基于因子图模型的动态图半监督聚类算法

张建朋 [1,2] 裴雨龙 [2] 刘 聪 [2,3] 李邵梅 [1] 陈鸿昶 [1]

摘 要 针对动态图的聚类主要存在着两点不足: 首先, 现有的经典聚类算法大多从静态图分析的角度出发, 无法对真实网络图持续演化的特性进行有效建模, 亟待对动态图的聚类算法展开研究, 通过对不同时刻图快照的聚类结构进行分析进而掌握图的动态演化情况. 其次, 真实网络中可以预先获取图中部分节点的聚类标签, 如何将这些先验信息融入到动态图的聚类结构划分中, 从而向图中的未标记节点分配聚类标签也是本文需要解决的问题. 为此, 本文提出进化因子图模型 (Evolution factor graph model, EFGM) 用于解决动态图节点的半监督聚类问题, 所提 EFGM 不仅可以捕获动态图的节点属性和边邻接属性, 还可以捕获节点的时间快照信息. 本文对真实数据集进行实验验证, 实验结果表明 EFGM 算法将动态图与先验信息融合到一个统一的进化因子图框架中, 既使得聚类结果满足先验知识, 又契合动态图的整体演化规律, 有效验证了本文方法的有效性.

关键词 半监督聚类, 进化因子图模型, 特征提取, 动态图

引用格式 张建朋, 裴雨龙, 刘聪, 李邵梅, 陈鸿昶. 基于因子图模型的动态图半监督聚类算法. 自动化学报, 2020, **46**(4): 670−680

DOI 10.16383/j.aas.c170363

A Semi-supervised Clustering Algorithm Based on Factor Graph Model for Dynamic Graphs

ZHANG Jian-Peng[1,2] PEI Yu-Long[2] LIU Cong[2,3] LI Shao-Mei[1] CHEN Hong-Chang[1]

Abstract There are two main deficiencies on clustering of dynamic graphs. Firstly, most of existing classical algorithms analyze such graphs from the perspective of static analysis. However, static analysis is not capable of modeling the continuous evolution of real-world networks. Therefore, it is a great need to research on clustering algorithms for dynamic graphs. Our goal is to capture the dynamic evolution characteristics by considering the clustering structure of multiple snapshots as a whole. Secondly, some clustering labels of some nodes in real-world graphs can be obtained in advance, thus how to integrate these priori information into clustering assignments of dynamic graphs and assign clustering labels to the unlabeled nodes of each snapshot should be resolved. In this paper, we propose an evolution factor graph model (EFGM) for the semi-supervised clustering of nodes in dynamic networks. EFGM is able to capture the node-attribute and edge-adjacency attribute of each node of the dynamic graphs, and also make full use of the snapshot information. We experiment with the real-world graphs and experimental results show that the EFGM integrates the prior knowledge and the dynamic graph into a unified framework (i. e., evolution factor graph model), which makes the clustering result satisfy the prior label information and conform to the overall evolution of dynamic graphs. It validates the effectiveness of the proposed approach.

Key words Semi-supervised clustering, evolution factor graph model (EFGM), feature extraction, dynamic graphs

Citation Zhang Jian-Peng, Pei Yu-Long, Liu Cong, Li Shao-Mei, Chen Hong-Chang. A semi-supervised clustering algorithm based on factor graph model for dynamic graphs. *Acta Automatica Sinica*, 2020, **46**(4): 670−680

收稿日期 2017-07-01 录用日期 2018-05-04
Manuscript received July 1, 2017; accepted May 4, 2018
国家自然科学基金群体项目 (61521003), 国家重点研发计划项目 (2016YFB0800101) 资助
Support by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (61521003), National Key Research and Development Program of China (2016YFB0800101)
本文责任编委 朱军
Recommended by Associate Editor ZHU Jun

1. 国家数字交换系统工程技术研究中心 郑州 450002 中国 2. 埃因霍温理工大学 埃因霍温 5600MB 荷兰 3. 山东科技大学计算机学院 青岛 266590 中国
1. National Digital Switching System Engineering & Technological R&D Center, Zhengzhou 450002, China 2. Eindhoven University of Technology, Eindhoven 5600 MB, the Netherlands
3. Department of Computer Science, Shandong University of Science and Technology, Qingdao 266590, China

如今越来越多的数据可以抽象为具有关联关系的图数据 (也称网络数据), 典型例子包括网页链接形成的 Web 网络图, 用户关联交互的社交网络, 蛋白质相互作用网络. 通常, 网络中的节点可以与相应的聚类标签相关联, 并且这些标签可以具有许多形式, 例如人口统计标签、兴趣、群组等. 如何向图中的未标记节点分配标签是经典的半监督聚类问题. 然而, 随着网络规模的增加和图节点之间复杂的关系使网络图的节点标签标注成本昂贵以及难以获得

的特点. 因此, 如何给网络中未标注的节点进行半监督聚类的问题最近引起了广泛的关注.

当前动态图聚类问题主要面临以下挑战: 1) 当前研究大多数集中于静态网络[1−3]. 事实上, 许多真实网络图结构是随时间变化而不断演变的, 即图中的节点和边随时间的变化而更新. 例如, 社交网络中的用户动态添加、删除朋友. 在这种动态场景中, 时间属性信息对未标注节点的半监督聚类问题有很重要的作用. 然而现有的经典算法诸如 Girvan-Newman (GN[2])、Fast Newman (FN[3])、标签传播[4] 等诸多方法大多从静态图分析的角度出发, 无法针对真实网络持续演化的特性进行有效的建模[5]. 2) 真实网络中我们是可以预先获取图中部分节点对应的聚类标签 (例如政治、兴趣, 加入群组标签), 如何将已有的先验信息融入到动态图的聚类划分中, 并对动态图进行半监督聚类的相关研究尚未成熟[6]. 3) 如何对图节点进行有效的特征提取仍未得到很好的解决. 文献 [7] 选择 5 种类型的特征, 即同源、三角闭合、范围、嵌入和结构洞, 用于节点的分类问题. 文献 [8] 计算几个预定义的网络属性作为特征表征, 包括归一化节点度、聚类系数、共同邻居等. 然而, 这些特征提取方法存在复杂度较高, 效果差的缺点. 因此, 难以在不同的聚类任务中推广使用.

为了解决这些问题, 本文提出进化因子图模型 (Evolution factor graph model, EFGM) 用于动态图节点的半监督聚类问题, 将动态图与先验信息融合到统一的进化因子图框架中, 这样既使得聚类结果满足先验知识, 又可以契合动态图的整体演化规律. 具体而言, 动态图以一系列图快照的形式组织, 本文在进化因子图模型中基于节点特征, 节点相关性和时间相关性分别设计三种类型的因子: 包括节点因子、边邻接因子和进化因子来对图的相应快照进行建模. 其中节点因子和边邻接因子可以捕获图结构的全局和局部性质, 而进化因子可以有效地利用时间快照信息. 为了解决图中特征提取的问题, 本文在动态图中引入 Node2vec[9] 图嵌入的特征提取方法, 此特征提取过程中并不需要复杂的特征工程与外部知识就可以实现图的特征提取.

本文工作的主要贡献可归纳如下:

1) 提出进化因子图模型 (EFGM) 用于动态图节点的半监督聚类问题, 本模型通过捕获动态图的节点属性、边邻接属性和时间属性作为相应的因子应用到进化因子图模型中, 有效解决了动态图的半监督聚类问题.

2) 采用真实数据集评估 EFGM 的聚类效果, 所提模型在不同的评价指标均取得了不错的聚类效果. 此外, 本文对特征维度、训练数据大小和图特征提取方法进行了敏感性分析, 进一步验证了算法的鲁棒性.

本文的后续部分安排如下: 第 1 节对相关工作进行简要回顾; 第 2 节对问题进行形式化定义; 第 3 节对所提的进化因子图模型进行了详细阐述; 第 4 节对所提方法的有效性进行了实验验证; 第 5 节对本文所做的工作进行了总结和展望.

1 相关工作

1.1 动态图的半监督聚类

动态网络图是由不同的快照时刻 $\{t_1, t_2, \cdots, t_m\}$ 所对应的静态图快照所构成的一系列静态图快照 $G = \{G_1, G_2, \cdots, G_m\}$ 所构成, 并且每个时刻的聚类结构由此时刻快照图对应的聚类簇 $C = \{C_1, C_2, \cdots, C_m\}$ 所构成. 通常情况下, 动态图中聚类结构会随着时间的变化而发生改变, 但是相邻时刻图的演化通常较为缓慢和平滑, 一般不会发生突变. 正式地, 动态图的聚类问题定义为: 给定当前时刻 t_i 图的连接关系 A_i 与之前时刻的所有连接关系 $\{A_1, A_2, \cdots, A_{i-1}\}$, 综合两者来获取当前时刻图的准确聚类划分[10]. 动态图的组织结构是随时间的变化而不断演化的, 其聚类结构也随之变化, 其中包括聚类簇的合并和分离, 节点/边的增加与消失等. 动态图的聚类问题尚处于起步阶段, 相应的聚类模型和方法并不多见. 按照动态图的处理方式的不同, 本文将其概括为两类方法: 一类是基于进化聚类的方法[11−13], 该方法基于动态图随时间变化相对缓慢的假设, 在对每个时刻的图进行聚类分析时, 既要使聚类结果尽量符合当前时刻的拓扑结构 (静态快照质量), 又要与历史时刻的聚类结构划分尽量吻合 (历史开销); 另一类是基于增量聚类的方法[14−16], 此类方法以历史时刻图划分为基础, 仅针对增量相关的节点和边进行处理, 算法运算速度较快, 但一般都会对聚类质量造成一些牺牲, 难以有效应对聚类数目发生变化的情况. 这两类方法都没有考虑已有的先验信息进行半监督聚类, 因此本文不在赘述相关方法. 特别地, 对于半监督聚类问题, 针对静态图的研究相对较多, 而对动态图的研究才刚刚起步, 现在本文梳理出相关研究成果如下:

1) 针对静态图的半监督聚类: 文献 [17] 通过联合非负矩阵分解和半监督聚类的方法, 融合成对约束进行半监督聚类划分. 文献 [18] 采用融合部分已知的聚类标签, 并分析了先验信息对聚类结果的影响. 文献 [19] 利用了部分已知部分节点的聚类标签以及成对约束限制的两类指导信息, 提出半监督的 SpinGlass 模型. 文献 [20] 基于离散势的半监督聚类算法, 将已知部分节点聚类标签的先验信息加入

到聚类过程中. 文献 [21] 在多个聚类模型中融入聚类标签的先验信息, 提出了统一的基于隐空间相似性的半监督框架. 文献 [22] 利用成对约束信息, 提出了基于极值优化的半监督聚类方法. 然而, 此类基于静态图的半监督方法难于扩展到动态场景中.

2) 针对动态图的半监督聚类: 目前已有少量文献解决动态网络中对节点进行标注的方法[23−24]. 在文献 [23] 提出了学习潜在特征表达并捕获动态模式的模型. 然而, 该方法需要所有历史快照数据对下一快照中的节点进行标注, 而实际上在历史数据中的一些标签可能已经丢失或不正确. 文献 [24] 使用支持向量机 (Support vector machine, SVM) 对每个快照中的节点进行标注, 并将来自上次快照的支持向量和当前训练数据组合起来进行分类. 然而, 此方法极大地依赖于 SVM 的性能, 并且仅使用先前快照的支持向量也可能丢失有用的动态信息.

综上所述, 本文从提升动态图半监督聚类精度的角度出发, 提出了进化因子图模型用于动态网络中的节点半监督聚类问题, 将动态图与先验信息融合到一个统一的进化因子图框架中, 既可以使得聚类结果满足先验知识, 又可以契合动态图的整体演化规律, 具备较好的理论基础和可行性.

1.2 图的特征提取

现有的图特征提取方法是利用特定的社会语义或用户特定属性来用于特征提取. 文献 [18] 基于特定社会关系来提取特征, 例如具有共同的指导老师的学生. 这些研究中也存在相应的限制, 因为: 1) 基于特定关系的预定义特征需要相关的网络外部知识, 且在实践中将难以满足; 2) 在任意网络中, 可能难以获得用户特定信息, 例如, 用户不公开私有信息. 还有一些研究是从网络的拓扑结构中提取特征. 在文献 [4] 中, 网络的特征包括基于链接数的本地自中心网络的相关属性等. 然而这类方法需要相应领域的专家信息, 导致其难以获取. 在文献 [24] 中, 5 种类型的网络属性被用作特征, 即同源、三角闭合、范围、嵌入和结构洞, 用于节点的分类问题. 类似地, 文献 [3] 计算归一化节点度和平均度、聚类系数、局部性指标等作为节点特征. 这些特征提取方法相对复杂度较高. 因此, 难以在不同的半监督任务中推广. 因此, 本文引入了一种无监督图嵌入的方法, 即 Node2vec[9], 从动态图中提取特征, 此方法在本文模型中取得了很好的应用效果.

2 问题定义

本文基于因子图模型对动态图的半监督聚类问题进行建模. 需要注意的是, 图的聚类方法按照划分方式大体可分为两大类: 1) 非重叠聚类结构检测; 2) 重叠聚类结构检测. 本文模型最终给出每个节点对应其所属聚类标签的概率分布, 因此能够给出节点隶属于各个聚类簇的概率化结果. 然而, 本文的侧重于给出非重叠聚类结构, 因此本文选取聚类标签概率最大的标签作为节点的聚类标签.

表 1 相关符号说明

Table 1 Description of symbols

符号	说明
G_L	部分标注网络
V_L	被标注的节点
V_U	未被标注的节点
E	图中的边集合
W	W_{ij} 为节点 V_i 的第 j_{th} 个属性值
f	映射函数, 将每个节点 i 赋予相应的标签, 记为 f_i
Ω	部分标注动态网络

首先, 本文介绍了相关概念和形式化描述, 部分符号描述见表 1, 并给出动态网络半监督聚类问题的定义.

定义 1. 部分标注网络: 给定有限非空标注集 $\mathbf{R}$, 部分标注网络被定义为 $G_L = \{V_L, V_U, E, W, f\}$, 其中:

V_L 是被标注的节点, V_U 是未被标注的节点, 使其满足 $V_L \cup V_U = V$;

E 是图中的边集合, 即 $E \subseteq V \times V$;

W 是节点 V 的特征矩阵, 其中每行对应相应节点 v, 每列对应相应的属性, W_{ij} 为节点 V_i 的第 j_{th} 个属性值;

f 是映射函数, 将每个节点赋予相应的标签, 即: $f : V \rightarrow \mathbf{R}$.

文中, 为了简明起见, 本文假设边是无向的, 且节点的总数是固定的, 其边结构可以随时间改变而改变. 相应的节点在不同的快照中被部分标记, 并且此部分标记的动态网络被定义为此问题输入. 更确切地说,

定义 2. 部分标注动态网络: 令 V 是有限节点集合, 部分标注动态网络 $\Omega = \{G^t | \ t = 1, \cdots, T\}$ 包括一系列的图快照 $G^t = \{V_L^t, V_U^t, E^t, W^t, f^t\}$, 其中每个快照 G^t 是部分标注网络.

基于以上定义, 本文可以定义动态网络中节点的半监督聚类问题. 给定一个部分标记的动态网络 Ω, 目标是推断网络中所有未被标记节点的聚类归属.

定义 3. 动态网络半监督聚类: 部分标注动态网络 $\Omega = \{G^t | t = 1, \cdots, T\}$, 动态网络的半监督聚类

的目标是学习预测函数 $f: V \to \mathbf{R}$, 使得所有未标注的节点得到相应的聚类归属, 即 $f(V_U) \to \mathbf{R}$.

3　基于因子图模型的动态图半监督聚类算法

3.1　因子图模型介绍

概率图模型 (Probabilistic graphical model, PGM[25]) 已广泛用于模拟网络图中实体之间的依赖关系. 它为有效地推理和学习图模型提供了可能. 概率图模型刻画了模型的随机变量的依赖关系, 有效的给出了图论相关问题的概率模型结构和推断方法. 如今, 概率图模型的推断和学习已广泛应用于人工智能、用户推荐、自然语言处理等研究领域, 研究热度和应用范围仍在继续增长. 因子图模型 (Factor graphical model, FGM[26−27]) 作为 PGM 的典型代表, 可以灵活地与特定任务集成, 并显式地表示出构造概率分布的因子, 因此特别适合在变量与因子之间传递消息的推理算法. 由此在本文中, 本文首次提出进化因子图模型 (EFGM), 它是传统因子图模型的扩展, 用于动态网络中节点的半监督聚类问题.

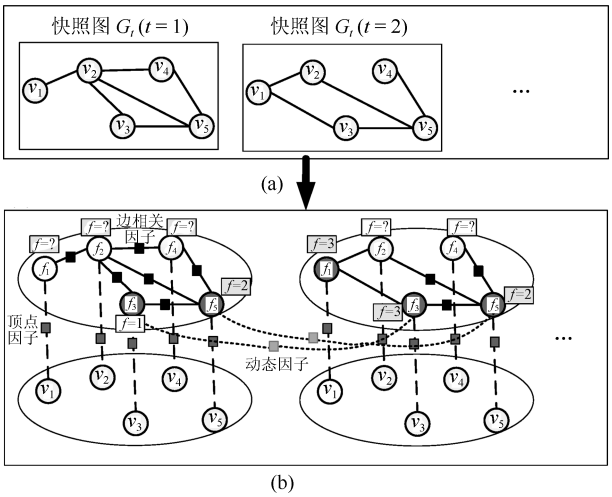


图 1　进化因子图模型示意图

Fig. 1　Diagram of the evolution factor graph model

3.2　进化因子图模型建立

在本节中, 本文考虑融合先验信息到动态图的聚类划分中, 将动态图与先验信息融合到一个统一的进化因子图框架中. 直观地, 动态图中的节点的聚类标签由三个特征属性来确定, 包括节点属性、边邻接属性、动态进化属性.

为了更清晰地理解本文模型, 本文接下来给出进化因子图模型的一个简单例子, 在图 1 中, 包含了 5 个节点 $\{v_1, v_2, v_3, v_4, v_5\}$ 和对应的连接关系. 图 1(b) 是由图 1(a) 的动态图所建立的进化因子图模型, 观测变量是网络中给定的节点 $\{v_1, v_2, v_3, v_4, v_5\}$, 模型的隐变量即每个节点的聚类标签 $\{f_1, f_2, f_3, f_4, f_5\}$, 假设整个网络中的用户划分为三个聚类簇, 即聚类标签为 $f \in \{1, 2, 3\}$, 图中 G_t 在 $t = 1$ 已知节点 $\{v_3, v_5\}$ 的聚类标签, 分别为 $\{f_3 = 1, f_5 = 2\}$, 类似地, G_t 在 $t = 2$ 已知节点 $\{v_1, v_3, v_5\}$ 的聚类标签, 此时相应的标签已经演化为 $\{f_1 = 3, f_3 = 3, f_5 = 2\}$, 本文的目标是需要预测各个快照图中其余未标注节点的聚类标签. 本文对模型的相关元素进行详细阐述:

3.2.1　变量节点说明

因子图模型中包含了两类变量节点, 分别为: 隐变量和观测变量.

隐变量节点: 这些变量是不能直接观测到的, 在图中用 f 标记的圆圈表示, 其中深色代表本文已知先验的聚类标签. 模型中包含一个隐变量的集合 $\{f_1, f_2, f_3, \cdots, f_N\}$, 其中, N 是节点的总数, 本章的任务是预测未被标记节点的标签.

观测变量节点: 在图中用 v 标记的圆圈表示, 模型中包含一个观测变量的集合 $\{v_1, v_2, v_3, \cdots, v_N\}$.

3.2.2　节点的相应属性

1) 节点属性: 即节点在当前时间戳下的全局和局部特性. 该因子对应于从网络提取的节点特征 (例如节点的出/入链路的数目), 或者是节点的固有特性 (例如用户的简档), 特征提取好坏也制约着模型效果的好坏.

2) 边邻接属性: 即在当前时间快照下节点之间的交互关系. 这个因子对应于网络中节点的连接关系. 例如, 在 Facebook 中, 用户关注 Trump 和 Clinton 可能被标记为政治爱好者. 如果作者与 Jure 和 Christos Faloutsos 合作, 有很高概率他同样也是研究数据挖掘领域的学者.

3) 动态进化属性: 本文假设节点的标签不会发生突变. 这个因子可以捕获网络的动态信息. 例如, 在 KDD 会议上发表过论文的研究人员很可能继续在数据挖掘领域发表论文.

3.2.3　进化因子图模型

基于上述介绍, 本文提出了进化因子图模型 (EFGM), 它由三个因子组成, 分别称为节点因子, 边邻接因子和进化因子, 它们分别对应于节点属性, 边邻接属性和进化属性. 详细地, 这三个因素定义如下.

1) 节点因子 $v(f_i, w_i)$: 这个因子表示为给定节点 v_i 的特征向量 w_i, 节点 v_i 标签 f_i 的后验概率分布.

2) 边邻接因子 $e(f_i, N(f_i))$: 这个因子反映了节点 v_i 与邻居中有标签节点的邻接关系, 其中, $N(f_i)$

是节点 v_i 的邻居中有标签节点的标签集合.

3) 进化因子 $d(f_i^t, f_i^{t-1})$: 这个因子定义为在两个连续的快照中, 节点 v_i 的标签的变化情况.

3.3 目标函数

基于以上所述, 给定动态图 $\Omega = \{G^t|t = 1, \cdots, T\}$, 本文定义标签 F 联合分布概率为:

$$\mathrm{P}(F|\Omega) = \prod_t \prod_i v(f_i, w_i) e(f_i, N(f_i)) d(f_i^t, f_i^{t-1}) \tag{1}$$

需要注意的是, 这些因子都可以以不同的方式进行实例化, 为了简化模型的学习, 根据 Hammersley-Clifford 定理, 本文选择了吉布斯分布的指数线性函数[28] 来实例化这三个参数. 详细来说,

1) 节点因子定义:

$$v(f_i, w_i) = \frac{1}{Z_1} \exp\{a^{\mathrm{T}} \zeta(f_i, w_i)\} \tag{2}$$

其中, a 是加权向量, $\zeta(f_i, w_i)$ 是 v_i 的特征函数的向量 (本文采用 Node2vec 求得), Z_1 是归一化因子, 也称为划分函数 (Partition function), 其形式为:

$$Z_1 = \sum_W \exp\{a^{\mathrm{T}} \zeta(f_i, w_i)\} \tag{3}$$

2) 边邻接因子定义:

$$e(f_i, N(f_i)) = \frac{1}{Z_2} \exp\{\beta^{\mathrm{T}} \delta(f_i, f_j)\} \tag{4}$$

其中类似地, Z_2 是归一化因子, β 是加权向量, $\delta(f_i, f_j)$ 是节点之间是否有连接的指示函数, 且被定义为:

$$\delta(f_i, f_j) = \begin{cases} 0, & \text{若 } e_{ij} \notin E \\ 1, & \text{否则} \end{cases} \tag{5}$$

3) 进化因子定义:

$$d(f_i^t, f_i^{t-1}) = \frac{1}{Z_3} \exp\{\eta^{\mathrm{T}} \delta(f_i^t, f_i^{t-1})\} \tag{6}$$

其中, Z_3 是归一化因子, η 是加权向量, $\delta(f_i^t, f_i^{t-1})$ 是进化属性的指示函数, 且被定义为:

$$\delta(f_i^t, f_i^{t-1}) = \begin{cases} 0, & \text{若 } f_i^t \neq f_i^{t-1} \\ 1, & \text{否则} \end{cases} \tag{7}$$

3.4 参数估计

为了简单起见, 本文给出等式 (1) 中定义的联合概率:

$$\begin{aligned} \mathrm{P}(F|\Omega) = & \frac{1}{Z} \prod_t \prod_i \exp\{\lambda^{\mathrm{T}} s_i^t\} = \\ & \frac{1}{Z} \exp\{\lambda^{\mathrm{T}} \sum_t \sum_i s_i^t\} = \\ & \frac{1}{Z} \exp\{\lambda^{\mathrm{T}} S\} \end{aligned} \tag{8}$$

其中, $Z = Z_1 Z_2 Z_3$ 是整体的归一化因子, $\lambda = (\alpha, \beta, \eta)$ 是加权向量, $S = (\zeta(f_i, w_i)^{\mathrm{T}}, \delta(f_i, f_j)^{\mathrm{T}}, \delta(f_i^t, f_i^{t-1})^{\mathrm{T}})$ 是连接相应的因子函数, 因此, 模型学习的目标就是要估计所有参数 λ. 然而, 要计算归一化因子 Z 是非常棘手的, 因为它是对所有节点的所有的可能状态的似然, 而部分节点的标签是未知的. 为了解决这个问题, 本文使用已经标注的节点来推断未知节点, 特别地, 本文需要从已知标签中推断预测未被标注节点的聚类标签. 同时, 相应的似然目标函数定义为: 已知部分标注动态网络 $\Omega = \{G^t|t = 1, \cdots, T\}$, 使得所观测到的已知节点的标签的似然 $\log p(R^L|\Omega)$ 达到最大, 正式地:

$$\begin{aligned} \mathrm{P}(\lambda) = & \log p(F^L|\Omega) = \\ & \log \sum_{F|F^L} \frac{1}{Z} \exp\{\lambda^{\mathrm{T}} S\} = \\ & \log \sum_{F|F^L} \exp\{\lambda^{\mathrm{T}} S\} - \log Z = \\ & \log \sum_{F|F^L} \exp\{\lambda^{\mathrm{T}} S\} - \log \sum_F \exp\{\lambda^{\mathrm{T}} S\} \end{aligned} \tag{9}$$

其中, Z 是整体的归一化因子, 即: $\sum_F \exp\{\lambda^{\mathrm{T}} S\}$. 通过采用 log 最大似然估计方法, 得到 log 似然目标函数 $\log p(F^L|\Omega)$, 为了解决此无约束的最优问题, 梯度下降法是最简单的选择. 为了使用梯度下降法, $\mathrm{P}(\lambda)$ 相对于 λ 的梯度计算如下:

$$\begin{aligned} \nabla_\lambda = & \frac{\delta \mathrm{P}(\lambda)}{\delta \lambda} = \\ & \frac{\sum\limits_{F|F^L} \exp\{\lambda^{\mathrm{T}} S\} S}{\sum\limits_{F|F^L} \exp\{\lambda^{\mathrm{T}} S\}} - \frac{\sum\limits_F \exp\{\lambda^{\mathrm{T}} S\} S}{\sum\limits_F \exp\{\lambda^{\mathrm{T}} S\}} = \\ & \mathrm{E}_{\mathrm{P}(F|F^L, \Omega)}[S] - \mathrm{E}_{\mathrm{P}(F|, \Omega)}[S] \end{aligned} \tag{10}$$

其中, $\mathrm{E}[S]$ 表示 S 的期望值. 由此可见, 此剃度值的计算需要在两个概率分布 $\mathrm{P}(F|F^L, \Omega)$ 和 $\mathrm{P}(F, \Omega)$ 下的期望值, 其中, $\mathrm{P}(F, \Omega)$ 是仅仅给定用户观测变量的情况下, 包括动态图中所有节点的聚类划分的

条件概率分布. 类似地, $\mathrm{P}(F|F^L,\Omega)$ 是给定节点观测变量和已知的相应节点聚类标签的情况下, 动态图聚类划分的条件概率分布. 然后, 参数 λ 使用梯度下降法被迭代更新:

$$\lambda_{\text{new}} = \lambda_{\text{old}} - \varepsilon\nabla_{\lambda} \tag{11}$$

其中, ε 为学习率和 ∇_{λ} 是相应的梯度值. 然而对于梯度的求解, 由于 EFGM 中的图结构可以是任意的且可包含循环, 使得直接计算期望值较困难. 不同的近似算法可以应用到梯度的求解中, 本文使用迭代信念传播算法[29] (Loopy belief propagation, LBP) 用以求解近似梯度值. 在每次迭代过程中, 算法利用和积 (Sum-product) 方法计算节点的边缘概率 $P(f_i|F^L,\Omega)$ 和 $P(f_i|\Omega)$, 又使用最大乘积 (Max-product) 计算整个网络中隐变量的当前最佳配置 F. 学习算法总结在算法 1 中.

算法 1.　进化因子图模型学习 $\boldsymbol{L_i}$

输入. 部分标注动态网络: Ω
　　快照图节点的 Node2vec 特征向量: W
　　学习速率: ε

输出. 学习得到的参数:$\lambda = (\alpha,\beta,\eta)$

While 最终结果未收敛, 则执行:
　计算 $\mathrm{E}_{\mathrm{P}(F|F^L,\Omega)}[S]$ 利用 LBP 梯度下降;
　计算 $\mathrm{E}_{\mathrm{P}(F|,\Omega)}[S]$ 利用 LBP 梯度下降;
　计算梯度更新 ∇_{λ};
　计算当前的梯度值 λ_{new};
End while

3.5　模型推理

模型根据所学习得到的参数 $\lambda = (\alpha,\beta,\eta)$, 进一步推断动态图中每个未知聚类标签的节点的聚类标签, 具体方法是, 给定目标函数, 然后寻找使目标函数最大化的一个聚类标签的设置, 即:

$$F_{\max} = \arg\max_{F|F^L} \mathrm{P}(F|F^L,\Omega) \tag{12}$$

本文再次使用 LBP 算法, 通过此算法计算出未知聚类标签节点的标签, 从而获得动态图的最终聚类划分.

4　图节点的特征提取

在本文中, 本文引入了一种无监督图嵌入[30−33] 的方法, 即 Node2vec[9], 从网络中提取特征. Node2vec 抽取图中节点的特征转化成最优化一个似然最大化的目标函数问题, 这个 "似然最大化" 是最大化其可观测到的邻居节点的信息. 此方法有效的避免了基于广度优先搜索 (Breath-first search, BFS) 和深度优先搜索 (Depth-first search, DFS) 来采样节点的周边节点所带来的相关问题. 简单来说, BFS 仅能够探究图中的结构性质, 而 DFS 则能够探究出内容上的相似性. 其中结构相似性不一定要相连接, 甚至可能相距很远. 而 Node2vec 方法创新的利用了一种有监督的随机游走 (Random walk) 方法可以有效的综合广度优先搜索 (BFS) 和深度优先搜索 (DFS) 的优势. 每次节点转移到下一个节点的概率取决于邻接节点的连接关系. 此有监督的游走方式拥有两个参数 p 与 q, 来控制反复经过节点的概率及所谓的向内型节点 (Inward) 和向外型节点 (Outward). 总之, 整个随机游走的目的就是在 DFS 和 BFS 之间采取某种平衡. 本文应用其作为本文的图特征提取方法, 并在本文模型中取得了很好的应用效果.

5　实验评估

在本节中, 为了验证 EFGM 的性能, 本文在真实网络数据集上进行实验验证和评估. 此外, 对模型中的参数进行敏感性分析, 包括特征维度大小的影响、标注训练数据规模的影响以及不同特征提取方法的影响. 实验硬件配置如下: Linux CentOS 操作系统, 内存 16 GB, 处理器 2.3 GHz, 实验编程语言: Python 2.7.

5.1　真实网络图实验分析

1) DBLP 会议论文网络. 本文提取 DBLP 数据集的子集. 其中包括了 6 个研究领域的会议集, 即人工智能和机器学习 (Artificial intelligence & machine learning, AI&ML)、算法和理论 (Algorithms & theory, AL&TH)、数据库 (Databases, DB)、数据挖掘 (Data mining, DM)、计算机视觉 (Computer vision, CV)、信息检索 (Information retrieval, IR). 具体来说, 我们提取 2001 年至 2010 年会议中的合著作者关系, 每年的数据都以图快照的形式组织. 表 2 为各个聚类簇中所包含的会议名称; 表 3 给出相应的作者和合著关系的统计信息.

表 2　DBLP 数据集的会议名称和聚类簇标签

Table 2　Conference names and their clustering labels of DBLP dataset

聚类簇标签	会议名称
AI & ML	IJCAI, AAAI, ICML, UAI, AISTATS
AL & TH	FOCS, STOC, SODA, COLT
CV	CVPR, ICCV, ECCV, BMVC
DB	EDBT, ICDE, PODS, SIGMOD, VLDB
DM	KDD, SDM, ICDM, PAKDD
IR	SIGIR, ECIR

2) HEPCitation 引文网络: 其数据来源于 arXiv.org, 涵盖的论文入库 (arXiv 数据库) 时间从 1993 年 1 月到 2003 年 4 月. 边的构建方式如下: 如果一篇论文 i 引用了另一篇论文 j, 在两者之间加上一条从 i 指向 j 的有向边. 对最初 1993 ~ 1997 年数据进行处理以获取初始聚类结构. 本文利用 1998 年 2 月 ~ 2003 年 2 月的连接关系, 将其按照每两个月的时间间隔构成 30 个图快照. 此数据集中论文所发表的研究领域设置为其所属聚类标签.

表 3 DBLP 会议论文网络的统计信息

Table 3 Statistics of DBLP conference network

年份	作者关系	关系数量
2001	3 074	5 743
2002	2 557	5 343
2003	3 836	7 700
2004	3 464	7 132
2005	5 198	11 171
2006	4 494	9 392
2007	7 294	15 708
2008	5 780	12 398
2009	6 405	14 321
2010	5 757	12 738

5.2 实验设置

为全面分析比较所提方法的性能, 实验采用以下三种类型的基线方法:

1) 基于因子图模型方法 (FGM)

为了验证进化属性的有效性, 我们比较了 EFGM 与两种类型的 FGM, 即 FFGM (仅使用特征属性) 和 CFGM (仅使用特征和相关属性). 此类方法都选用 Node2vec 方法提取特征向量.

2) 基于链接的半监督方法: (Majority voting, MV) 的多数表决法[34]. 多数表决法在实践中是最为普遍运用的投票规则. 这意味着未标注节点的聚类标签是由其邻居节点中出现频率最高的聚类标签决定的. 具体而言, 如果在先前快照中标记了节点标签, 则从先前的快照中复制相应的标签; 否则, 节点通过当前快照中相邻节点标签的多数表决来标记.

3) 基于特征的分类方法: 使用多分类支持向量机 (SVM) 作为分类基线[24]. SVM 从研究线性二元分类最优分类面发展而来. 最优分类面不但要求分类线能够将两类明显分开, 且使分类间隔达到最大. SVM 考虑寻找一个满足分类要求的超平面, 并且使训练集中的点距离分类面尽可能的远. 多分类 SVM 则是主要是通过组合多个二元分类器来实现多分类器的构造, 最常见的方法有一对一法 (One-versus-one SVMs) 和一对多法 (One-versus-rest SVMs) 两类方法. a) 一对一 SVM 是在任意两类样本点之间设计一个 SVM, 因此 k 个类别的样本点就需要设计 $k(k-1)/2$ 个 SVM. 当对一个未知样本点进行分类时, 最后得票最多的类别即为该未知样本点的类别. 这种方法简单易行, 聚类效果较好. b) 一对多法则是训练时依次把某个类别的样本点归为一类, 其他剩余的样本点归为另一类, 这样 k 个类别的样本点就构造出了 k 个 SVM. 分类时将未知样本点分类为具有最大分类函数值的那一类, 这个方法的不足之处在于识别过程中容易产生属于多类别的点和不能被分类的点. 因此本文采用 Scikit-learn 库的 1-v-1 方法构建多类 SVM 分类器, 其默认设置为高斯核, 参数惩罚因子 $C = 1.0$, 核参数 $\gamma = 0.2$.

5.3 聚类评价指标

本文实验中首先选取了归一化互信息指标 (Normalized mutual information, NMI)[5] 来评测聚类结果好坏. NMI 值越大表明算法所得到的聚类簇结构与真实的聚类簇结构越相似.

其次, 为了更好地评估性能, 特别是对于论文属于多个研究领域的情况, 本文采用了另一个评估指标, 即概率误差. 其优点是: 1) 它可以匹配 EFGM 的概率化输出, 此输出表示节点属于各个聚类簇的概率; 2) 它可以概率化评估聚类标签的好坏. 相对概率误差 (Relative error, RE) 定义为

$$\mathrm{RE} = \frac{1}{N \times C} \sum_{i=1}^{N} \sum_{j=1}^{C} \frac{|p_{ij} - \widetilde{p_{ij}}|}{\max(p_{ij}, \widetilde{p_{ij}})}, \quad p_{ij} \neq \widetilde{p_{ij}} \tag{13}$$

其中, N 是图中的未标记节点的数目, C 是总的聚类簇数目, 其取值范围是 [0, 1]. p_{ij} 和 $\widetilde{p_{ij}}$ 分别是节点 i 的聚类标签 j 的预测概率和真实概率. 节点 i 的聚类标签 j 的真实概率是节点 i 在研究领域 j 中的论文数量与节点 i 的论文的总数量的比值.

再次, 本文对算法的运行时间进行实验对比. 本文算法和对比算法都运行在 CentOS 服务器上且都基于 Python 2.7 语言进行算法实现, 因此尽可能降低了系统平台和编程语言导致的差异性带来的影响.

5.4 实验结果和分析

本文利用真实数据集进行对比试验. 为了与分类算法进行比较, 本文将图中节点划分成有标签的训练集和隐去标签的训练集进行实验. 本次实验则使用 70 %的数据作为有标签的训练集, 30 %作为未标注节点的测试集. 同时采用 10 折交叉验证的方法, 得到各个算法的聚类指标的平均值. EFGM 算法和不同基线方法的聚类性能如表 4 所示, 从结果

中可以得出一些结论:

1) EFGM 算法的两个评估指标优于其他方法, 表明提出模型的有效性和动态信息的重要性; 为此, 本实验对比了 FFGM 和 CFGM 算法, 本文发现如果删除节点相关性信息, 即在 FFGM 中, 聚类效果还不及 SVM 的聚类效果, 但是基于边连接信息的 CFGM 的聚类效果要明显好于 FFGM, 表明连接信息对聚类结果有着重要的影响.

2) MV 多数表决法的聚类效果好于仅使用特征的 SVM 算法, 但是聚类效果不及同时利用节点和连接信息结合的 EFGM 方法.

表 4　真实网络图的实验结果比较

Table 4　Comparison of results on real-world graphs

网络集	相应算法	NMI	RE
HEPCitation	EFGM	**0.845**	**0.203**
	FFGM	0.393	0.478
	CFGM	0.824	0.245
	MV	0.578	0.450
	SVM	0.502	0.423
DBLP	EFGM	**0.885**	**0.196**
	FFGM	0.493	0.280
	CFGM	0.814	0.235
	MV	0.678	0.350
	SVM	0.560	0.323

3) 由此可见, 对于动态图的半监督聚类问题, 节点属性、边邻接属性和时间属性对聚类结果都起到着非常重要的作用, 本算法能很好地融合已有标签信息并利用三个属性进行动态建模, 实验结果表明本文算法的有效性和鲁棒性.

5.5　处理时间对比

本节针对各个算法在这两个真实数据集上的平均运行时间进行统计, 各个算法 5 次运行的平均处理时间如表 5 所示.

表 5　各个算法在真实网络数据集上的处理时间比较 (秒)

Table 5　Comparison of the execution time on a real-world networks (s)

运行时间 (s)	EFGM	FFGM	CFGM	MV	SVM
HEPCitation	282.8	269.2	272.6	220.3	394.4
DBLP	123.8	110.3	108.2	84	232.3

从实验结果可以看出, 多数表决法 (MV) 的效率最高, 这是因为 MV 算法仅单独考虑每个快照图中各个节点的邻居信息, 如果在先前快照中标记了节点, 则从先前的快照复制相应的标签; 否则, 节点通过当前快照中相邻节点标签的多数表决来标记, 因此计算效率最高. 所提 EFGM 算法效率好于 SVM 分类器, 这是因为 One-versus-one SVMs 需要在任意两类样本点之间设计一个 SVM, 因此 k 个类别的样本点就需要设计 $k(k-1)/2$ 个 SVM. 这种方法计算量大、效率较低. FFGM (仅使用特征属性) 和 CFGM (仅使用特征和相关属性) 仅考虑部分因子信息, 因此算法效率要高于 EFGM 模型.

5.6　参数分析

在本实验中, 本文在 DBLP 会议论文合作网络中分析了在 EFGM 中参数的影响, 包括对特征维度、训练数据大小和图特征提取方法进行了敏感性分析.

1) 特征维度的影响: 为了分析特征维度对节点标签标注的影响, 通过将 Node2vec 生成的从 30～80 的特征维度以间隔 10 设置进行实验, 结果如图 2 和图 3 所示. 本文观察到各个算法在特征维度为 40 时均取得了相对较高的 NMI 值和较低的相对概率误差, 随后 NMI 的值随着特征维度的升高而下降, 而概率误差逐渐增大. 因此, 对于 DBLP 会议论文合作网络数据集, 特征维度为 40 将是这个数据集的相对较好的选择. 这是因为 Node2vec 将图节点的特征提取问题转化成最优化一个似然最大化目标函数的问题, 这个 "似然最大化" 是最大化其可观测到的邻居节点信息. 其获取的节点特征既能表征节点的内容相似性, 也能探索节点之间的结构相似性. 当节点特征维数过大或过小时, Node2vec 不能有效地表征节点的结构相似性和结构相似性信息.

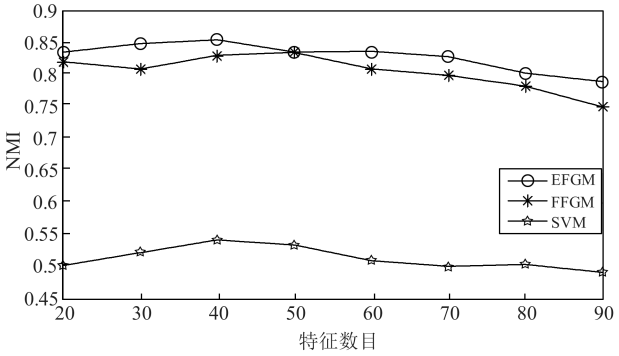


图 2　特征数目对 NMI 指标的影响

Fig. 2　The impact of the number of features on NMI score

2) 训练数据大小的影响: 在本组实验中将已知标注数据的规模设置为从 10 %～80 %, 来测试对算法性能的影响, 实验结果如图 4 和图 5 所示. 总体来说, 当给出更多的训练数据时, 则可获得更好的聚类结果. 这是因为训练数据越多, 已知的聚类标签信息能够训练出更加契合数据的模型, 使得模型更

加准确. 结果还表明 EFGM 在不同规模的训练数据具有较强的鲁棒性. 此外, 值得注意的是, 当训练数据规模相对较小时 (例如, 训练集规模小于 20 %), EFGM 的聚类效果不佳, 这是因为训练数据过少, 节点的边邻接属性和动态时间属性将会受到影响, 从而进一步影响到 EFGM 的聚类性能. 而当训练数据超过 0.4 以后, EFGM 聚类表现最好, 验证了本文算法的有效性.

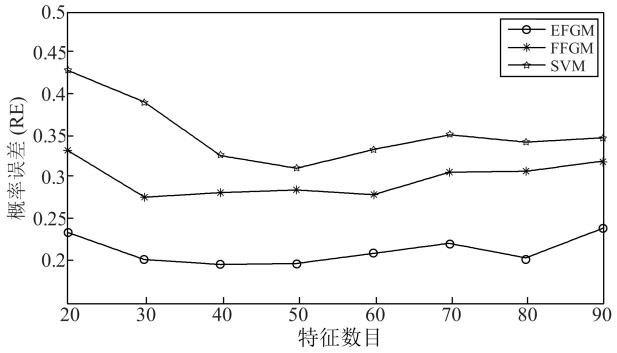


图 3 特征数目对概率误差的影响

Fig. 3 The impact of the number of features on RE score

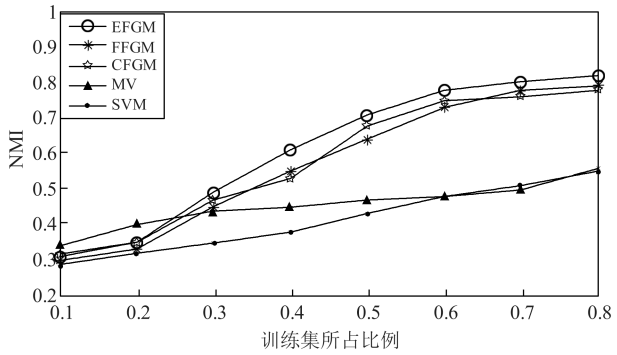


图 4 训练集所占比例对 NMI 指标的影响

Fig. 4 The impact of the training set percentage on NMI score

3) 不同特征提取方法比较: 本文实验比较了本文选用的 Node2vec 的提取特征方法和一种广泛使用的图特征提取方法 ReFeX 的特征提取效果, 实验结果如表 6 所示. 从实验结果中可以看出, Node2vec 提取特征相对于 ReFeX 更加准确, 更能有效地表征动态图节点的特征属性和局部结构信息, 这是因为 Node2vec 改进了随机游走的策略, 同时考虑到节点局部和宏观的连接信息, 具有很高的适应性. 验证了无监督图特征提取方法能够有效应用于节点的半监督聚类问题, 且不需要相应的特征工程和外部知识.

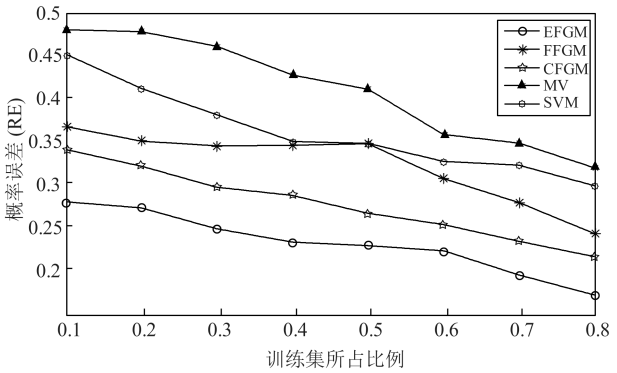


图 5 训练集所占比例对概率误差的影响

Fig. 5 The impact of the training set percentage on RE score

表 6 不同特征提取方法的实验结果比较

Table 6 Comparison of results on different feature extraction methods

特征提取方法		NMI	RE
ReFeX	EFGM	**0.837**	**0.222**
	FFGM	0.372	0.427
	CFGM	0.799	0.253
Node2vec	EFGM	**0.852**	**0.193**
	FFGM	0.402	0.392
	CFGM	0.819	0.235

6 结论

本文提出了一个进化因子图模型 (EFGM), 通过在动态图中融合已有的先验信息对动态图中的节点进行半监督聚类分析. 该模型通过捕获动态图的节点属性、边邻接属性和时间属性作为相应的因子应用到进化因子图模型中. 此外, 为了突破图特征提取的限制, 本文引入了 Node2vec 图嵌入方法从图中提取特征. 在真实网络数据集上进行实验, 所提模型可以有效地提高动态图聚类的精度. 同时本文对特征维度、训练数据大小和图特征提取方法所产生的影响进行了分析, 进一步验证了算法的鲁棒性.

未来, 本文将研究如何更好的图提取特征方法. 由于用户生成内容 (UGC) 易于访问, 因此动态场景中如何整合 UGC 信息对节点的标注非常重要. 此外, 随着网络规模的急剧增加, 研究针对大规模网络的高效聚类算法具有重要意义.

References

1 Huang Li-Wei, Li Cai-Ping, Zhang Hai-Su, Liu Yu-Chao, Li De-Yi, Liu Yan-Bo. A semi-supervised community detection method based on factor graph model. *Acta Automatica Sinica*, 2016, **42**(10): 1520−1531
(黄立威, 李彩萍, 张海粟, 刘玉超, 李德毅, 刘艳博. 一种基于因

子图模型的半监督社区发现方法. 自动化学报, 2016, **42**(10): 1520−1531)

2 Girvan M, Newman M E J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2002, **99**(12): 7821−7826

3 Newman M E J. Fast algorithm for detecting community structure in networks. *Physical Review E*, 2004, **69**(6): Article No. 066133

4 Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 2007, **76**(3): Article No. 036106

5 Chakrabarti D, Kumar R, Tomkins A. Evolutionary clustering. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, PA, USA: ACM, 2006. 554−560

6 Pan S R, Zhu X Q, Zhang C Q, Yu P S. Graph stream classification using labeled and unlabeled graphs. In: Proceedings of the IEEE 29th International Conference on Data Engineering (ICDE). Brisbane, QLD, Australia: IEEE, 2013. 398−409

7 Zhao Y C, Wang G, Yu P S, Liu S B, Zhang S. Inferring social roles and statuses in social networks. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA: ACM, 2013. 695−703

8 Choobdar S, Ribeiro P, Parthasarathy S, Silva F. Dynamic inference of social roles in information cascades. *Data Mining and Knowledge Discovery*, 2015, **29**(5): 1152−1177

9 Grover A, Leskovec J. Node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: ACM, 2016. 855−864

10 Chang Zhen-Chao. Research on Key Technologies of Community Detection in Online Social Networks [Ph. D. dissertation], Information Engineering University, China, 2016
(常振超. 在线社会网络社团检测关键技术研究 [博士学位论文], 解放军信息工程大学, 中国, 2016)

11 Lin Y R, Chi Y, Zhu S, Sundaram H, Tseng B L. Analyzing communities and their evolutions in dynamic social networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2009, **3**(2): Article No. 8

12 Chi Y, Song X D, Zhou D Y, Hino K, Tseng B L. Evolutionary spectral clustering by incorporating temporal smoothness. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, California, USA: ACM, 2007. 153−162

13 Dinh T N, Nguyen N P, Thai M T. An adaptive approximation algorithm for community detection in dynamic scale-free networks. In: Proceedings of the 2013 IEEE International Conference on Computer Communications. Turin, Italy: IEEE, 2013. 55−59

14 Sun J M, Faloutsos C, Papadimitriou S, Yu P S. Graphscope: parameter-free mining of large time-evolving graphs. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, California, USA: ACM, 2007. 687−696

15 Xiao Jie-Bin, Zhang Shao-Wu. An algorithm of integrating random walk and increment correlative vertexes for mining community of dynamic networks. *Journal of Electronics & Information Technology*, 2013, **35**(4): 977−981
(肖杰斌, 张绍武. 基于随机游走和增量相关节点的动态网络社团挖掘算法. 电子与信息学报, 2013, **35**(4): 977−981)

16 Ning H Z, Xu W, Chi Y, Gong Y H, Huang T S. Incremental spectral clustering by efficiently updating the eigen-system. *Pattern Recognition*, 2010, **43**(1): 113−127

17 Ma X K, Gao L, Yong X R, Fu L D. Semi-supervised clustering algorithm for community structure detection in complex networks. *Physica A: Statistical Mechanics and its Applications*, 2010, **389**(1): 187−197

18 Allahverdyan A E, Ver Steeg G, Galstyan A. Community detection with and without prior information. *EPL (Europhysics Letters)*, 2010, **90**(1): Article No. 18002

19 Eaton E, Mansbach R. A spin-glass model for semi-supervised community detection. In: Proceedings of the 26th AAAI Conference on Artificial Intelligence. Toronto, Ontario, Canada: AAAI, 2012. 900−906

20 Liu D, Liu X, Wang W J, Bai H Y. Semi-supervised community detection based on discrete potential theory. *Physica A: Statistical Mechanics and its Applications*, 2014, **416**: 173−182

21 Yang L, Cao X C, Jin D, Wang X, Meng D. A unified semi-supervised community detection framework using latent space graph regularization. *IEEE Transactions on Cybernetics*, 2015, **45**(10): 2585−2598

22 Li L, Du M, Liu G, Hu X G, Wu G Q. Extremal optimization-based semi-supervised algorithm with conflict pairwise constraints for community detection. In: Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Beijing, China: IEEE, 2014. 180−187

23 Li K, Guo S X, Du N, Gao J, Zhang A D. Learning, analyzing and predicting object roles on dynamic networks. In: Proceedings of IEEE 13th International Conference on Data Mining (ICDM). Dallas, TX, USA: IEEE, 2013. 428−437

24 Yao Y B, Holder L. Scalable SVM-based classification in dynamic graphs. In: Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM). Shenzhen, China: IEEE, 2014. 650−659

25 Koller D, Friedman N. *Probabilistic Graphical Models: Principles and Techniques*. Cambridge: MIT Press, 2009.

26 Tang W B, Zhuang H L, Tang J. Learning to infer social ties in large networks. In: Proceeding of the 2011 Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg: Springer, 2011. 381−397

27 Yang Z, Tang J, Li J Z, Yang W J. Social community analysis via a factor graph model. *IEEE Intelligent Systems*, 2011, **26**(3): 58−65

28 Xu H, Yang Y J, Wang L W, Liu W H. Node classification in social network via a factor graph model. In: Proceedings of the 2013 Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin, Heidelberg: Springer, 2013. 213−224

29 Murphy K P, Weiss Y, Jordan M I. Loopy belief propagation for approximate inference: An empirical study. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. Stockholm, Sweden: Morgan Kaufmann Publishers Inc., 1999. 467−475

30 Mao Q, Wang L, Tsang I W, Sun Y J. Principal graph and structure learning based on reversed graph embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(11): 2227−2241

31 Chen S H, Niu S F, Akoglu L, Kovačević J, Faloutsos C. Fast, Warped Graph Embedding: Unifying Framework and One-Click Algorithm. arXiv preprint arXiv: 1702.05764, 2017.

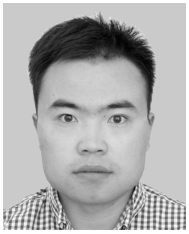
32 Shijia E, Jia S B, Xiang Y, Ji Z L. Knowledge graph embedding for link prediction and triplet classification. In: Proceedings of the 2016 Knowledge Graph and Semantic Computing: Semantic, Knowledge, and Linked Big Data. Singapore: Springer, 2016. 228−232

33 Hu W M, Gao J, Xing J L, Zhang C, Maybank S. Semi-supervised tensor-based graph embedding learning and its application to visual discriminant tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(1): 172−188

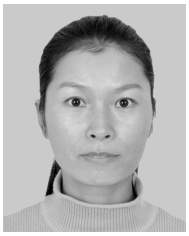
34 Lueckenga J, Engel D, Green R. Weighted vote algorithm combination technique for anomaly based smart grid intrusion detection systems. In: Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN). Vancouver, BC, Canada: IEEE, 2016. 2738−2742

张建朋 荷兰埃因霍温理工大学博士研究生, 中国国家数字交换系统工程技术研究中心助理研究员. 主要研究方向为数据流挖掘.
E-mail: zhangjianpeng0309@gmail.com
(**ZHANG Jian-Peng** Ph. D. candidate at Eindhoven University of Technology, Netherlands and lecturer at the National Digital Switching System Engineering & Technological R&D Center, China. His main research interest covers data stream mining.)

裴玉龙 荷兰埃因霍温理工大学博士生. 主要研究方向为数据挖掘, 机器学习. 本文通信作者.
E-mail: feilong0309@sina.com
(**PEI Yu-Long** Ph. D. candidate at Eindhoven University of Technology, Netherlands. His research interest covers data mining and machine learning. Corresponding author of this paper.)



刘 聪 荷兰埃因霍温理工大学博士生. 主要研究方向为流程挖掘, 软件工程.
E-mail: liucongchina@163.com
(**LIU Cong** Ph. D. candidate at Eindhoven University of Technology, Netherlands. His research interest covers process mining and software engineering.)



李邵梅 中国国家数字交换系统工程技术研究中心副研究员. 主要研究方向为图像处理与模式识别.
E-mail: lishaomei@ndsc.com.cn
(**LI Shao-Mei** Associate professor at the National Digital Switching System Engineering & Technological R&D Center, China. Her research interest covers image process and pattern recognition.)

陈鸿昶 中国国家数字交换系统工程技术研究中心教授. 主要研究方向为电信网信息关防.
E-mail: chenhongchang@ndsc.com.cn
(**CHEN Hong-Chang** Professor at the National Digital Switching System Engineering & Technological R&D Center, China. His main research interest is telecommunication network protection.)