软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(3):663−675 [doi: 10.13328/j.cnki.jos.005164]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

动态信息网络中基于角色的结构演化与预测*

李 川[1], 冯冰清[1,2,3], 李艳梅[1], 胡绍林[3], 杨 宁[1], 唐常杰[1]

[1](四川大学 计算机学院,四川 成都 610065)

[2](西安卫星测控中心 厦门测控站,福建 厦门 361023)

[3](航天器故障诊断与维修重点实验室,陕西 西安 710043)

通讯作者: 杨宁, E-mail: yangning@scu.edu.cn

摘 要: 动态信息网络是当前复杂网络领域中一个极具挑战的问题,其动态的演化过程具有时序、复杂、多变的特点.结构是网络最基本的特征,也是进行网络建模和分析的基础,研究网络结构的演化过程,对全面认识复杂系统的行为倾向具有重要意义.使用角色来量化动态网络的结构,得到动态网络的角色模型,应用并改进多类标分类问题的问题转换思想,将动态网络的角色预测问题视为多目标回归问题,以历史网络数据作为训练数据构建模型,预测未来时刻网络可能的角色分布情况,提出基于多目标回归思想的动态网络角色预测方法 MTR-RP(multi-target regression based role prediction).该方法不仅克服了基于转移矩阵方法忽略时间因素的不足,还考虑了多个预测目标之间可能存在的依赖关系.实验结果表明,提出的 MTR-RP 方法具有更准确且更稳定的预测效果.

关键词: 动态信息网络;结构演化;结构预测

中图法分类号: TP311

中文引用格式: 李川,冯冰清,李艳梅,胡绍林,杨宁,唐常杰.动态信息网络中基于角色的结构演化与预测.软件学报,2017,28(3):663−675. http://www.jos.org.cn/1000-9825/5164.htm

英文引用格式: Li C, Feng BQ, Li YM, Hu SL, Yang N, Tang CJ. Role-Based structural evolution and prediction in dynamic networks. Ruan Jian Xue Bao/Journal of Software, 2017,28(3):663−675 (in Chinese). http://www.jos.org.cn/1000-9825/5164.htm

Role-Based Structural Evolution and Prediction in Dynamic Networks

LI Chuan[1], FENG Bing-Qing[1,2,3], LI Yan-Mei[1], HU Shao-Lin[3], YANG Ning[1], TANG Chang-Jie[1]

[1](School of Computer, Sichuan University, Chengdu 610065, China)

[2](Xiamen Station, China Xi'an Satellite Control Center, Xiamen 361023, China)

[3](State Key Laboratory for the Spacecraft Fault Diagnosis and Maintenance, Xi'an 710043, China)

Abstract: Dynamic information network is a new challenging problem in the field of current complex networks. The evolution of dynamic networks is temporal, complex and changeable. Structure is the basic characteristics of the network, and is also the basis of network modeling and analysis. The study of the network structure evolution is of great importance in getting a comprehensive understanding of the behavior trend of complex systems. This paper introduces "role" to quantify the structure of dynamic network and proposes a role-based model. To predict the role distributions of dynamic network nodes in future time, the presented framework views role prediction as a multi-target regression problem, extracts properties from historical snapshot sub-network, and predicts the future role distributions of dynamic network nodes. The paper then proposes a multi-target regression based role prediction (MTR-RP) method for dynamic network. This method not only overcomes the drawback of the existing methods which operate on transfer matrix while ignoring the time factor, but also takes into account of possible dependencies between multiple forecast targets. Experiments results show that MTR-RP has better and more stable prediction capability compared with the existing methods.

* 基金项目: 国家自然科学基金(61473222, 91646108)
Foundation item: National Natural Science Foundation of China (61473222, 91646108)
收稿时间: 2016-07-31; 修改时间: 2016-09-14; 采用时间: 2016-11-01; jos 在线出版时间: 2016-11-29
CNKI 网络优先出版: 2016-11-29 13:35:13, http://www.cnki.net/kcms/detail/11.2560.TP.20161129.1335.016.html

Key words: dynamic information network; structural evolution; structural prediction

复杂网络随着信息技术的发展不断被应用到各个领域,而对复杂网络的研究通常需要对其进行建模和简化.Han 等人在 EDBT 2009 和 SIGMOD 2010 会议上提出一种新的复杂网络表现形式,即信息网络(information network).信息网络属于复杂网络范畴,是对现实空间中海量、多维、复杂结构和问题更具一般性的抽象[1,2].信息网络一般都有动态的演化过程,网络中的节点和边随着时间不断改变,这里统称为动态信息网络.传统复杂网络研究多将复杂系统建模为静态网络,在揭示社交网络、生物网络等复杂系统的性质与特征等方面都已取得重要成果.而在现实世界中,几乎所有的网络结构是随时间不断变化的,社交网络[3,4]、科研合作者网络[5]、蛋白质网络[6]等都是动态信息网络的实例.动态网络是当前复杂网络领域中极具挑战的新问题,其动态的演化过程具有时序、复杂、多变的特点,蕴含着丰富的潜在信息和商业价值.

结构[7]是网络的基本特征,分析动态信息网络结构的演化过程,对全面认识复杂系统的行为倾向具有重要意义.针对网络结构的表示问题,KDD 2012 会议上的文献[8]首次提出用潜在的角色(role)来刻画节点的结构行为.角色代表网络结构的某种类型,结构类型相似的节点属于同一种角色,如中心节点、边缘节点等.动态网络的结构预测是网络演化分析的重要问题,它旨在利用历史网络信息预测未来时刻节点的拓扑结构,帮助人们提前进行预警和决策.例如:生物学家需要预测分子网络的拓扑结构何时发生变化,因为细胞分子的结构与生物体的生长状况息息相关;在商品销售网络中,挖掘未来可能成为关键节点的用户可以实现广告的精准投放;而针对通信网络,通过预测通信网络节点的重要性来优化网络资源的均衡分配,是网络管理人员的迫切需求.

动态网络的结构演化分析的目的是挖掘网络随时间演化的特点,由于网络结构本身比较复杂,难以表示和量化,动态网络时序、多变的演化过程更增加了分析的难度.目前,动态网络研究还处于起步阶段,Rossi 等人以角色为基础进行了一系列动态网络演化分析相关的研究[9−11],是本文研究工作的基础.文献[9]提出一种基于自学习方法挖掘动态网络角色的混合模型 DBMM(dynamic behavioral mixed-membership model);文献[10]在 DBMM 模型的基础上分析了网络整体角色随时间的动态变化情况,提出角色演化的几种特点.然而,以上模型并未针对动态网络的特点给出角色匹配以及角色解释的方法,因此可解释性不强.WSDM 2013 会议的文献[11]进一步将上述模型进行扩展,应用模型进行未来时刻网络角色的预测,其思想是将动态网络的多种角色视为网络的多个状态,通过计算网络在相邻时刻的状态转移矩阵来进行角色演化的分析与预测.但是,由于该方法的转移矩阵模型只由相邻两个时刻的网络数据得到,忽略了历史时间因素对角色演化的影响,因此效果并不理想.

为解决以上问题,本文使用角色来量化动态网络的结构,以得到的动态网络角色模型为基础,为网络结构预测问题提供了新思路.主要贡献如下:

(1) 提出动态网络的角色模型.使用角色来表示动态网络的结构,将静态网络的角色发现方法扩展至动态网络,用相对简单的角色序列来量化复杂的结构演化,得到动态网络角色模型;对得到的模型进行分析,给出角色匹配与角色解释的方法.

(2) 提出了基于多目标回归思想的动态网络角色预测方法 MTR-RP(multi-target regression based role prediction).以历史网络数据作为训练数据构建模型,预测未来时刻网络可能的角色分布情况.该方法不仅克服了已有基于转移矩阵方法忽略时间因素的不足,还考虑了多个预测目标之间可能存在的依赖关系.

1 动态网络的角色模型

为解决动态网络进行结构演化分析的问题,本文将角色引入动态网络场景,提出适合网络结构演化分析的动态网络角色模型.根据动态网络的演化特点,将静态网络的角色发现方法扩展至动态网络,通过将复杂的网络结构量化为相对简单的角色分布,减小直接分析网络结构演化的难度.

1.1 用角色建模动态网络

在用角色建模动态网络之前,首先要构建动态网络.动态网络的定义如下.

定义 1(动态网络). $D=\langle N,E\rangle$表示一个动态网络,$N=\langle N_1,N_2,\ldots,N_T\rangle$为节点集合,$E=\langle E_1,E_2,\ldots,E_T\rangle$为边集合.将 D 看做一个时间有序的子图序列 $D=\langle S_1,S_2,\ldots,S_T\rangle$,其中,$S_t=\langle N_t,E_t\rangle$是动态网络 D 在 t 时刻的子图快照,N_t 为 S_t 的节点集合,E_t 为 S_t 的边集合,T 为动态网络长度.本文研究网络的结构演化,故只考虑无向网络.

将动态网络表示为有序的子图序列后,可以对每个时刻的网络快照分别进行角色发现,即:将每个子网络都转化为节点的角色信息,然后根据节点在各个时刻的角色分布情况分析节点结构的演化过程.本文使用 KDD 2012 会议的 RloX 方法进[8]行角色发现,包括特征提取和角色发现.图 1 给出了对动态网络进行角色建模的基本框架.本文将节点的角色看成一种软概率分布,所有角色共同表示节点的结构,从节点的视角刻画每个节点在网络所表现出的结构身份.

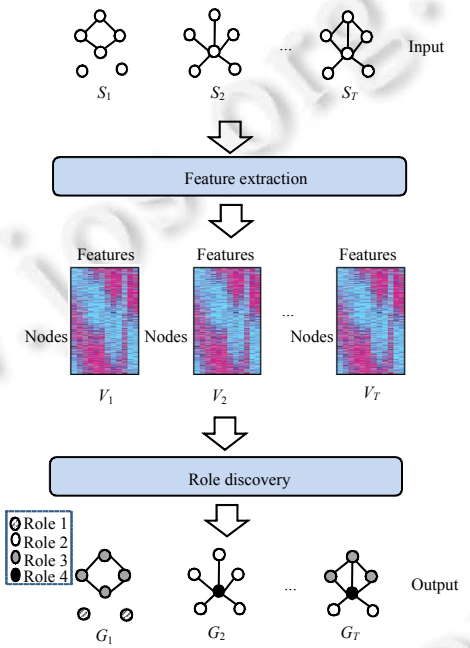


Fig.1 General process of dynamic network role modeling

图 1 动态网络角色建模的一般过程

1.1.1 角色模型的框架

(1) 特征提取

特征提取旨在尽可能多地为每个节点提取表示结构的特征,力求用一个高维的特征取值向量来保存节点完整的拓扑结构.采用 ReFex[12]的迭代特征产生方法,为每个节点提取基本特征和递归特征,基本特征指节点局部结构的特征,如节点的度、自网络包含的边数、参与三角形的个数等;得到节点的基本特征后,使用聚集函数递归地对其邻居节点的基本特征进行聚集计算得到递归特征,此过程由节点的一阶邻居开始,按照逐层扩散的方式向外围蔓延,将得到的新特征与已有的特征进行对比,若新特征与已有特征相差不大,认为不再有新特征产生并将其抛弃,终止递归[12].以图 2 的网络为例,使用度、自网络包含的边数、参与三角形的个数作为 3 个基本特征,使用 *sum* 和 *mean* 两种聚集函数来产生递归特征,为每个节点计算特征取值.如对节点 n_1,可得 n_1 的基本特征向量为$\langle f_1,f_2,f_3\rangle=\langle 5,7,2\rangle$.接着计算递归特征,直到没有新特征产生终止.

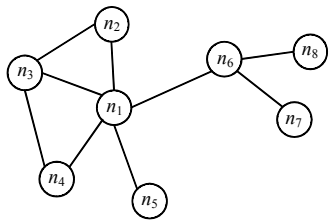


Fig.2 An example of role discovery

图 2 角色发现的示例网络

定义 2(节点-特征矩阵序列 $V=\langle V_1,V_2,\ldots,V_T\rangle$). 给定动态网络子图序列 $D=\langle S_1,S_2,\ldots,S_T\rangle$,对 S_t 进行特征提取得到节点的特征矩阵 $V_t\in\Re^{N\times f_t}$, 其中,N 为网络的节点个数,f_t 表示 t 时刻得到的特征个数.对每个网络快照 $D=\langle S_1,S_2,\ldots,S_T\rangle$分别进行特征提取,得到节点-特征矩阵序列 $V=\langle V_1,V_2,\ldots,V_T\rangle$.

(2) 角色发现

根据节点的特征矩阵进一步进行角色发现,NMF(non-negative matrix factorization)是一种解决高维非负矩阵分解的有效方法,对特征矩阵 $V_t\in\Re^{N\times f_t}$,给定一个正整数 $r<<\min(N,f_t)$,NMF 可以寻找非负矩阵 $G_t\in\Re^{N\times r}$ 及 $F\in\Re^{r\times f_t}$ 满足 $G_tF\approx V_t$,使得下面的函数值最小.

$$f(G_t,F)=\frac{1}{2}\|V_t-G_tF\|_F^2 \tag{1}$$

其中,$\|\cdot\|_F^2$ 表示矩阵 F 范数的平方,分解目标 G_t 便是节点的角色矩阵.

G_t 是一个 N 行 r 列的矩阵,每一行表示该节点在各个角色上的概率取值情况,称 G_t 为节点的角色矩阵.G_t 的每个元素 $g_t(i,j)$均为正实数,表示节点 i 属于角色 j 的概率.以中心节点、桥梁节点以及边缘节点这 3 种网络结构为例,有 r=3 的第 n 行为节点 n 在 t 时刻的角色分布,例如{R_1:0.1,R_2:0.1,R_3:0.8}.本文对 G_t 进行了行归一化处理,即:

$$\forall G_t\in G,\sum_{j=1}^{r}g_t(i,j)=1(1\leqslant i\leqslant N) \tag{2}$$

由节点的特征序列 $V=\langle V_1,V_2,\ldots,V_T\rangle$可以得到全部节点的角色序列 $G=\langle G_1,G_2,\ldots,G_T\rangle$.

1.1.2 动态网络进行角色建模算法

对动态网络进行角色建模的伪代码如下.

算法 1. 动态网络角色建模算法.

输入:动态网络 $D=\langle S_1,S_2,\ldots,S_T\rangle$;

输出:节点角色序列 $G=\langle G_1,G_2,\ldots,G_T\rangle$.

```
1.  G=∅;
2.  for (t←1 to T) do
3.      G_t=∅;   //G_t 为 t 时刻节点的角色矩阵
4.      V_t=∅;   //V_t 为 t 时刻节点的特征矩阵
5.      for (n∈N_t) do
6.          V_t(n)=V_t(n)∪ReFex_basic(n);   //ReFex_basic()提取节点基本特征,V_t(i)为特征矩阵 V_t 的第 n 行
7.      end for
8.      while (true) do
9.          for (n∈N_t) do
10.             Nei(n)=getNeighbor(n);   //计算一阶邻居节点集合
11.             f_Nei=∅;   //保存由当前一阶邻居所产生的递归特征
```

12. $f_{Nei}=f_{Nei}\cup ReFex\_recurse(Nei(n))$; //*ReFex_recurse*()产生递归特征,使用聚集函数(*sum*,*mean*)递归地对其邻居节点的基本特征进行聚集
13. **end for**
14. *count*=0;
15. **for** ($f\in f_{Nei}$) **do**
16. **if** (*ReFex_isNewFeature*(f_{Nei})) //判断是否为新特征
17. $V_t(n)=V_t(n)\cup f$; //添加新特征
18. **else** *count*++;
19. **end for**
20. **if** (*count*==f_{Nei}.*size*()) **break**;
21. **end while**
22. $G_t=NMF(V_t)$; //非负矩阵分解得到角色矩阵
23. **end for**

算法第 2 行~第 21 行是对单个网络快照进行特征提取的过程,其中,第 2 行~第 7 行得到节点的基本特征,第 8 行~第 21 行计算递归特征.第 15 行~第 19 行判断本次递归过程中所得到的特征是否与已有特征相似:若不相似则为新特征,并入特征集合;若得到的所有特征都与已有特征相似,则终止递归.第 22 行用非负矩阵分解得到节点的角色矩阵.

1.2 角色匹配

要识别相邻时刻的角色是否一致,可以将某时刻的角色与下一时刻的所有角色一一匹配.由于角色模型将每个节点在每个时刻都表示为一个角色概率的向量,向量的每一维表示节点在相应角色上的概率取值,假定概率取值最大的角色作为节点的角色,当节点本身属性比较模糊,可以将节点同时划分到相对应的两个角色集合中.通常来说,网络中大部分节点的角色在短时间内不会剧烈变化,因此,可以根据比对相邻时刻两角色节点集合的公共节点来进行匹配,匹配得多的两个角色认为是同一种角色.基于以上思路,接下来定义 t 时刻角色 $R_{t,i}$ 与 t+1 时刻角色 $R_{t+1,j}$ 的匹配程度.

$$M(R_{t,i},R_{t+1,i})=\frac{|R_{t,i}\cap R_{t+1,i}|}{|R_{t,i}\cup R_{t+1,i}|} \tag{3}$$

1.3 角色解释

得到角色序列 $G=\langle G_1,G_2,\ldots,G_T\rangle$后,节点结构表示为一系列角色的取值分布.由前面介绍可知,节点的角色可以认为是节点的某种相似性度量,能否用熟知的其他度量(如度、介数等)来从侧面对角色做出解释?

给定动态网络的子图快照 S_t 与角色矩阵 G_t,为每个节点选定 m 种熟知度量,可以得到矩阵 $M_t\in\Re^{N\times m}$,M_t 为 t 时刻节点的度量矩阵,m 为度量个数.要量化角色与度量之间的关系.再次使用非负矩阵分解方法 NMF 求解一个新的矩阵 E_t,使得 $G_tE_t\approx M_t$,其中,$E_t\in\Re^{r\times m}$.因此,E_t 的行对应 r 个角色,列对应 m 个度量,E_t 的每一行则表示该角色在各个度量上的概率取值.如此,将角色与一系列熟知度量的关系进行量化,可以形成对角色的直观感受.

2 动态网络角色预测

2.1 问题定义

动态网络的角色预测,就是要利用若干历史网络数据来预测下一时刻网络的节点角色矩阵.形式化表示为:给定动态网络 $D=\langle S_1,S_2,\ldots,S_T\rangle$,得到动态网络角色模型 $G=\langle G_1,G_2,\ldots,G_T\rangle$,$G_i$ 为 i 时刻节点角色矩阵,角色预测就是要得到 t+1 时刻网络的节点角色矩阵 G'_{t+1}.

若将动态网络角色预测问题视作回归问题,要得到 t+1 时刻网络的节点角色矩阵,就是要预测每个节点 n 在 t+1 时刻的角色分布向量 g_{t+1}(G'_{t+1} 的第 n 行).因此,对每个节点而言,预测目标是一个向量.回归问题分为模型训

练和测试两个阶段,设用于学习模型的历史快照数为 k+1 个,即,用$\langle G_{t-k},\ldots,G_{t-1},G_t\rangle$训练模型,其中,提取$\langle G_{t-k},\ldots,G_{t-1}\rangle$用于预测的属性,$G_t$ 为预测目标,采用滑动窗口,向后推移一个时刻,得到测试数据$\langle G_{t-k+1},\ldots,G_t,G_{t+1}\rangle$,其中,$\langle G_{t-k+1},\ldots,G_t\rangle$提取属性,$G'_{t+1}$ 为预测目标,如图 3 所示.k 可以视作窗口往前滑动,如此完成动态网络预测任务.

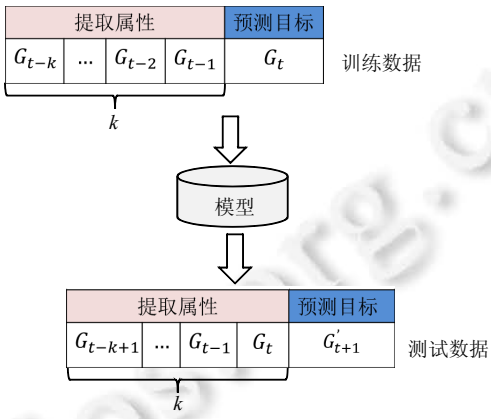


Fig.3 Model of multi-objective prediction

图 3 多目标预测模型

2.2 属性选择

将动态网络角色预测转化为回归问题,首先要为每个节点选择合适的属性.为了尽可能提高预测效果,本文不仅使用了网络的节点角色信息,还为节点计算多种度量.对时刻 i 的网络快照,为每个节点选择属性集合 $F=F_R\cup F_M$,其中:F_R 表示节点的角色属性,是节点在各个角色上的取值,由节点角色矩阵 G_i 得到;F_M 是节点的多种度量取值.本文计算 6 种常见的节点度量(度、带权度、介数、特征向量中心性、带权特征向量中心性以及 *PR* 值),角色特征为 4 个,见表 1.

Table 1 Attribute explanations

表 1 属性说明

时刻	类别	属性	说明
i	角色属性	R_1	i 时刻,节点 n 在角色 R_1 上的取值
		R_2	i 时刻,节点 n 在角色 R_2 上的取值
		R_3	i 时刻,节点 n 在角色 R_3 上的取值
		R_4	i 时刻,节点 n 在角色 R_4 上的取值
	度量属性	Degree	i 时刻,节点 n 的度
		Degree_wei	i 时刻,节点 n 的加权度
		Betweeness	i 时刻,节点 n 的介数
		Eigenvector	i 时刻,节点 n 的特征向量中心性
		Eigenvector_wei, *PR*	i 时刻,节点 n 的带权特征向量中心性;i 时刻,节点 n 的 *PageRank* 值

完成属性选择后,为每个节点得到一个 $k\times f$ 维的属性集合 $x=[x_1,\ldots,x_d]$,预测目标为节点在下一时刻的角色分布向量,记为 $y=[y_1,\ldots,y_m]$(本文中 m=4).因此,用于训练模型的 k+1 个历史网络快照可以表示为 $D=\{(x^1,y^1),\ldots,(x^N,y^N)\}$,$D$ 是训练集,其中,N 为网络节点的个数,也是 D 包含的实例条数.那么,角色预测问题就是要学习模型 $h:X\rightarrow Y$,能为给定的属性向量 x^q 预测目标向量 $y'^q=h(x^q)$,使得 y'^q 与真实值 y^q 最接近.

3 基于多目标回归的动态网络角色预测模型

3.1 预测模型

本文用问题转换思路来解决多类标分类的方法,为每个类标分别建立数据集,每个数据集只包含一个类标,

然后使用单类标分类器完成分类.在预测类标时,并没有直接用单类标分类器作为最终模型,而是把得到的所有单目标分类结果重新并入属性集合,如此更新训练数据的属性集合.本文提出解决动态网络角色预测问题的多目标回归模型 MTR-RP 包括训练和测试两个阶段(如图 4 所示),其中,训练阶段又包括了训练一阶模型和训练二阶模型两步.

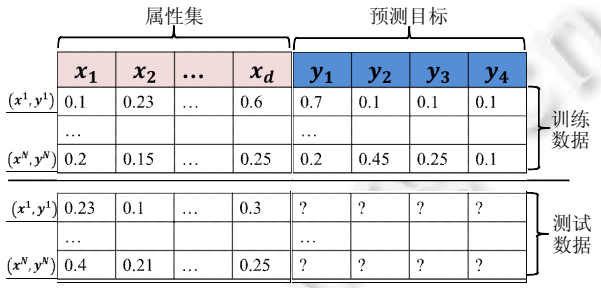


Fig.4 Prediction model

图 4 预测模型

(1) 训练一阶模型

根据问题转换思想,首先将数据集 $D=\{(x^1,y^1),\ldots,(x^N,y^N)\}$ 划分为 m 份,每份表示为 $D_j=(x^1,y_j^1),\ldots,(x^N,y_j^N)$,对 D_j 应用已有的单目标回归方法(如线性回归)可得到 $h_j:X\to R$.如此,可得到 m 个预测模型 $h_1,\ldots,h_m$.本文称这些模型为一阶模型,过程如图 5 所示.

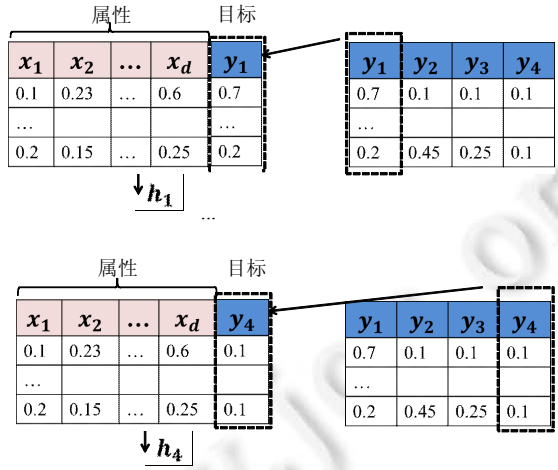


Fig.5 First-order model

图 5 产生一阶模型

虽然用单目标回归算法已经为每个目标得到了一阶模型,但本文没有直接将一阶模型作为最终模型.为方便实验对比,本文将这些由单目标回归方法得到的一阶模型记为单目标回归的角色预测方法 STR-RP (single-target regression based role prediction).

(2) 训练二阶模型

由一阶模型 $h_1,\ldots,h_m$ 可以得到各目标的一阶预测值,如图 6 所示.为了利用多个预测目标之间可能存在的依赖关系,此处将得到的一阶预测值重新并入数据集的属性集合中用于预测,因此,原始数据集 $D_j=(x^1,y_j^1),\ldots,(x^N,y_j^N)$ 更新为 $D_j^*=\{(x^{*1},y_j^1),\ldots,(x^{*N},y_j^N)\}$,其中,$x_j^{*n}=[x_1^n,\ldots,x_d^n,y_1'^n,\ldots,y_{j-1}'^n,y_j'^n,\ldots,y_m'^n]$ 为扩充后的属性集合,$y_j'^n$ 为

一阶模型 h_j 得到的一阶预测值 $y'^n_j=h_j(x^n)$,扩充后属性集合的属性个数为[d+(m−1)]个.由扩充后数据集 D^*_j 使用单目标回归方法得到新的预测模型 $h^*_j: X^* \to R$.如此,可得到 m 个更新后的模型 $h^*_1,...,h^*_m$,如图 7 所示.这里称其为二阶模型,也就是最终的预测模型.在得到一阶模型 $h_1,\ldots,h_m$ 与二阶模型 $h^*_1,...,h^*_m$ 后,使用 $G=\langle G_{t-k+1},\ldots,G_t, G_{t+1}\rangle$数据为测试集,记每个节点从时刻($t$–$k$+1)到时刻 t 得到的属性集合为 x^s,由一阶模型 $h_1,\ldots,h_m$ 得到一阶预测结果 $y'^s=[y'^s_1,y'^s_2,y'^s_3,y'^s_4]=[h_1(x^s),h_2(x^s),h_3(x^s),h_4(x^s)]$,然后利用一阶预测结果为每个目标更新属性集,对目标 h_j,有 $x^{*s}_j=[x^s_1,...,x^s_d,y'^s_1,...,y'^s_{j-1},y'^s_{j+1},...,y'^s_m]$,由此可得最终预测结果(如图 8 所示).

$$y''^s=[y''^s_1,y''^s_2,y''^s_3,y''^s_4]=h^*_1(x^{(*s)}_1),h^*_2(x^{(*s)}_2),h^*_3(x^{(*s)}_3),h^*_4(x^{(*s)}_4)].$$

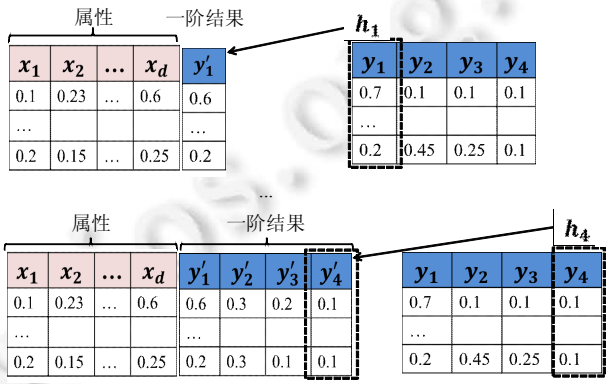


Fig.6 Results of first-order prediction model

图 6 一阶预测结果

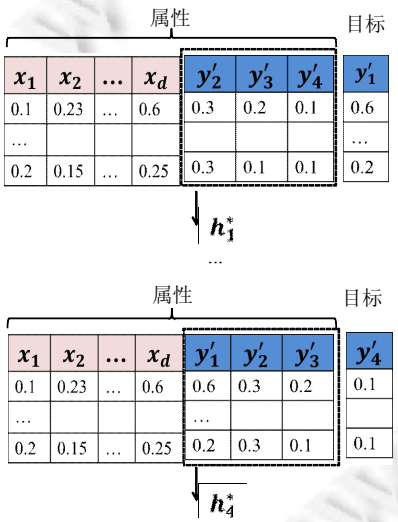


Fig.7 Second-Order model

图 7 产生二阶模型

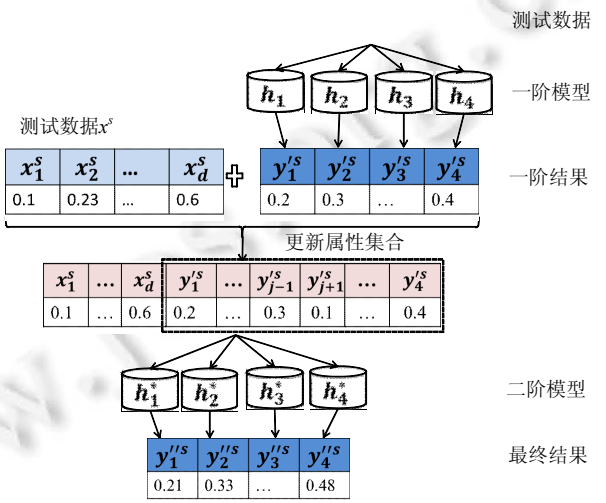


Fig.8 Testing stage

图 8 测试阶段

3.2 MTR-RP的训练过程算法

由此,MTR-RP 的训练过程算法可抽象如下.

算法 2. MTR-RP 的训练过程.

输入:数据集 $D=\{(x^1,y^1),\ldots,(x^N,y^N)\}$,其中,$x=[x_1,\ldots,x_d]$为属性集,$y=[y_1,\ldots,y_m]$为预测目标;

输出:一阶模型 $h_1,\ldots,h_m$ 与二阶模型 $h_1^*,\ldots,h_m^*$.

1. //产生一阶模型
2. **for** j=1 to m
3. $D_j=(x^1,y_j^1),\ldots,(x^N,y_j^N)$; //复制 m 个数据集
4. $h_j:D_j\rightarrow R$; //学习单目标回归模型
5. $D_j^*=\varnothing$; //初始化新数据集
6. **end for**
7. //更新数据集
8. **for** j=1 to m **do**
9. **foreach** $x_j\in D_j$ **do**
10. $x_j^*=x_j$;
11. **for** k=1 to m, $k\neq j$ **do**
12. $y_k'=h_k(x_j)$; //一阶预测结果
13. $x_j^*=[x_j^*,y_k']$; //合并一阶预测结果到属性集
14. **end for**
15. $D_j^*=D_j^*\cup x_j^*$; //添加新数据
16. **endforeach**
17. **end for**
18. //产生二阶模型
19. **for** j=1 to m **do**
20. $h_j^*=D_j^*\rightarrow R$;
21. **end for**

实验所用单目标回归为线性回归,调用自 WEKA 函数包 weka.classifiers.functions.LinearRegression.class. 训练过程的时间复杂度主要在更新数据集时的循环,取决于预测目标的个数 m、数据集实例个数(即节点个数)N,时间复杂度为 $O(Nm^2)$.预测目标一般较小,本文为 4,因此,训练过程的时间消耗主要取决于网络节点个数.

4 实验及分析

4.1 数据集

本文选取具有代表意义的 3 个动态网络数据集(Enron,Facebook,DBLP).

(1) Enron 公司邮件网络(Enron(http://konect.uni-koblenz.de/networks/enron))

Enron 公司内部员工的邮件网络.Enron 公司 2001 年宣告破产,也引来众多的研究关注.本文选取 2001 年数据用于构建动态网络,每月为一个快照间隔,共得到 12 个网络快照.节点表示员工、边为员工之间的邮件往来.

(2) Facebook 社交网络(Facebook-wall(http://konect.uni-koblenz.de/networks/facebook-wosn-wall))

Facebook-wall 数据源自社交网站 Facebook 的“用户墙”应用.本文选取的时间段为 2008 年(1 月~12 月),每月为一个快照间隔,节点表示 Facebook 用户,边表示用户间的一次留言.网络中一直存在的节点数为 5 111 个.

(3) 科研合作者网络(DBLP(http://dblp.uni-trier.de/xml/))

DBLP 是大规模的科研合作关系的数据集.作者为节点,作者共同发表一篇论文则产生一条合作关系.本文选取 2002 年~2011 年的数据构建动态网络,每年为一个快照间隔,网络节点表示作者,边表示作者之间的合作关系.在初始化该网络时,去除发表论文总数小于 10 篇论文的作者.网络中一直存在的节点数为 29 747 个.

为简化起见,本文假设网络的节点数目保持不变,因而只考虑在研究时间段内一直出现的节点,忽略中途出

现或消失的节点.对以上 3 个网络,本文均建模为无向加权网络,权重的计算使用按时间衰减的思想,即:离当前时刻越远的网络快照,对当前时刻节点权重影响越小.权重计算公式如下:

$$w_{a,b}(t)=\sum_i w_i e^{-\lambda(t-t_i)} \tag{4}$$

其中,w_i 为边 ab 在 t_i 时刻的权重,相应的邻接矩阵序列为$\langle A_1,A_2,\ldots,A_r\rangle$.

3 个数据集的详细信息见表 2.

Table 2 Datasets details

表 2 数据集详细信息

数据集	节点数	*边数	*特征数	角色数	快照数	快照长度	时间区间
Enron	2 114	16 413	70	4	12	1 month	2001.1~2001.12
Facebook	5 111	14 438	144	4	12	1 month	2008.1~2008.12
DBLP	29 747	96 874	159	4	10	1 year	2002~2011

*表示平均值

4.2 对比方法

本文使用 3 种方法作为对比.

(1) PRE(Baseline):直接用当前时刻的节点角色取值作为下一时刻的预测值,即,用 t 时刻的节点角色矩阵作为 t+1 时刻的预测结果,即, $G'_{t+1}=G_t$;

(2) TM:TM 方法是前面提到的转移矩阵方法[11],该方法利用 t–1 和 t 时刻的角色矩阵 G_{t-1} 和 G_t,根据非负矩阵分解得到角色转移矩阵 T:$G_{s(t-1)}T\approx G_{s(t)}$,由 G_t 和 T 相乘得到 t+1 时刻的目标角色矩阵.

$$G'_{t+1}:G'_{t+1}=G_tT;$$

(3) STR-RP:以基于单目标回归算法得到的一阶模型为最终模型, G'^n_{t+1} 表示节点 n 在 t+1 时刻的预测向量.

$$G'^n_{t+1}=[h_1(G^n_t),\ldots,h_4(G^n_t)].$$

4.3 评估方法

给定预测值 G'_{t+1} 和真实值 G_{t+1},本文使用 F 范数来度量两个矩阵的差异.

$$Frobenious\ Loss=\|G'_{t+1}-G_{t+1}\|_F \tag{5}$$

$\|\cdot\|_F$ 为矩阵的 F 范,对矩阵 A 计算 F 范数,显然,F 范数值越小,预测效果越好.

$$\|A\|_F\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{i,j}|^2} \tag{6}$$

4.4 实验效果

在本文模型中,参数 k 是用于提取属性集合的时刻个数,直接决定用于预测的数据,因而首先验证参数 k 对预测结果的影响.下面以 Facebook 数据集 2008 年 12 月的角色矩阵为预测目标,k 取值从 2 到 7,每次分别计算预测矩阵 G'_{t+1} 和真实矩阵 G_{t+1} 的 F 范数值.图 9 为对每次实验重复 10 遍后取平均的结果,可以看出,k 的值从 2 变化到 7,预测效果并没有随之变好,这说明用作预测的属性个数并非越多越好.事实上这也是合理的:一方面,k 取值过大带来太多的预测属性,很可能造成数据过拟合;另一方面,可以认为只有离当前时刻较近的历史数据才能有效帮助预测下一时刻网络状态,太远的历史数据可能反而不利于预测效果.当 k 取 3 时,预测效果相对理想.

下面分别在 Enron,Facebook 以及 DBLP 这 3 个数据集上验证 MTR-RP 模型的有效性.此处参数 k 均设定为 3,因而预测目标从时刻 5 开始,分别表示 Enron 数据集中 2001 年 5 月~12 月(如图 10 所示)、Facebook 数据集中 2008 年 5 月~12 月(如图 11 所示)、DBLP 数据集中 2006 年~2011 年(如图 12 所示).

对图 10~图 12 的分析包括以下几点.

(1) 总体分析.

可以看到,MTR-RP 模型在规模不同的 3 个真实数据集中都取得了很好的效果.与直接用一阶模型方法

STR-RP 相比,MTR-RP 得到更准确的预测值.这说明节点在 4 种角色上分别对应的取值是具有一定联系的.本文的节点角色取值在归一化处理后,节点在各个角色上的取值存在求和为 1 的关系.PRE 和 TM 都只使用了前一个时刻的数据做预测,在 3 个数据集上的预测效果都不如 MTR-RP,STR-RP;并且从曲线形状来看,PRE 和 TM 模型的曲线都比较波动.相比之下,MTR-RP 和 STR-RP 能够得到相对稳定的预测结果.

(2) DBLP 网络的有趣现象.

观察 PRE 曲线与其他 3 条曲线的距离可以发现一个有趣的现象:相比 Enron 和 Facebook 网络,在 DBLP 数据集中,PRE 的预测效果与其余 3 种方法更为接近.PRE 直接使用当前时刻节点的角色作为下一时刻的预测值,这是基于相邻时刻网络结构不会发生剧烈变化的假设.继续分析可知,在 Enron 和 Facebook 网络中 PRE 表现很差.这说明 Enron 和 Facebook 网络结构并不稳定.事实上,2001 年 7 月~10 月间,Enron 公司发生了巨大的人事调动,年底公司破产,Enron 网络结构理应是不稳定的;而 2008 年的 Facebook 网络正处于超速发展中,6 月正式成为全球最大、增长最快的社交网络,Facebook 的网络结构也不会是稳定的;但在 DBLP 网络中,学者一般具有稳定的研究兴趣和合作学者,网络结构自然相对稳定.

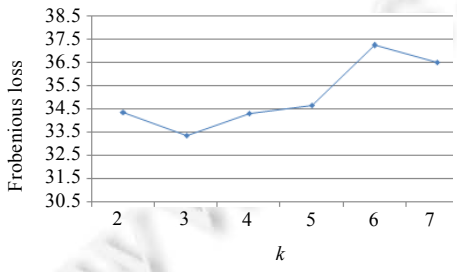


Fig.9 Influence of k to the prediction result—Facebook dataset

图 9 参数 k 对预测结果的影响——Facebook 数据集

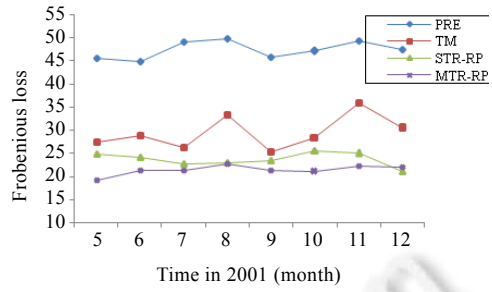


Fig.10 Predicition in Enron dataset

图 10 Enron 数据集预测效果

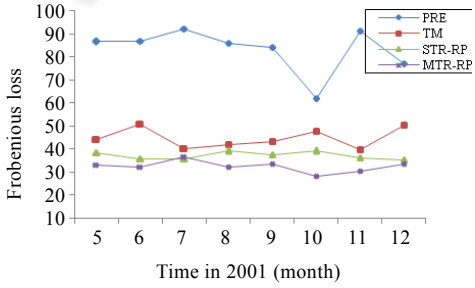


Fig.11 Predicition in Facebook dataset

图 11 Facebook 数据集预测效果

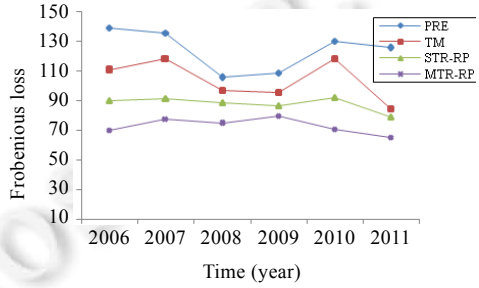


Fig.12 Predicition in DBLP dataset

图 12 DBLP 数据集预测效果

5 相关工作

目前,动态网络的结构演化分析的相关研究可以分为动态网络结构预测、动态网络的演化模式挖掘以及动态社团发现等.动态网络结构预测的相关研究主要包括网络的链接预测以及节点中心性预测两部分.链接预测主要有:Sarkar 等人[13]提出了一种基于隐藏空间模型(latent space model)的方法对动态网络的边进行预测,该方法基于马尔可夫假设,即,下一时刻节点的边至于当前时刻有关系;Liben 等人[14]研究动态网络的链接预测问题,提出了基于分类思想的网络链接预测方法;Huang 等人[15]将动态网络的链接数据视为时间序列,通过定义新的时间序列相似性度量进行链接预测.动态网络的模式挖掘主要集中于网络频繁子图模式挖掘问题的研究[16].Borgwardt 等人[17]提出了动态频繁子图的概念,提出了 Dynamic GREW 算法用于挖掘动态频繁子图,通过应用

后缀树挖掘单边动态频繁子图,迭代地合并小的动态频繁子图得到规模更大的动态频繁子图;以 Borgwardt 等人的工作为基础,Wachersreuther 等人[18]提出了 DFSD 算法挖掘动态图中的动态频繁子图,采用字符串方式将静态频繁子图编码,应用字符串的公共子串发现方法得到这些静态图共有的边出现模式,从而获得动态图中的动态频繁子图;Berlingerio 等人[19]提出一种基于频繁模式挖掘的演化规则挖掘算法 GREM,解决了动态网络的局部演化规律挖掘问题.动态网络社团挖掘方法主要分增量聚类和演化聚类两类:增量聚类是分别对每个时刻的网络快照进行聚类;演化聚类多采用时间平滑假设,在多个时刻上对聚类进行连续分析.Toyoda 等人[20]给出了挖掘社团演化的 6 种类型(出现、消失、增长、收缩、分裂和合并),为每种类型分别定义相应的度量,这是社团演化问题较早的研究之一;Hopcroft 等人[21]使用演化聚类的方法究 NEC 的文献引文网络中社团结构的演化情况,发现网络中可供跟踪的稳定存在社团.

6 总 结

本文解决了动态网络的结构表示、结构演化及角色预测等问题,通过将静态网络的角色发现扩展至动态网络,提出动态网络的角色模型,并以此为基础,将结构预测问题角色预测,提出基于多目标回归思想的动态网络角色预测算法 MTR-RP.动态网络是目前复杂网络研究领域中极具活力的新兴研究方向,相比于静态网络的研究成果,目前动态网络的研究还处于起步阶段.本文只针对其中的演化和预测问题进行研究,传统静态网络中许多问题都需要在动态网络中得到进一步研究与扩展,未来的研究工作将继续关注动态网络的演化问题.

References:

[1] Albert R, Barabasi AL. Statistical mechanics of complex networks. Reviews of Modern Physics, 2002,74(1):47−97. [doi:10.1103/RevModPhys.74.47]

[2] Wang XF, Li X, Chen GR. The Theory and Application of Complex Networks. Beijing: Tsinghua University Press, 2006 (in Chinese).

[3] Biggs N, Lloyd EK, Wilson RJ. Graph Theory 1736~1936. Oxford University Press, 1976.

[4] Fang BX. Online social network Analysis. Beijing: Electronic Industry Press, 2014 (in Chinese).

[5] Kim J, Wilhelm T. What is a complex graph? Physica A: Statistical Mechanics and its Applications, 2008,387(11):2637−2652. [doi: 10.1016/j.physa.2008.01.015]

[6] Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. Nature, 2001,411:41−42. [doi: 10.1038/35075138]

[7] Fang JQ, Wang XF, Zheng ZG, Bi Q, Di ZR, Li X. A new cross science: Network science (I). Progress in Physics, 2007,21(3):239−337 (in Chinese with English abstract).

[8] Henderson K, Gallagher B, Eliassi-Rad T, Tong HH. Rolx: Structural role extraction & mining in large graphs. In: Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2012. 1231−1239. [doi: 10.1145/2339530.2339723]

[9] Rossi R, Gallagher B, Neville J, Henderson K. Dynamic behavioral mixed-membership model for large evolving networks. arXiv preprint arXiv:1205.2056, 2012.

[10] Rossi R, Gallagher B, Neville J, Henderson K. Role-Dynamics: Fast mining of large dynamic networks. In: Proc. of the 21st Int'l Conf. on Companion on World Wide Web. ACM Press, 2012. 997−1006. [doi: 10.1145/2187980.2188234]

[11] Rossi RA, Gallagher B, Neville J, Henderson K. Modeling dynamic behavior in large evolving graphs. In: Proc. of the 6th ACM Int'l Conf. on Web Search and Data Mining. ACM Press, 2013. 667−676. [doi: 10.1145/2433396.2433479]

[12] Henderson K, Gallagher B, Li L, Akoglu L, Eliassi-Rad T, Tong HH, Faloutsos C. It's who you know: Graph mining using recursive structural features. In: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2011. 663−671. [doi: 10.1145/2020408. 2020512]

[13] Sarkar P, Moore AW. Dynamic social network analysis using latent space models. ACM Special Interest Group on Knowledge Discovery and Data Mining, 2005,7(2):2330.

[14] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology, 2007,58(7):1019−1031. [doi: 10.1002/asi.20591]

[15] Huang Z, Lin DKJ. The time-series link prediction problem with applications in communication surveillance. INFORMS Journal on Computing, 2009,21(2):286−303. [doi: 10.1287/ijoc.1080.0292]

[16] Gao L, Yang JY, Qin GM. Methods for patternmining in dynamic networks and applications. Ruan Jian Xue Bao/Journal of Software, 2013,24(9):2042−2061. http://www.jos.org.cn/1000-9825/4439.htm [doi: 10.3724/SP.J.1001.2013.04439]

[17] Borgwardt KM, Kriegel HP, Wackersreuther P. Pattern mining in frequent dynamic subgraphs. In: Proc. of the 6th IEEE Int'l Conf. on Data Mining. 2006. 818−822. [doi: 10.1109/ICDM.2006.124]

[18] Wackersreuther B, Wackersreuther P, Oswald A, Böhm C, Borgwardt KM. Frequent subgraph discovery in dynamic networks. In: Proc. of the 8th Workshop on Mining and Learning with Graphs. 2010. 155−162. [doi: 10.1145/1830252.1830272]

[19] Berlingerio M, Bonchi F, Bringmann B, Gionis A. Mining graph evolution rules. In: Proc. of the Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg: Springer-Verlag, 2009. 115−130. [doi: 10.1007/978-3-642-04180-8_25]

[20] Toyoda M, Kitsuregawa M. Extracting evolution of Web communities from a series of Web archives. In: Proc. of the 14th ACM Conf. on Hypertext and Hypermedia. ACM Press, 2003. 28−37. [doi: 10.1145/900051.900059]

[21] Hopcroft J, Khan O, Kulis B, Selman B. Tracking evolving communities in large linked networks. Proc. of the National Academy of Sciences, 2004,101(Suppl. 1):5249−5253. [doi: 10.1073/pnas.0307750100]

附中文参考文献:

[2] 汪小帆,李翔,陈关荣.复杂网络理论及其应用.北京:清华大学出版社,2006.

[4] 方滨兴.在线社交网络分析.北京:电子工业出版社,2014.

[7] 方锦清,汪小帆,郑志刚,毕桥,狄增如,李翔.一门崭新的交叉科学:网络科学(上).物理学进展,2007,21(3):239−337.

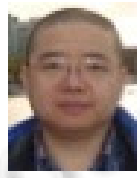
[16] 高琳,杨建业,覃桂敏.动态网络模式挖掘方法及其应用.软件学报,2013,24(9):2042−2061. http://www.jos.org.cn/1000-9825/4439.htm [doi: 10.3724/SP.J.1001.2013.04439]

李川(1977－),男,河南郑州人,博士,副教授,CCF 专业会员,主要研究领域为数据库,数据挖掘,信息网络数据分析,社会网络分析,生物信息学.

冯冰清(1990－),女,助理工程师,主要研究领域为数据库,数据挖掘,信息网络.

李艳梅(1994－),女,硕士,主要研究领域为数据库,数据挖掘,信息网络.

胡绍林(1964－),男,博士,教授,博士生导师,主要研究领域为航天安全与大数据分析技术,过程监控与故障诊断技术,复杂系统建模与仿真.



杨宁(1974－),男,博士,讲师,CCF 专业会员,主要研究领域为时空数据挖掘,时态序列挖掘,异构信息网络挖掘,网络上的信息处理.

唐常杰(1946－),男,教授,博士生导师,CCF 杰出会员,主要研究领域为数据库系统,数据挖掘和数据仓库,知识工程,计算机安全.