络拓扑图也称之为属性图,记为 $G(V,E,F)$,其中 F 为结点属性表.

定义 2. 结点属性表. 属性图 $G(V,E,F)$ 中任意结点 $u\in V$,除了具有与外部结点的链接关系外,还具有描述其各个侧面的特征 f_i 组成的属性集 $\{f_i\}_1^m$. 结点与特征之间具有如下映射关系:

$$f_i: u\rightarrow v^{f_i};\ \forall u\in V, f_i\in\{f_i\}_1^m, v^{f_i}\in V^{f_i}, \tag{2}$$

其中,v^{f_i} 为结点 v 在特征 f_i 映射值域 V^{f_i} 的一个具体属性值. 具有属性值的网络图结点可以表示为属性值向量 $\boldsymbol{v}^f=(v^{f_1},v^{f_2},\cdots,v^{f_m})$,所有结点属性向量集合为结点属性表,记为 F.

定义 3. 网络社区. 任意给定一个网络拓扑图 $G(V,E)$,其中的网络结点根据一定的划分规则潜在地属于某个类簇中心,即 $v\in C_i, i\in\{1,2,\cdots,k\}$ (k 为类簇总数). 在复杂网络分析上下文环境中,类簇中心也称之为网络社区.

特别地,由网络拓扑数据得到的网络社区也称之为拓扑网络社区,记为 $\{C_i\}, i\in\{1,2,\cdots,k\}$.

同理,给定结点的属性表,那么根据结点的属性相似度,潜在的存在着一个由所有结点组成的聚类族(属性表诱导的网络社区),记为 $\{Q_i\}, i\in\{1,2,\cdots,k\}$,这类网络社区称之为属性网络社区. 若结点可以同时属于多个网络社区,那么由这些结点所形成的社区之间,会存在结点的重叠,即 2 个或以上社区存在公共结点,这种类型的社区也称之为重叠社区.

3 层次粒度表示的属性图链接预测模型

3.1 问题描述

在介绍模型之前,首先将所研究的问题做个简单的描述. 一般地,假设给定属性图 $G(V,E,F)$,则链接预测问题的主要任务是根据观测到的现有结点间的链接关系集 $\{e_{ij}\mid \forall e_{ij}\in E\wedge e_{ij}>0\}$ 与结点属性表 F,试图建立任意结点对之间链接关系的似然估计.

如图 2 所示,在系统处理的原始数据中,拓扑结构图观测到的链接边与结点属性相似度诱导的潜在链接边存在不一致的情况,即:

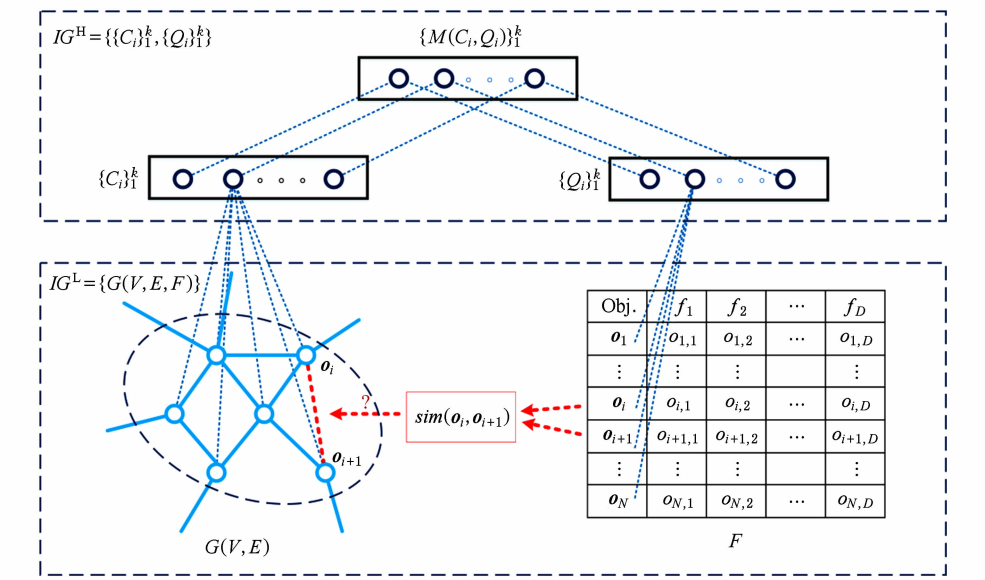

Fig. 2 The framework of the HGRLPM

图 2 层次粒度表示的属性图链接预测模型框架图

1) 拓扑结构图中结点对没有观测到链接关系,而属性表提供的结点相似度暗示着这对结点存在一条隐式链接;

2) 根据结点属性表计算出结点对相似度较低,表示存在链接关系的可能性较低,而拓扑结构图却观测到了结点对之间存在链接.

如果将所有对象映射至高层信息粒,即拓扑网络社区与属性网络社区,则可以在这种层次粒度结构的表示下,通过粒度转换融合异构数据,最大化地消除低层信息粒的不一致.信息粒度表示的粒层数量依赖于问题的规模及领域特点.

现有的属性图链接预测模型都忽略了这种潜在的数据不一致导致的冲突问题.基于此,不同于现有链接预测模型,本文提出层次粒度表示链接预测模型(hierarchical granular representation link prediction model, HGRLPM)将从数据的层次粒度表示出发,通过提升数据的粒度层次,消除低层次粒度的不一致性,最大化地融合异构的数据源,降低链接预测的不确定性,提升链接预测的准确性与精确性.下面引入基于层次信息粒表示的属性图链接预测模型.

3.2　层次粒度表示的属性图链接预测模型

本文的模型主要基于 Breiger 等人[25]与 Jaewon 等人[26]的工作,即每个结点依据不同的隶属度包含在不同的网络社区(每个结点与每个社区潜在地都具有包含关系,区别在于隶属度不同).如果任意 2 个结点所属的公共社区越多,那么它们将会以更高的概率建立链接关系.

同时,我们也注意到每一个网络社区在结点对建立链接关系中的重要程度是不同的.换句话说,每个社区对每个结点对建立链接的贡献度是不一致的.

最后,由于结点的属性表获取代价昂贵,所以研究数据对象以网络拓扑结构图为主体,而属性表诱导的结点社区隶属关系作为拓扑网络结构统计模型对应随机变量的先验分布.

定义 4. 结点-隶属关系矩阵.$\boldsymbol{B}$ 为一个非负矩阵,表示结点-隶属关系.$\boldsymbol{B}$ 的维度为 $|C|\times|V|$.每一个矩阵元素 $\boldsymbol{B}_{cu}$ 用于表示结点 $u(\forall u\in V)$ 与网络社区 $c(c\in\{C_i\}_1^k\triangleq\{C_1,C_2,\cdots,C_k\})$ 之间的隶属程度,k 为网络社区总数.

引理 1. 层次粒度表示链接预测模型(HGRLPM)通过相应的产生概率 $p(u,v)$ 建立任意结点链接边 (u,v),$\forall u,v\in V$,生成拓扑结构图 $G(V,E)$,且

$$p(u,v)=1-\exp(-\boldsymbol{s}^{\mathrm{T}}(\boldsymbol{B}_u\odot\boldsymbol{B}_v)),\tag{3}$$

其中,s_c 是一个表示网络社区 c 的贡献度的随机变量,向量 $\boldsymbol{s}=(s_1,s_2,\cdots,s_k)$;$\boldsymbol{B}_u,\boldsymbol{B}_v$ 分别是结点 u,v 的结点-隶属向量,代表结点-隶属矩阵的一列;"$\odot$"为逐元素乘法.

证明. 式(3)隐含的表示拓扑网络图中每一对结点都具有一个潜在的交互.本文假设任意一对结点(u,v)在任意网络社区 c 产生一个非负的交互作用 $T_{uv}^{(c)}$,且服从均值为 $s_cB_{cu}B_{cv}$ 的泊松分布,即

$$T_{uv}^{(c)}\sim Poi(s_cB_{cu}B_{cv}),\tag{4}$$

其中,$Poi(\cdot)$为泊松分布.根据泊松分布的性质,可知,网络社区对结点对(u,v)的相互作用总量 T_{uv} 为

$$T_{uv}=\sum_{c\in\{C_i\}_1^k}T_{uv}^{(c)}\sim Poi(\boldsymbol{s}^{\mathrm{T}}(\boldsymbol{B}_u\odot\boldsymbol{B}_v)),\tag{5}$$

可以得到链接概率 $p(u,v)=P(T_{uv}>0)$,即

$$P(T_{uv}>0)=1-P(T_{uv}=0)=1-\exp(-\boldsymbol{s}^{\mathrm{T}}(\boldsymbol{B}_u\odot\boldsymbol{B}_v)).\tag{6}$$

证毕.

3.2.1　拓扑结构图对数似然

假设给定潜在因子矩阵,即结点-隶属关系矩阵 $\boldsymbol{B}$ 与网络社区贡献度因子向量 $\boldsymbol{s}$,拓扑结构图 G 生成模型的似然概率记为 $\mathcal{L}(\boldsymbol{B},\boldsymbol{s})$,那么有:

$$\mathcal{L}(\boldsymbol{B},\boldsymbol{s})=\ln P(G\mid\boldsymbol{B},\boldsymbol{s})=\sum_{(u,v)\in E}\ln(1-\exp(-\boldsymbol{s}^{\mathrm{T}}(\boldsymbol{B}_u\odot\boldsymbol{B}_v))-\sum_{(u,v)\notin E}\boldsymbol{s}^{\mathrm{T}}(\boldsymbol{B}_u\odot\boldsymbol{B}_v).\tag{7}$$

在没有考虑结点属性表的情况下,通过求解以下问题就可以得到最优模型,即

$$\boldsymbol{B}^*,\boldsymbol{s}^*=\arg\max_{\boldsymbol{B}\geq0,\boldsymbol{s}\geq0}\mathcal{L}(\boldsymbol{B},\boldsymbol{s}).\tag{8}$$

然而,在这种情况下,模型没有考虑结点属性以及拓结构图数据与属性数据之间潜在的不一致性.一般地,在建构一个全面、鲁棒的链接预测模型时,不仅需要考虑集成异构数据,同时也需要考虑消除异构数据的不一致性.

3.2.2　结点-隶属关系先验

当我们考虑结点的属性信息时,这表示根据某种相似度测度可以将所有的结点按照它们之间的亲疏程度,划分为 k 个聚类簇.这里为了保持数据整体上是一致的,假设拓扑结构以及属性信息各自产生的数据概括是相同的,即拓扑图产生的类簇与属性图产生的类簇个数是相等的.在本文中,属性表产生的结点与聚类簇的隶属程度记为

$$D_{cu}(\forall u\in V,c\in\{Q_i\}_1^k\triangleq\{Q_1,Q_2,\cdots,Q_k\}),$$

并将此信息作为拓扑结构图的结点-隶属关系矩阵的一个先验信息.

下面是矩阵 $\boldsymbol{D}$ 的产生过程.首先,我们使用某种相似度测度 $\mathcal{M}(\cdot,\cdot)$计算结点相似度,并以此测度为基础,使用某种聚类算法对结点属性表进行聚

类.这将产生一个原有结点集的一个划分;其次,计算每个结点 u 与所有聚类簇中心$\{c|c\in\{Q_i\}_1^k\}$的相似度,即

$$D_{cu}=\mathcal{M}(u,c),\ \forall u\in V,c\in\{Q_i\}_1^k, \quad (9)$$

$\mathcal{M}(\cdot,\cdot)$一般使用欧氏距离,也可以根据数据特点选择余弦距离、马氏距离等.

3.2.3 社区贡献度先验

我们假设每一个网络社区在结点对建立链接的过程中的贡献度是不同的.例如当我们在分析中国的社交网络链接问题时,如果所有结点所属的社区都为中国,那么这个社区对于链接预测任务而言,贡献度是可以忽略的;而根据某种爱好划分的社区,如音乐、体操、乒乓球等社区对于结点间的链接关系建立具有很强的驱动力.

显然,网络社区的重要性可以从当前社区的链接数在全体网络链接所占的比例观察出来.这也是为什么使用拓扑结构图的链接密度作为社区贡献度的先验,即

$$\mu_c=\frac{\varphi(c)}{\varphi(G)},\ \forall c\in\{Q_i\}_1^k, \quad (10)$$

其中,函数 $\varphi(\cdot)$用于计算结点集所具有的链接边的个数.

3.2.4 结点-隶属关系重要度

由于社区的贡献度不一样,所以结点-隶属关系成员的重要程度也不一样.在系统建模时,应当考虑结点-隶属关系成员的重要度.在本文中,结点-隶属关系重要度为

$$W_{ij}=\pi\mu_j, \quad (11)$$

其中,$i\in\{1,2,\cdots,|V|\}$,$j\in\{1,2,\cdots,k\}$,π 为自定义常数.W_{ij}表示当前结点与社区中心的隶属关系的重要度与社区在结点建立链接关系时的重要度成正比关系.

3.2.5 层次粒度表示

在获得先验信息$\{\{Q_i\}_1^k,\boldsymbol{\mu},\boldsymbol{D}\}$与拓扑网络结构图$G(V,E)$之后,我们可以获得数据的层次粒度表示.

定义 5. 原始信息粒、高层信息粒.一般地,信息系统所采集的经过简单数据清洗后得到的数据被认为是信息系统输入的原始数据.原始数据称之为原始信息粒,记为 IG^{L}.对原始数据依据某种规则抽象之后形成原有数据的一个概括描述,即高层信息粒,记为 IG^{H}.

定义 6. 层次信息粒化表示.原始信息粒和高层信息粒,以及由这些不同层次信息粒形成的层次结构,称之为数据的层次粒化表示,记为 R^{HI}.本文将属性图 $G(V,E,F)$归属于原始信息粒 IG^{L},拓扑网络社区$\{C_i\}_1^k$与属性网络社区$\{Q_i\}_1^k$归属于高层信息粒 IG^{H},即

$$\begin{aligned}IG^{\mathrm{L}}&=\{G(V,E,F)\},\\ IG^{\mathrm{H}}&=\{\{C_i\}_1^k,\{Q_i\}_1^k\}.\end{aligned} \quad (12)$$

由此,$R^{\mathrm{HI}}=\{IG^{\mathrm{L}},IG^{\mathrm{H}}|\boldsymbol{B},\boldsymbol{D},\boldsymbol{W},\boldsymbol{s},\boldsymbol{\mu}\}$为原始数据的层次粒度表示,其中,$\{\boldsymbol{B},\boldsymbol{D},\boldsymbol{W},\boldsymbol{s},\boldsymbol{\mu}\}$为层次粒度表示的参数集.

在层次粒度表示 R^{HI}的结构基础上,根据粒计算范式,我们知道在原始信息粒、结点的拓扑结构与结点的属性是结点的各个侧面的信息,会产生差异,也就是不一致性,参考图 2.而在高层信息粒,也就是拓扑网络社区$\{C_i\}_1^k$与属性网络社区$\{Q_i\}_1^k$,作为原始信息粒的一个较为抽象的全貌,其差异性要小于原始信息粒.模型的框架图在图 2 中描述.为了消除不一致性提高结点链接预测质量,我们需要达到 3 个目标:

1) 最小化高层信息粒 $IG^{\mathrm{H}}=\{\{C_i\}_1^k,\{Q_i\}_1^k\}$的差异,也就是矩阵 $\boldsymbol{B}$ 与 $\boldsymbol{D}$ 的差异应该最小化.

2) 属性表诱导的社区重要度与拓扑结构图生成过程中的社区的重要度的差异也应最小化.

3) 根据最大似然估计法,我们需要最大化拓扑结构图的对数似然.

因此,本文提出层次粒度表示链接预测模型,即

$$\begin{aligned}&\max_{\boldsymbol{B}\geq0,\boldsymbol{s}\geq0}\mathcal{L}(\boldsymbol{B},\boldsymbol{s}),\\ &\text{s.t. }\psi(\{C_i\}_1^k,\{Q_i\}_1^k)=0,\\ &\eta(\boldsymbol{s},\boldsymbol{\mu})=0,\end{aligned} \quad (13)$$

其中,$\psi(\cdot,\cdot)$为高层信息粒 IG^{H} 成员粒之间距离度量函数,$\eta(\cdot,\cdot)$为 $\boldsymbol{s}$ 与 $\boldsymbol{\mu}$ 的距离度量,用于测量 2 个输入变量的相似程度.矩阵 $\boldsymbol{B}$ 和 $\boldsymbol{D}$ 为构建高层信息粒层的参数.

在本文中,选取$\|\boldsymbol{W}\odot(\boldsymbol{B}-\boldsymbol{D})\|_{\mathrm{F}}^2$ 作为高层粒的信息距离度量函数 $\psi(\{C_i\}_1^k,\{Q_i\}_1^k)$,$\|\cdot\|_{\mathrm{F}}$ 为 Frobenius 范数,W_{ij}为权值,表示当前结点-隶属关系差分在不同社区的重要度.同时,选择 $\eta(\cdot,\cdot)=\|\boldsymbol{s}-\boldsymbol{\mu}\|_2^2$,其中$\|\cdot\|_2$为向量的 2 范数.由此,式(13)转化为

$$\begin{aligned}&\max_{\boldsymbol{B}\geq0,\boldsymbol{s}\geq0}\mathcal{L}(\boldsymbol{B},\boldsymbol{s}),\\ &\text{s.t. }\|\boldsymbol{W}\odot(\boldsymbol{B}-\boldsymbol{D})\|_{\mathrm{F}}^2=0,\\ &\|\boldsymbol{s}-\boldsymbol{\mu}\|_2^2=0.\end{aligned} \quad (14)$$

根据拉格朗日乘子法,可以将式(14)转化为以下优化问题:

$$\begin{aligned}\min_{\boldsymbol{B}\geq0,\boldsymbol{s}\geq0}&-\mathcal{L}(\boldsymbol{B},\boldsymbol{s})+\frac{\alpha}{2}\|\boldsymbol{W}\odot(\boldsymbol{B}-\boldsymbol{D})\|_{\mathrm{F}}^2+\\ &\frac{\beta}{2}\|\boldsymbol{s}-\boldsymbol{\mu}\|_2^2.\end{aligned} \quad (15)$$

为方便计算,将式(15)的目标函数记为 $\ell(\boldsymbol{B},\boldsymbol{s})$.

4　模型参数学习

模型所有需要学习的参数为矩阵 $\boldsymbol{B}$ 与向量 $\boldsymbol{s}$,记为 $\Theta=\{\boldsymbol{B},\boldsymbol{s}\}$. 对于式(15)这个优化问题,当固定参数 $\boldsymbol{s}$ 与其他的 $\boldsymbol{B}_v(\forall v\in V\wedge v\neq u)$ 时,我们发现 $\ell(\boldsymbol{B},\boldsymbol{s})$ 是 $\boldsymbol{B}_u(\forall u\in V)$ 的凸函数(convex function). 同理,$\ell(\boldsymbol{B},\boldsymbol{s})$ 是 $\boldsymbol{s}$ 的凸函数. 因此,选择块座标梯度下降算法[27](block coordinate gradient descent algorithm, BCGDA)来得到模型参数的最优解.

原式(13)可以分解为

$$\boldsymbol{B}_u^*=\arg\min_{\boldsymbol{B}_u}\ell_1(\boldsymbol{B}_u,\boldsymbol{B}_{-u}^{+},\boldsymbol{s}^{+}),\forall u\in V, \quad (16)$$

与

$$\boldsymbol{s}^*=\arg\min_{\boldsymbol{s}}\ell_2(\boldsymbol{B}^{+},\boldsymbol{s}), \quad (17)$$

其中,$\boldsymbol{B}_u$ 为列向量,表示 $\boldsymbol{B}$ 的第 u 列. $\boldsymbol{B}^{+},\boldsymbol{s}^{+}$ 分别表示矩阵 $\boldsymbol{B}$ 或向量 $\boldsymbol{s}$ 的值是固定的,在当前优化过程中. $\boldsymbol{B}_{-u}^{+}$ 表示的是除了 $\boldsymbol{B}_u$,其他的列都是固定的.

式(16)的目标函数为

$$\ell_1(\boldsymbol{B}_u,\boldsymbol{B}_{-u}^{+},\boldsymbol{s}^{+})=-\sum_{v\in\mathcal{N}(u)}\ln(1-\exp(-\boldsymbol{s}^{+\mathrm{T}}(\boldsymbol{B}_u\odot\boldsymbol{B}_v^{+})))+\sum_{v\notin\mathcal{N}(u)}\boldsymbol{s}^{+\mathrm{T}}(\boldsymbol{B}_u\odot\boldsymbol{B}_v^{+})+\frac{\alpha}{2}\|\boldsymbol{W}_u\odot(\boldsymbol{B}_u-\boldsymbol{D}_u)\|_{\mathrm{F}}^2+\frac{\beta}{2}\|\boldsymbol{s}^{+}-\boldsymbol{\mu}\|_2^2, \quad (18)$$

其中,$\mathcal{N}(u)$ 为结点 u 在拓扑结构图中具有链接关系的结点集,$\boldsymbol{B}_v^{+}$ 为 $\boldsymbol{B}^{+}$ 的第 v 列.

式(17)的目标函数为

$$\ell_2(\boldsymbol{B}^{+},\boldsymbol{s})=-\sum_{(u,v)\in E}\ln(1-\exp(-\boldsymbol{s}^{\mathrm{T}}(\boldsymbol{B}_u^{+}\odot\boldsymbol{B}_v^{+})))+\sum_{(u,v)\notin E}\boldsymbol{s}^{\mathrm{T}}(\boldsymbol{B}_u^{+}\odot\boldsymbol{B}_v^{+})+\frac{\alpha}{2}\|\boldsymbol{W}\odot(\boldsymbol{B}^{+}-\boldsymbol{D})\|_{\mathrm{F}}^2+\frac{\beta}{2}\|\boldsymbol{s}-\boldsymbol{\mu}\|_2^2. \quad (19)$$

BCGDA 将迭代地按某一个顺序,循环地优化式(14)和式(15). BCGDA 的一个最基本的要求是必须计算式(16)和式(17)的梯度.

下面,给出每个目标函数各自相对应的梯度,它们分别为

$$\frac{\partial\ell_1}{\partial\boldsymbol{B}_u}=\sum_{v\in\mathcal{N}(u)}\frac{-\exp(-\boldsymbol{s}^{+\mathrm{T}}(\boldsymbol{B}_u\odot\boldsymbol{B}_v^{+}))}{1-\exp(-\boldsymbol{s}^{+\mathrm{T}}(\boldsymbol{B}_u\odot\boldsymbol{B}_v^{+}))}(\boldsymbol{s}^{+}\odot\boldsymbol{B}_v^{+})+\sum_{v\notin\mathcal{N}(u)}(\boldsymbol{s}^{+}\odot\boldsymbol{B}_v^{+})+\alpha(\boldsymbol{W}_u\odot(\boldsymbol{B}_u-\boldsymbol{D}_u)), \quad (20)$$

与

$$\frac{\partial\ell_1}{\partial\boldsymbol{s}}=\sum_{(u,v)\in E}\frac{-\exp(-\boldsymbol{s}^{\mathrm{T}}(\boldsymbol{B}_u^{+}\odot\boldsymbol{B}_v^{+}))}{1-\exp(-\boldsymbol{s}^{\mathrm{T}}(\boldsymbol{B}_u^{+}\odot\boldsymbol{B}_v^{+}))}(\boldsymbol{B}_u^{+}\odot\boldsymbol{B}_v^{+})+\sum_{(u,v)\notin E}(\boldsymbol{s}\odot\boldsymbol{B}_v^{+})+\beta(\boldsymbol{s}-\boldsymbol{\mu}). \quad (21)$$

下面给出本文提出的层次粒度表示链接预测模型(HGRLPM)的参数学习算法.

算法 1. HGRLPM 参数学习算法.

输入:精度 $\varepsilon^{\mathrm{tol}}$、属性图 $G(V,E,F)$、社区数 k、学习率 γ、重要度系数 π、最大迭代次数 $MAXITER$;

输出:结点-隶属关系矩阵 $\boldsymbol{B}$、社区贡献度 $\boldsymbol{s}$.

① 根据属性表 F,使用 K-means 聚类算法,计算 k 个聚类中心 $\{Q_i\}_1^k$;
② for $u=1,2,\cdots,|V|$ do
③　for $c=1,2,\cdots,k$ do
④　　$D_{cu}\leftarrow$式(9);
⑤　end for
⑥ end for
⑦ for $c=1,2,\cdots,k$ do
⑧　$\mu_c\leftarrow$式(10);
⑨ end for
⑩ for $u=1,2,\cdots,|V|$ do
⑪　for $c=1,2,\cdots,k$ do
⑫　　$W_{cu}\leftarrow$式(11);
⑬　end for
⑭ end for
⑮ 初使化 $\boldsymbol{B},\boldsymbol{s}$;$t\leftarrow1$;计算 $r^{(t)}\leftarrow$式(13);
⑯ while $t<MAXITER$ do
⑰　for $u=1,2,\cdots,|V|$ do
⑱　　$\boldsymbol{B}_u\leftarrow\gamma\times$式(18);
⑲　end for
⑳　$\boldsymbol{s}\leftarrow\gamma\times$式(19);$t\leftarrow t+1$;
㉑　计算 $r^{(t+1)}\leftarrow$式(13);
㉒　if $|r^{(t+1)}-r^{(t)}|<\varepsilon^{\mathrm{tol}}$
㉓　　break;
㉔　end if
㉕ end while
㉖ 返回 $\boldsymbol{B},\boldsymbol{s}$.

5　层次粒度表示模型的链接预测

经过数据训练后,获得当前数据的一个层次粒度表示 $R^{\mathrm{HI}}=\{IG^{\mathrm{L}},IG^{\mathrm{H}}\mid\boldsymbol{B}^*,\boldsymbol{D}^*,\boldsymbol{s}^*,\boldsymbol{\mu}^*\}$,其中 $\boldsymbol{B}^*,\boldsymbol{D}^*,\boldsymbol{s}^*,\boldsymbol{\mu}^*$ 是从数据中学习的参数最优值. 模型 HGRLPM 将在学习好的层次粒表示 R^{HI} 基础上,执行链接预测任务. 为此我们设计了以下预测模型.

5.1　协同预测

给定层次粒度表示 R^{HI},以及查询结点 (u,v),对于信息缺失的结点 u,根据 u 的属性信息,寻找与

u 最相似的 n 个结点,并以这些结点的潜在结点-隶属向量的期望作为潜变量 $\boldsymbol{B}_u$ 的值,称之为预测向量,记为 $\bar{\boldsymbol{B}}_u$,即

$$\bar{\boldsymbol{B}}_u \triangleq \frac{1}{n}\sum_{t\in nnb(u)}\boldsymbol{B}_t^*, \tag{22}$$

其中,$nnb(u)$为结点 u 最相似的 n 个结点集合. 若 v 也是信息缺失,那么执行与结点 u 相同的处理过程.

在此基础上,结点对(u,v)建立链接关系的协同预测概率 $p(u,v)$,即

$$p(u,v)=1-\exp(\boldsymbol{s}^{\mathrm{T}}(\bar{\boldsymbol{B}}_u\odot\bar{\boldsymbol{B}}_v)). \tag{23}$$

特别地,当 $n=1$ 时,这种协同预测策略称之为基础预测.

5.2 预测算法

下面给出基于层次粒度表示协同链接预测算法的详细步骤:

算法 2. 协同链接预测算法.

输入:层次粒度表示 R^{HI}、预测结点对(u,v)、协同数 n;

输出:结点对链接概率 $p(u,v)$.

① 根据属性表 F,协同数 n,计算结点对(u,v)的期望预测向量:$\bar{\boldsymbol{B}}_u\leftarrow$ 式(22),$\bar{\boldsymbol{B}}_v\leftarrow$ 式(22);

② 计算链接概率:$p(u,v)\leftarrow$ 式(23);

③ 输出 $p(u,v)$.

6 实验与结果

在本节中,我们设计了关于引言部分介绍的示例数据集,以及 2 个真实数据数据集 AmazonFail[28] 和 Lazega[29] 的实验. 同时,我们对比了 HGRLPM 与其他的算法的预测性能,实验结果显示 HGRLPM 模型相对于其他的方法具有较强的优越性.

6.1 数据集

示例数据集的拓扑结构图已在引言部分介绍过了,表 1 给出了示例数据的结点属性值,表 2 给出了所有数据集的一些基本统计信息.

AmazonFail 数据集是从 Amazon 网站所收集的. 该数据集总共有 1 418 个结点,每一个都代表 Amazon 网站出售的一件产品. 3 695 条链接边建立了这些结点之间的相关关系. 此外,这个数据集还提供了产品的属性以及标签信息. 标签信息用于标记用户对该产品的不满意度.

Lazega 数据集是一个公司法律事务所关于公司法合伙关系的属性图. 它包括该公司的 71 名律师(合伙人和同事)之间的网络关系. 这个数据集常用于社区网络分析,例如有限团结、横向控制、质量控制、知识共享、权力平衡、监管等等.

Table 1 The Attribute Information of the Toy Dataset

表 1 示例数据集结点的属性信息

Node	Age	Math	English	Art	Physics
Node_1	13	70	71	88	95
Node_2	14	81	88	93	85
Node_3	14	75	89	87	79
Node_4	14	86	85	95	81
Node_5	13	95	86	90	92
Node_6	14	89	91	92	88
Node_7	13	87	89	88	83
Node_8	15	70	72	90	81
Node_9	14	71	73	70	87
Node_{10}	14	92	88	90	93
Node_{11}	13	70	75	81	71
Node_{12}	14	69	89	87	84
Node_{13}	14	65	70	90	69

Table 2 Basic Statistics of Datasets

表 2 数据集基本统计信息

Dataset	Nodes	Edges	Attributes
Toy	13	17	5
AmazonFail	1 418	3 695	28
Lazega	71	650	8

6.2 对比标准

对于模型输出的链接概率,设置一个阈值 ε,如果 $p(u,v)>\varepsilon$,那么认为结点对(u,v)存在一条边,即 $E(u,v)=1$,反之为 0. 那么,使用 3 个指标来对比算法的性能:

$$\begin{aligned} Precision&=\frac{TP}{TP+FP},\\ Recall&=\frac{TP}{TP+FN},\\ Accuracy&=\frac{TP+TN}{TP+TN+FP+FN}. \end{aligned} \tag{24}$$

其中,TP(true positive)为正确正例,TN(true negative)为正确负例,FP(false positive)为错误正例,FN(false negative)为错误负例.

6.3 案例分析

在本节将详细分析算法在示例数据集上的性能,以及与其他算法的对比. 对比的算法为第 1 部分介绍的杰卡得系数(JI)和资源分配系数(resource allocation index, RA)和 LHNI (Leicht Holme Newman index)系数.

由上述示例数据集的背景知识,假设社区数量 $k=3$,学习率 $\gamma=0.001$,正则项系数 $\alpha=0.03$,$\beta=0.005$,在这些参数设定条件下,算法经过 50 次迭代后目标函数逐渐稳定,最终在第 59 次迭代后收敛于 6 020.49.具体的收敛过程如图 3 所示:

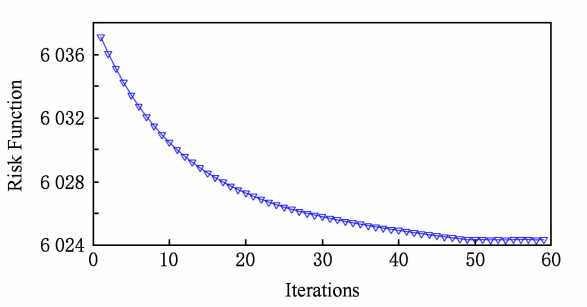

Fig. 3　The convergence of the HGRLPM model training processes with the Toy dataset

图 3　HGRLPM 模型在示例数据集上的训练收敛过程

训练过程结束后,我们得到了用于表示模型 HGRLPM 的最优参数值 $\boldsymbol{B}^*$,$\boldsymbol{s}^*$.图 4 给出了 $\boldsymbol{B}^*$ 和 $\boldsymbol{s}^*$ 的参数值分布.从图 4 我们可以发现每个社区参与建立结点对之间的链接关系的贡献度 $\boldsymbol{s}^*$ 是不同的,社区 1 的贡献度最小,社区 2 的贡献度最大,社区 3 的贡献介于社区 1 和社区 2 之间.同时,社区贡献度 $\boldsymbol{s}^*$ 还影响着结点与社区之间的隶属关系 $\boldsymbol{B}^*$.模型 HGRLPM 的最优参数值也代表着:属性表导出的结点与类簇的隶属关系 $\boldsymbol{D}$ 与拓扑结构图产生的 $\boldsymbol{B}$ 在 $\boldsymbol{B}^*$ 时差异最小.这也表明通过最优化技术,低层信息粒的不一致性在高层信息粒得到了最大化的消除.

通过数据训练得到最优模型 HGRLPM 后,我们根据链接预测算法,对属性图的所有结点,预测其链接的生成概率.图 5 显示了模型 HGRLPM 预测的属性图中的示例训练集观测到的所有链接生成概率,以及潜在链接(生成概率大于 50%的边)的生成概率.在图 5 中,虚线代表为错误正例,实线为正确正例,链接边的数字为建立链接的概率.

另外,表 3 给出了所有的结点对的链接边(观测到的以及未观测到的结点链接关系)预测概率.

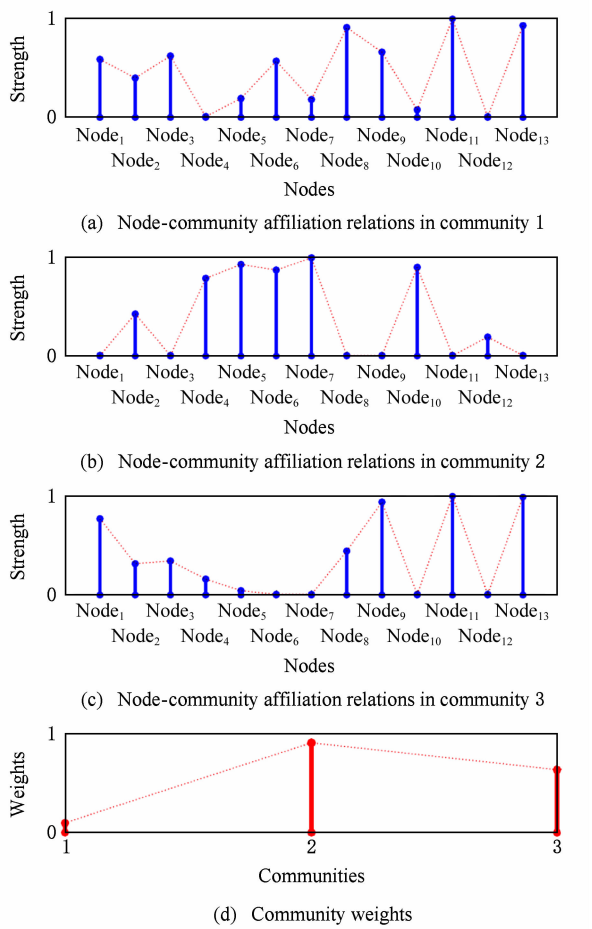

Fig. 4　The community weights and node-community affiliation relations learnt by the HGRLPM

图 4　HGRLPM 得到的社区重要度与结点-社区关系

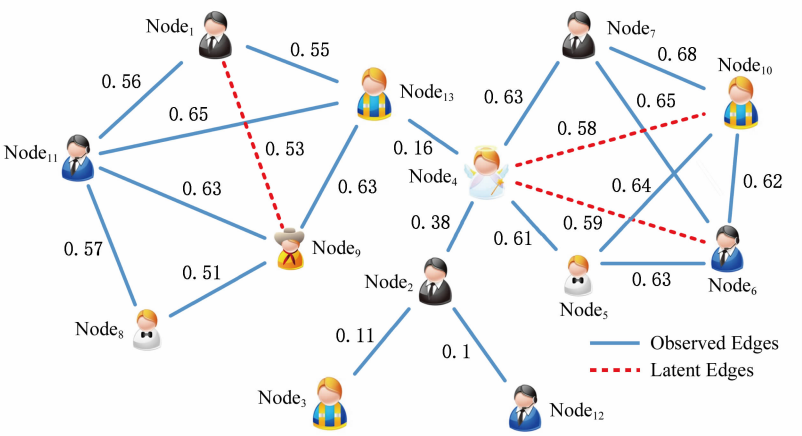

Fig. 5　The link prediction result of the HGRLPM model

图 5　HGRLPM 模型链接预测结果

Table 3 All of the Link Probability Predicted by the HGRLPM for the Node Pairs

表 3 HGRLPM 预测的所有结点对链接概率

Nodes	$Node_1$	$Node_2$	$Node_3$	$Node_4$	$Node_5$	$Node_6$	$Node_7$	$Node_8$	$Node_9$	$Node_{10}$	$Node_{11}$	$Node_{12}$	$Node_{13}$
$Node_1$	0.00	0.12	0.26	0.11	0.09	0.10	0.09	0.25	0.53	0.01	0.56	0.02	0.55
$Node_2$	0.12	0.00	0.11	0.38	0.36	0.38	0.36	0.13	0.27	0.37	0.28	0.10	0.25
$Node_3$	0.26	0.11	0.00	0.06	0.03	0.01	0.02	0.15	0.29	0.00	0.30	0.01	0.30
$Node_4$	0.11	0.38	0.06	0.00	0.61	0.59	0.63	0.07	0.16	0.58	0.15	0.17	0.16
$Node_5$	0.09	0.36	0.03	0.61	0.00	0.63	0.01	0.02	0.05	0.64	0.05	0.21	0.03
$Node_6$	0.10	0.38	0.01	0.59	0.63	0.00	0.65	0.01	0.01	0.62	0.01	0.17	0.03
$Node_7$	0.09	0.36	0.02	0.63	0.01	0.65	0.00	0.01	0.00	0.68	0.00	0.16	0.04
$Node_8$	0.25	0.13	0.15	0.07	0.02	0.01	0.01	0.00	0.51	0.03	0.57	0.00	0.38
$Node_9$	0.53	0.27	0.29	0.16	0.05	0.01	0.00	0.51	0.00	0.00	0.63	0.00	0.63
$Node_{10}$	0.01	0.37	0.00	0.58	0.64	0.62	0.68	0.03	0.00	0.00	0.00	0.23	0.01
$Node_{11}$	0.56	0.28	0.30	0.15	0.05	0.01	0.00	0.57	0.63	0.00	0.00	0.00	0.65
$Node_{12}$	0.02	0.10	0.01	0.17	0.21	0.17	0.16	0.00	0.00	0.23	0.00	0.00	0.01
$Node_{13}$	0.55	0.25	0.30	0.16	0.03	0.03	0.04	0.38	0.63	0.01	0.65	0.01	0.00

从表 3 可以看出,图 5 中有 4 条边划分为错误的负例,即{($Node_4$,$Node_{13}$):0.16,($Node_2$,$Node_4$):0.38,($Node_2$,$Node_3$):0.11,($Node_2$,$Node_{12}$):0.10};同时,有 3 条未能观测到的边划分为正例,即错误的正例{($Node_1$,$Node_9$):0.53,($Node_4$,$Node_6$):0.59,($Node_4$,$Node_{10}$):0.58}.可以得到以下评价分:

$$Precision=\frac{14}{14+3}=82.35\%,$$

$$Recall=\frac{14}{14+4}=77.78\%,$$

$$Accuracy=\frac{14+57}{78}=91.02\%.$$

为了验证算法的性能,我们将原有的拓扑相似索引 JI[6],RA[8],LHNI[8]进行扩展,使之能够同时利用拓扑和属性表信息.然后与 HGRLPM 进行对比.具体扩展如下:

1) 计算 JI,RA,LHNI 的相似度;

2) 计算属性表中结点对的修正余弦相似度(adjust cosine similarity, ACOS),即

$$\delta^{\mathrm{ACOS}}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\frac{(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\cdot(\boldsymbol{x}_j-\bar{\boldsymbol{x}})}{\|\boldsymbol{x}_i-\bar{\boldsymbol{x}}\|_2\times\|\boldsymbol{x}_j-\bar{\boldsymbol{x}}\|_2}. \tag{25}$$

3) 融合 2 个相似度,以 JI 索引为例,融合 ACOS 与 JI,即

$$\delta^{\mathrm{JIA}}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\theta\times\delta^{\mathrm{JI}}(\boldsymbol{x}_i,\boldsymbol{x}_j)+(1-\theta)\times\delta^{\mathrm{ACOS}}(\boldsymbol{x}_i,\boldsymbol{x}_j), \tag{26}$$

其中,$\boldsymbol{x}_i$,$\boldsymbol{x}_j$ 为属性表中的任意结点的特征向量,θ 为权值参数,可以设置属性表与拓扑结构图的重要性.扩展后的 3 个算法分别记为 JIA,RAA,LHNIA.

我们使用 3 个扩展算法和相同的对比标准,在示例数据集上可以得到表 4 的实验结果:

Table 4 Performance Comparison of Toy

表 4 示例数据集的性能对比

Algorithms	*Precison*	*Recall*	*Accuracy*
HGRLPM	0.8235	0.7778	0.9102
JIA	0.4222	0.4485	0.7590
RAA	0.889	0.1881	0.7436
LHNIA	0.3111	0.3353	0.7590

表 4 的 3 个方法(JIA,RAA,LHNIA)的结果为 θ 取值 0.3:0.05:0.7 的 9 次计算结果的均值.这 3 个方法的 *Precision*,*Recall*,*Accuracy* 的方差分别为(JIA:0.2243,0.0924,0.0263),(RAA:0.1152,0.1946,0.0128)和(LHNIA:0.2955,0.2274,0.0190).实验结果表明:无论是 *Accuracy* 还是 *Precision* 以及 *Recall* 指标,HGRLPM 模型预测的结果对比 JIA,RAA,LHNIA 有着显著的提升.而且,当 θ 的值发生变化时,算法的性能波动较大,这主要体现在方差的变化.原因在于 JIA,RAA,LHNIA 等算法是建立在原始信息上的单粒度表示之上,拓扑结构图与属性表的原始信息融合容易出现偏差,异构数据源的冲突较为明显.这也证明了在原始信息粒上直接处理信息融合是具有挑战性的问题.HGRLPM 充分挖掘了潜在社区变量的分布,以及社区作用的不平衡,在多层信息粒的表示下,将数据的不一致性上升到高层信息粒,可以最大化地消除原始信息粒上

较难处理的融合问题.这一对比结果映证了我们的假设,也显示了 HGRLPM 的优越性.

6.4　其他数据集上的结果

采取与示例数据一致的评价标准,表 5 和表 6 分别显示了在数据集 AmazonFail 和 Lazega 数据集上的算法性能.

Table 5　Performance Comparison of AmazonFail

表 5　AmazonFail 的性能对比

Algorithms	*Precison*	*Recall*	*Accuracy*
HGRLPM	0.235 1	0.339 2	0.846 2
JIA	0.555 0	0.205 6	0.995 7
RAA	0.000 0	0.000 0	0.996 1
LHNIA	0.000 0	0.000 0	0.996 1

Table 6　Performance Comparison of Lazega

表 6　Lazega 的性能对比

Algorithms	*Precison*	*Recall*	*Accuracy*
HGRLPM	0.409 1	0.641 2	0.815 7
JIA	0.185 0	0.582 2	0.739 9
RAA	0.000 0	0.000 0	0.738 4
LHNIA	0.000 0	0.000 0	0.738 4

虽然 HGRLPM 能够取得比 JIA,RAA,LHNIA 要好的 *Precison* 和 *Recall* 成绩,但 *Accuracy* 却提升不是很显著,甚至在 AmazonFail 数据集上,*Accuracy* 指标在还有较大的差距离,主要原因在于拓扑图数据的结点链接的稀疏性.当数据集规模不大时稀疏性不会对算法性能产生很严重的影响,然而当数据集规模扩大到一定程度时,稀疏性将严重影响预测的准确性.这一现象也称之为类别不平衡问题[30].最直观的影响在于正例淹没在负例的海洋.对于 *Precision* 和 *Recall* 为 0,这说明 RAA 和 LHNIA 在数据集 AmazonFail 和 Lazega 上所有的判例都为负,不能识别正例,这也证实了底层原始数据源的不一致性.同时,这也说明了当数据规模扩大时,HGRLPM 应该在建模时对数据的稀疏性这一数据因子加以考虑.

7　结　　论

本文提出了一种融合异构数据(网络拓扑图与结点属性表)的层次粒度表示模型,根据粒计算理论,对于低层信息粒中的数据不一致性,通过提升粒层的方法,在高层信息粒最大化的消除异构数据的不一致性.该方法的最大优势在于它能捕捉数据潜在的层次粒度结构,同时也最大化的捕捉了数据的语义.实验结果表明,层次信息粒表示的链模型是有效的,对比其他方法有较大优势.

参　考　文　献

[1] Lü Linyuan, Zhou Tao. Link prediction in complex networks: A survey [J]. Physica A: Statistical Mechanics & Its Applications, 2011, 390(6): 1150-1170

[2] Zhao Shu, Liu Xiaoman, Duan Zhen, et al. A survey on social ties mining [J]. Chinese Journal of Computers, 2017, 40(3): 535-555 (in Chinese)
(赵姝, 刘晓曼, 段震, 等. 社交关系挖掘研究综述[J]. 计算机学报, 2017, 40(3): 535-555)

[3] Libennowell D, Kleinberg J. The link prediciton problem for social networks [J]. Journal of the Association for Information Science & Technology, 2007, 58(7): 1019-1031

[4] Kim M, Leskovec J. Multiplicative attribute graph model of real world networks [C] //Proc of the Int Workshop on Algorithms and Models for the Web-Graph. Berlin: Springer, 2010: 62-73

[5] Newman M E J. Clustering and preferential attachment in growing networks [J]. Physical Review E-Statistical, Nonlinear and Soft Matter Physics, 2001, 64(2): 025102

[6] Jaccard P. Etude de la distribution florale dans une portion des Alpes et du Jura [J]. Bulletin De La Societe Vaudoise Des Sciences Naturelles, 1901, 37(142): 547-579

[7] Salton G, Mcgill M J. Introduction to Modern Information Retrieval[M]. New York: McGraw-Hill, 1983: 305-306

[8] Zhou Tao, Lü Linyuan, Zhang Yicheng. Predicting missing links via local information [J]. European Physical Journal B, 2009, 71(4): 623-630

[9] Adamic L A, Adar E. Friends and neighbors on the Web [J]. Social Networks, 2003, 25(3): 211-230

[10] Clauset A, Moore C, Newman M E. Hierarchical structure and the prediction of missing links in networks [J]. Nature, 2008, 453(7191): 98-101

[11] Friedman N, Getoor L, Koller D, et al. Learning probabilistic relational models [C] //Proc of the 16th Int Joint Conf on Artificial Intelligence. San Francisco: Morgan Kaufmann, 1999: 1300-1309

[12] Taskar B, Wong M F, Abbeel P, et al. Link prediction in relational data [C] //Proc of the 16th Int Conf on Neural Information Processing Systems. Cambridge: MIT Press, 2003: 659-666

[13] Neville J, Jensen D. Relational dependency networks [J]. Journal of Machine Learning Research, 2007, 8(2): 653-692

[14] Liu Zhen, Zhang Qianming, Lü Linyuan, et al. Link prediction in complex networks: A local naive bayes model [J]. Europhysics Letters, 2011, 96(4): 8007

[15] Wang Chao, Satuluri V, Parthasarathy S. Local probabilistic models for link prediction [C] //Proc of the 7th IEEE Int Conf on Data Mining. Los Alamitos, CA: IEEE Computer Society, 2007: 322-331

[16] Menon A K, Elkan C. Link prediction via matrix factorization [C] //Proc of the 2011 European Conf on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2011: 437-452

[17] Dunlavy D M, Kolda T G, Acar E. Temporal link prediction using matrix and tensor factorizations [J]. ACM Transactions on Knowledge Discovery from Data, 2011, 5 (2): 10:1-10:27

[18] Fan Xuhui, Xu Yida, Cao Longbing, et al. Learning nonparametric relational models by conjugately incorporating node information in a network [J]. IEEE Transactions on Cybernetics, 2017, 47(3): 589-599

[19] Miller K T, Griffiths T L, Jordan M I. Nonparametric latent feature models for link prediction [C] //Proc of the 22nd Int Conf on Neural Information Processing Systems. Vancouver: Curran Associates, 2009: 1276-1284

[20] Wang Xin, Wang Ying, Zuo Wanli. Exploring interactional opinions and status theory for predicting links in signed network [J]. Journal of Computer Research and Development, 2016, 53(4): 764-775 (in Chinese)
(王鑫, 王英, 左万利. 基于交互意见和地位理论的符号网络链接预测模型[J]. 计算机研究与发展, 2016, 53(4):764-775)

[21] Liu Ye, Zhu Weiheng, Pan Yan, et al. Multiple sources fusion for link prediction via low-rank and sparse matrix decomposition [J]. Journal of Computer Research and Development, 2015, 52(2): 423-436 (in Chinese)
(刘冶, 朱蔚恒, 潘炎,等. 基于低秩和稀疏矩阵分解的多源融合链接预测算法[J]. 计算机研究与发展, 2015, 52(2): 423-436)

[22] Zhang Zehua, Miao Duoqian, Qian Jin. Detecting overlapping communities with heuristic expansion mehtod based on rough neighborhood [J]. Chinese Journal of Computers, 2013, 36(10): 2078-2086 (in Chinese)
(张泽华, 苗夺谦, 钱进. 邻域粗糙化的启发式重叠社区扩张方法[J]. 计算机学报, 2013, 36(10): 2078-2086)

[23] Yao Yiyu. Interpreting concept learning in cognitive informatics and granular computing [J]. IEEE Transactions on Systems Man & Cybernetics , Part B (Cybernetics), 2009, 39(4): 855-866

[24] Yao Yiyu, Zhao Liquan. A measurement theory view on the granularity of partitions [J]. Information Sciences, 2012, 213(2012): 1-13

[25] Breiger R L. The duality of persons and groups [J]. Social Forces, 1974, 53(2):181-190

[26] Yang Jaewon, Leskovec J. Overlapping community detection at scale: A nonnegative matrix factorization approach [C] // Proc of the 6th ACM Int Conf on Web Search & Data Mining. New York: ACM, 2013: 587-596

[27] Hsieh C J, Dhillon I S. Fast coordinate descent methods with variable selection for non-negative matrix factorization [C] // Proc of the 17th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2011: 1064-1072

[28] Sanchez P I, Muller E, Laforet F, et al. Statistical selection of congruent subspaces for mining attributed graphs [C] // Proc of the 13th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2013: 647-656

[29] Snijders T A B, Pattison P E, Robins G L, et al. New specifications for exponential random graph models [J]. Sociological Methodology, 2010, 36(1): 99-153

[30] Xiong Bingyan, Wang Guoyin, Deng Weibin. Under-sampling method based on sample weight for imbalanced data [J]. Journal of Computer Research and Development, 2016, 53(11): 2613-2622 (in Chinese)
(熊冰妍, 王国胤, 邓维斌. 基于样本权重的不平衡数据欠抽样方法[J]. 计算机研究与发展, 2016, 53(11): 2613-2622)

Luo Sheng, born in 1982. PhD candidate in Tongji University. Student member of CCF. His main research interests include granular computing, machine learning.

Miao Duoqian, born in 1964. PhD, professor, PhD supervisor in Tongji University. Distinguished member of CCF. His main research interests include rough sets, granular computing and machine learning.

Zhang Zhifei, born in 1986. PhD and lecturer in Tongji University. Member of CCF. His main research interests include natural language processing and machine learning.

Zhang Yuanjian, born in 1990. PhD candidate in Tongji University. Student member of CCF. His main research interests include rough sets, granular computing and machine learning.

Hu Shengdan, born in 1982. PhD candidate in Tongji University. Student member of CCF. Her main research interests include rough sets and machine learning.