计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(1):20−27 [doi: 10.15888/j.cnki.csa.006157]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

局部协同选择聚类的多视角社区发现研究①

于　悦, 卢　罡, 郭俊霞

(北京化工大学 信息科学与工程学院, 北京 100029)

摘　要: 近年来, 随着各种网络应用平台愈演愈烈, 多种关系网络中用户之间往往存在大量相似的局部社区结构. 鉴于传统单视角社区发现算法在划分时无法同时考虑多种因素, 本文将在多视角原理上提出一种基于局部协同选择聚类的多视角社区发现模型, 该模型中主要解决了传统多视角聚类算法的条件限制问题 (节点, 聚类个数和充分的属性信息) 和过度调整问题. 首先, 构建选择调节矩阵来训练各视角中的共同部分节点集, 并集成其共同节点的社团结构, 然后, 构建局部优化矩阵将共同节点结构做为训练集, 利用核岭回归 (KRR) 原理完成各视角中孤立节点的划分, 最后通过 UCI 数据集和 DBLP 数据集来分别验证聚类精度和算法适用性.

关键词: 多关系网络; 社区发现; 多视角聚类; 局部协同选择

引用格式: 于悦,卢罡,郭俊霞.局部协同选择聚类的多视角社区发现研究.计算机系统应用,2018,27(1):20–27. http://www.c-s-a.org.cn/1003-3254/6157.html

Research on Multi-View Community Detection Based on Local Co-Selecting Clustering

YU Yue, LU Gang, GUO Jun-Xia

(School of Information Science and Engineering, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract: In recent year, with the development of various network platforms, there are always a lot of similar local community structures between users in different networks. In consideration of some single-view community detection algorithms cannot find the multi-factor community structures, in this paper we present a Multi-view Local collaborative Selecting Clustering model (called co-MLSC). This model can solve many constraints problems (like nodes, clusters, and sufficient information) and over adjustment problems. Firstly, the model can build a choice regulate matrix that can train the common part of the node set, and converge its common structure. Then we also build a local optimization matrix that regards the node structure as a training set, and uses the KRR algorithm to complete the division of isolated nodes. Finally, we use the UCI and DBLP data sets to demonstrate the effectiveness and applicability of our algorithm.

Key words: multi-relational network; community detection; multi-view clustering; local selecting clustering

1　引言

如今, 随着大量网络应用平台的产生, 用户会根据自身的不同特点 (社会角色) 来选择不同的应用平台并与他人进行交互和信息资源的共享, 所以, 同一个用户可以在多种平台上处于不同的关系网络结构中, 这样的网络结构也被称为多维关系网络, 每个维度代表一种不同的关系视角[1], 例如, 不同学术期刊论文可以构建不同的作者协作网络, 在某些期刊中两个同社团的作者 (研究领域相同), 在其他期刊中也认为应该在同一个社团; 又如, 两个用户在某些社交平台是好友关系, 在其他社交平台中则认为这两个用户是好友的可能性很高, 所以利用不同关系网络的信息来发觉某个应用中隐藏的社团结构并进行有利的信息整合是现代信息挖掘的主要方向.

① 基金项目: 国家自然科学基金 (61602026); 国家基础科研项目 (JCKY2016212C005)
收稿时间: 2017-04-09; 修改时间: 2017-04-26; 采用时间: 2017-05-10; csa 在线出版时间: 2017-11-14

在传统的单视角社区发现算法中, 我们可以将人物关系整合到图的拓扑结构中, 用户代表图中的节点, 边表示用户间的交互关系和关联属性, 社区发现的过程被看作是图聚类分析的过程, 因此, 基于图划分的单视角聚类算法被相继提出, 如 K 均值 (K-means) 算法[2]、谱聚类算法 (Spectral Clustering) 算法[3]、规范割集准则 (Ncut) 算法[4]、对称非负矩阵因式分解 (symNMF) 算法[5,6]等. 这些算法有明显的两个特点[7]: (1) 切割式聚类, 即节点被聚类到没有交集的聚簇中; (2) 简单的图形构造, 即两个节点之间最多只有一条代表关系的连边. 然而在很多实际应用中, 表示社区结构的图数据往往来自于不同的视角 (领域), 单视角的划分和构造难以综合考虑多种因素. 例如, 在现实生活中, 企业需要寻求某个研究领域中的专家. 根据论文作者之间的引用及合作关系可以得到社团划分结果为 R1=({A, B, C, D, G, K}, {L, M, N, O, P, Q}, {P}); 同时, 根据项目合作和协助关系的社团划分结果为 R2=({A, C, D, E, F, P}, {L, M, N, O, R}, {Q}). 可以看出, 第一个视角不能将 P 进行有效的划分, 同理, 第二个视角不能将 Q 进行有效划分, 所以, 这种划分结果考虑的用户关系较为单一, 很难有效解决这种多维关系网络的问题.

针对以上这种同源异构的数据, 近年来基于多视角聚类方法比较受关注, 文献[8,9]的算法都是图切割原理结合不同的视角关系来定义的, 它们根据切割聚簇内节点代价大、切割聚间节点代价小的原理, 提出相应的目标函数公式, 将离散问题转换为连续问题, 对其进行优化, 求出函数收敛后所对应的聚类结果. 这些算法的关键在于收敛条件的确定, 而极值点和最值点不易确定. 文献[10]的 CCA 算法, 是用其它视角的聚类结果来更新本视角下节点的初始状态, 用联合训练的方法, 通过对拉普拉斯矩阵不断的更新迭代来改善聚类精度. 文献[11]的 CSC 算法是对 CCA 的改进, 利用拉普拉斯矩阵的特征向量对相似矩阵进行迭代更新, 使矩阵中同属一个聚簇的节点权值相近, 簇间节点的权值不同, 最后串联个视角的特征向量进行融合, 并用 K-means 算法聚类结果. 文献[12]的 SCSC 算法针对 CSC 算法进行改进, 利用选择投票的方式, 对强视角 (节点信息完全) 和弱视角 (显示部分节点信息) 做不同的选择处理, 最终实现多视角聚类. 但该处理方法较简单, 主要针对多稀疏实例集的稀疏关系聚类, 精度提升较差. 文献[12,13]都是使用正则化方法, 使用基于拉普拉斯特征向量来调节缺失函数, 这样由每个拉普拉斯矩阵得出的聚簇结果在所有的视角中是一致的. 这些算法假设前提为不同视角下的节点分布是相同的, 假设条件比较苛刻. 文献[14]提出的 CGC 算法利用节点之间的关联矩阵关系, 对视角中不同的实例集进行一对多和多对多的处理, 实现最后的聚类, 然而该算法中选取关联关系困难, 没有明确的公式定义.

传统多视角聚类都有两个限制: (1) 每个视角下的节点关系不同, 但每个视角要求使用相同的节点, 且每个视角中的节点聚簇个数要求相同. (2) 每个视角基本上必须拥有图结构中的充分划分信息. 以上限制条件, 导致这些方法往往不能反映真实的网络结构, 所以在实际应用中的适应性比较差.

本文针对传统的单视角社区发现算法的局限性和现有的多关系社区发现算法的不足提出一个协同选择聚类的多视角社区发现算法, 并从多个方面对算法进行改进, 本文主要贡献点包括以下 3 个方面.

(1) 使用局部协同训练方法解决大多数多视角聚类算法中实例和聚簇条件的限制, 允许不同视角中选择不同数量的节点, 每个视角的划分数量也可以不同.

(2) 解决传统多视角算法中其他视角的决定性修正而导致聚类不准确的问题. 本文只对各个视角中可识别节点 (在其他视角中有关系) 的聚类结果进行强促进处理和弱促进处理, 迭代更新每个视角中识别节点关系的相似矩阵, 促进各视角中部分结构的融合.

(3) 将局部学习和联合训练相结合, 利用每个视角中的不可识别节点 (在其他视角中无关系) 的邻居节点划分情况, 来得到非标记节点的最终归属.

本文的结构分为以下几个章节: 第 1 节是对现有社区发现的介绍和新算法的提出; 第 2 节提出多视角局部协同选择聚类算法 (co-MLSC) 并介绍其原理; 第 3 节介绍该算法中的核心公式; 最后, 通过实验, 验证算法的有效性.

2 局部协同选择多视角聚类算法

2.1 谱聚类

谱聚类[3](CS) 方法是利用拉普拉斯矩阵特征向量的性质来挖掘社团方法, 其定义为: 设 $G=(V(G), E(G))$ 是网络图结构, 其中节点集为 $V(G)=\{v_1, v_2, ...v_m\}$, 边集

为 E(G). 谱聚类的算法流程如算法 1 所示.

算法1. 谱聚类(CS)

输入: 相似矩阵$S \in R^{m \times m}$, 聚类个数$k \in N$

输出: 该相似矩阵所代表的点的聚类结果C

1. 计算S的对角矩阵$D_{(i,i)} = \sum_{j=1}^{m} S_{(i,j)}$
2. 计算拉普拉斯矩阵$L = D - S$
3. 计算L的前k个特征向量, $U = LapEig(L,k)$
4. 正规化U矩阵的每一行
5. 样例j属于聚簇c当且仅当通过K-means算法得到U的j行属于聚簇c
6. 计算并返回结果矩阵$C = Cls(U,k)$

聚类结果矩阵 C 表示聚类结果矩阵, 其中第 i 行 j 列的元素表示视角中实体 i 与实体 j 是否在同一个聚簇中, 在则为 1, 反之则为 0.

2.2 co-MLSC 算法

本节设计一种多视角协同选择的聚类算法, 该算法是基于多视角谱聚类算法的基础上提出的. 针对多视角聚类算法应用到实际网络关系中的限制问题, 该算法将每个视角中的实例分为两种, 当某个视角中的两个节点在其他视角中存在并属于同一社团结构, 则这两个节点在该视角中为可识别节点, 反之, 如果某一个节点在其他视角中不存在或存在但没有对应的关系, 则节点在该视角中为不可识别节点. 所以, co-MLSC 对实例的处理也分为两种方法, 其中包括多视角中可识别节点的强弱关系促进方法和各个单视角中不可识别节点的划分方法, 其流程如算法 2 所示.

首先, 根据实体集$x_{all} = \{x_1, x_2, ..., x_n\}$的不同属性, 我们可以使用像径向基核函数的方法得到初始相似度矩阵$K_{all} = \{K_1, K_2, ...K_n\}$, 这里 d 表示视角个数, 当每个视角下实例关系不同或者数量不相等时, 其构建的相似矩阵维数也不同. 然后通过谱聚类方法 (如算法 1), 对各视角下的实例关系进行初始划分, 得到初始的聚类结果矩阵$C_{all} = \{C_1, C_2, ..., C_d\}$, 在不同属性关系中, 我们将可识别节点和不可识别节点进行分别处理, 先将视角$\pi(\pi \in d)$ 与其他视角 $j(j \neq \pi \& j \in d)$ 结果矩阵中可识别节点的聚类结果进行集成处理, 可得到该视角的调节矩阵 R_π(3.1 节中介绍其构建公式), 并通过对原始相似矩阵进行更新得到修改后的相似矩阵 S_π, 即更新 π 视角下节点间联系的权重值. 再利用核岭回归估计的原理, 将不可识别节点的邻居节点集做为训练集, 来估计不可识别节点的最终划分, 并得到划分后的局部优化矩阵 L_π(3.2 节中介绍构建方法), 该构建过程也可以看成可识别节点利用多个视角协同集成后的聚类结构来确定单视角 π 中不可识别节点的划分. 然后, 通过每个视角中聚类后的节点集来更新各视角中的聚类中心. 最后, 不断迭代该过程直到所有视角中的聚类中心不变.

算法2. 协同式多视角选择聚类算法(co-MLSC)

输入: 各个view的相似度矩阵$K_{all} = \{K_1, K_2, ..., K_d\}$, 各个view中聚类个数$\{k_1, k_2, ..., k_d\}$.

输出: 最终聚类结果$C_{all} = \{C_1, C_2, ..., C_d\}$.

1. **Initialize:** $K_{all} = \{K_1, K_2, ..., K_d\}$, $\{k_1, k_2, ..., k_d\}$.
2. 用谱聚类CS算法得到各视角初始聚类结果矩阵$C_{all}^0 = \{C_j^0\}, j \in d$,
3. **for** i = 1 to *iter* **do** //iter表示迭代次数
4. **for** $\pi = 0$ to $d-1$ **do**
5. $R_\pi^i = co\_SO(C_{all}^{i-1}, \pi)$; //构建选择调节矩阵
6. $L_\pi^i = KRR(C_\pi^{i-1}, R_\pi^i)$; //构建局部优化矩阵
7. $S_\pi = R_\pi^i \circ K_\pi$; //修正相似矩阵
8. 对更新后的相似矩阵S使用谱聚类算法;
9. $C_\pi^i = \{C_\pi^i\}_{\pi \in d} \circ \tilde{L}_\pi^i \oplus \tilde{L}_\pi^i$//修正结果矩阵
10. 更新π视角聚类中心;
11. **end for**
12. **end for**
13. return C_{all};

3 核心矩阵的构建

这里我们提出针对两种实例的处理方法, 其中包括选择调节矩阵的构建和局部优化矩阵的构建, 选择调节矩阵主要是利用多个视角中可识别节点协同训练的作用, 对可识别节点间相似度权值进行强促进或弱促进处理, 其目的在于集成各视角中由可识别节点组成的部分社团结构. 局部优化矩阵则是对各视角中不可识别节点进行聚类, 这里利用机器学习中的核岭回归分析方法, 将每个不可识别节点的邻居节点社团划分结果做为训练集, 以估计不可识别节点最终的社团归属并对最终社团结构进行完善.

3.1 选择调节矩阵

在这里我们要定义可识别节点对视角 π 进行更新的调节矩阵, 利用共识矩阵 (co-association) 原理, 计算各视角中可识别节点的同簇划分概率, 通过概率大小值来调节视角 π 的原始相似度矩阵, 构建方法如公式 (1) 所示:

$$co\_SO(C_{all}, \pi) = \exp\left(\frac{sel\_Cl_\pi + \sum_{\pi'=0, \pi' \neq \pi}^{d} co\_Cl_{\pi,\pi'}}{d}\right) \quad (1)$$

公式 (1) 中, 不同于传统的共识矩阵构建方法, 为了解决了其他视角的过度调整问题, 使其计算同簇划分概率中分子的比重不同, 如果两个同社区的实例在

其他多个视角中也在同一个社区, 则我们将这种促进作用放大; 如果两个非同社区的实例在其他多个视角中属于同一个社区, 虽同样有促进作用, 其分子影响比重较小. 所以, 上式中可以认为选择调节矩阵是促进关系结构的过程, 不存在其他视角的抑制作用而导致的过度性调整, 只有当两个实例在所有视角中都属于同一个社团结构时, 式中的同簇划分概率为 1, 影响值 2.7(e^1), 属于强促进, 其他多种情况均为弱促进.

$$co\_Cl_{\pi,\pi'(i,j)} = \begin{cases} 1 & \text{if } C_{\pi(i,j)} = C_{\pi'(i,j)} = 1 \\ 0 & \text{else} \end{cases} \tag{2}$$

公式 (2) 表示 π 视角与其他视角可识别节点关系聚类情况的整合. 例如, 节点 i 和节点 j 在 π 视角中同一个社区, 并且在其他 $h(h<d)$ 个视角中也在同一个社区, 那么节点 i 和节点 j 在同一个社区的概率会变大, 促进作用会大幅度增强.

$$sel\_Cl_{\pi(i,j)} = \begin{cases} 1 & \sum_{\pi'=0,\pi'\neq\pi}^{d} C_{\pi'(i,j)} \geq d/2 \\ 0 & \text{else} \end{cases} \tag{3}$$

公式 (3) 表示其他视角中聚类效果明显的实例对, 对视角 π 是有促进作用的. 如果节点 i 和节点 j 在其他大多数视角中属于同一个社区结构, 那么节点 i 和节点 j 在同一个社区的概率也会变大, 促进作用会小幅度增强.

下面用 3 个实例的关系来展示该更新过程. 假设我们得到 3 个视角关系的聚类结果矩阵, 如图 1 所示. 其中 3 个视角的实例集聚类结果分别为{(*a*, *b*)(*c*)(*d*, *e*)}, {(*a*, *b*, *d*)(*c*)(*m*)(*e*, *t*)}, {(*a*, *b*, *c*)(*m*)(*d*, *e*)}.

$$C_1=\begin{bmatrix} & a & b & c & d & e \\ a & 1 & 1 & 0 & 0 & 0 \\ b & 1 & 1 & 0 & 0 & 0 \\ c & 0 & 0 & 1 & 0 & 0 \\ d & 0 & 0 & 0 & 1 & 1 \\ e & 0 & 0 & 0 & 1 & 1 \end{bmatrix} C_2=\begin{bmatrix} & a & b & c & d & e & m & t \\ a & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ b & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ c & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ d & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ e & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ m & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ t & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} C_3=\begin{bmatrix} & a & b & c & d & e & m \\ a & 1 & 1 & 1 & 0 & 0 & 0 \\ b & 1 & 1 & 1 & 0 & 0 & 0 \\ c & 1 & 1 & 1 & 0 & 0 & 0 \\ d & 0 & 0 & 0 & 1 & 1 & 0 \\ e & 0 & 0 & 0 & 1 & 1 & 0 \\ m & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

图 1 3 个视角的聚类结果矩阵

图 2 可以看出, 由于 3 个视角的共同影响, 实例 A 与实例 B 在同一个社区的可能性比较大, 在 3 个视角中都是强促进关系. 在第 1 个视角中, 尽管实例 *c* 和实例 *d* 没有与实例 *a*, *b* 划分到一个簇中, 但在其他视角的作用下对其相似度关系会起到弱促进的作用, 同理 *e* 和 *d*; 在视角 2 中, *m* 实例在视角 3 中也出现了但其聚类结构中没有和其他节点有任何促进关系, 所以我们也可以将其当作不可识别节点处理, 参加不可识别节点局部优化矩阵的构建, 这里 *t* 为不可识别节点; 在视角 3 中实例 *c* 在本视角中和 *a*, *b* 实例同簇, 在其他视角中无促进作用, 所以在该视角中仍然作不可识别节点处理来避免其划分不准确的问题, 如果节点 *c* 与节点 *a* 和节点 *b* 的关联度大则在下节优化矩阵构建中仍然为同簇.

$$R_1^1=\begin{bmatrix} & a & b & c & d & e \\ a & 2.7 & 2.7 & 1.4 & 1.4 & 1 \\ b & 2.7 & 2.7 & 1.4 & 1.4 & 1 \\ c & 1.4 & 1.4 & 2.7 & 1 & 1 \\ d & 1.4 & 1.4 & 1 & 2.7 & 1.95 \\ e & 1 & 1 & 1 & 1.95 & 2.7 \end{bmatrix} R_2^1=\begin{bmatrix} & a & b & c & d & e & m & t \\ a & 2.7 & 2.7 & 1.4 & 1 & 1 & 1 & 1 \\ b & 2.7 & 2.7 & 1.4 & 1 & 1 & 1 & 1 \\ c & 1.4 & 1.4 & 2.7 & 1 & 1 & 1 & 1 \\ d & 1 & 1 & 1 & 2.7 & 1.4 & 1 & 1 \\ e & 1 & 1 & 1 & 1.4 & 2.7 & 1 & 1 \\ m & 1 & 1 & 1 & 1 & 1 & \underline{2.7} & 1 \\ t & 1 & 1 & 1 & 1 & 1 & 1 & \underline{2.7} \end{bmatrix}$$

$$R_3^1=\begin{bmatrix} & a & b & c & d & e & m \\ a & 2.7 & 2.7 & 1 & 1 & 1 & 1 \\ b & 2.7 & 2.7 & 1 & 1 & 1 & 1 \\ c & 1 & 1 & \underline{2.7} & 1 & 1 & 1 \\ d & 1 & 1 & 1 & 2.7 & 1.95 & 1 \\ e & 1 & 1 & 1 & 1.95 & 2.7 & 1 \\ m & 1 & 1 & 1 & 1 & 1 & \underline{2.7} \end{bmatrix}$$

图 2 3 个视角的选择调节矩阵

3.2 局部优化矩阵

局部优化矩阵的构建是利用局部学习和联合训练相结合的方法, 应用监督学习的思想来解决无监督学习中的聚类问题. 我们将每个视角中无法通过其他视角进行联合聚类的节点集称为不可识别节点集, 那么局部优化矩阵构建的基本原理, 就是基于核岭回归函数[15], 利用节点的邻居节点来估计不可识别节点集的划分. 通过提出优化函数的方式来得到不可识别节点集的最好的聚类结果如公式 (4) 所示:

$$\min_{F\in R^{n\times c}} \sum_{l=1}^{c}\sum_{i=1}^{n} (f_i^l - o_i^l(x_i))^2 = \sum_{l=1}^{c} \|f^l - o^l\|^2 \tag{4}$$

其中, n 表示不可识别节点个数, c 表示聚簇个数, f_i^l表示节点 i 在 l 聚簇中的标签值, 这里 $F^{n\times c}$ 为标准划分矩阵 ($F^{\mathrm{T}}F = I$), 公式中$o_i^l(*)$ 表示核向量机的输出函数[16], 利用核函数的监督方法, 训练节点 i 的邻居节点标签集$\{(x_j, f_j^l)\}$, 得到其在聚簇 l 下的估测划分值. 其中$o_i^l(*)$ 的计算如公式 (5) 所示:

$$o_i^l(x_i) = \sum_{x_j\in N_i} \beta_{ij}^l K(x_i, x_j) \tag{5}$$

其中 $K(x_i, x_j)$ 是径向基核函数, 这里 x_i 为核函数中心, N_i 表示 x_i 的邻居节点集, β 为回归参数.

3.2.1 利用核岭回归计算 F

我们将求解公式 (4) 的问题转换成求解公式 (6) 的问题:

$$\min \lambda(\beta_i^l)^{\mathrm{T}} K_i \beta_i^l + \|K_i\beta_i^l - f_i^l\|^2 \tag{6}$$

其中, K_i 为 x_i 邻居节点集的相似矩阵, f_i^l表示向量$\left[f_j^l\right]^{\mathrm{T}}$. 这里利用核岭回归 (KRR) 算法求得公式 (6) 中回归参数为$\beta_i^l = (K_i + \lambda I)^{-1} f_i^l$带入公式 (5) 得公式 (7):

$$o_i^l(x_i) = k_i^{\mathrm{T}}(K_i + \lambda \mathrm{I})^{-1} f_i^l = \alpha_i^{\mathrm{T}} f_i^l \tag{7}$$

其中k_i^{T}表示核函数的输出向量$[K(x_i, x_j)]^{\mathrm{T}}(x_j \in N_i)$, $\alpha_i^{\mathrm{T}} = k_i^{\mathrm{T}}(K_i + \lambda I)^{-1}$. 将公式 (7) 转换为以下公式 (8):

$$o^l = A f^l \tag{8}$$

可以看出f_i^l是f^l的子向量, 在公式 (8) 中如果 x_j 为 x_i 的邻居节点则$a_{ij} = \alpha_i$, 否则$a_{ij} = 0$. 将公式 (8) 代入公式 (3) 中得到公式 (9):

$$\min_{F \in \Re^{n \times c}} \sum_{l=1}^{c} \|f^l - A f^l\|^2 = \sum_{l=1}^{c} (f^l)^{\mathrm{T}} T f^l = trace(F^{\mathrm{T}} T F) \tag{9}$$

上式中$T = (I - A)^{\mathrm{T}}(I - A)$, F 为 T 矩阵的拉普拉斯矩阵前 d 个特征值对应的特征向量构成的结果阵.

3.2.2 获得最终局部优化矩阵 L

根据文献[17,18]中的聚类算法, 使用 (PM)P 表示聚类划分结果, 其中$P = [p_{il}] \in \{0,1\}^{n \times c}$, 如果$p_{il} = 1$表示 x_i 在 l 聚簇中且$\sum_{l=0}^{c} [p_{il}] = 1$. 上文中我们先使用标准划分矩阵 (PM)$F$ 表示划分结果, 其中 F 的初始定义为 $F = P(P^{\mathrm{T}}P)^{-\frac{1}{2}}$, 这里通过公式 (10) 可得到划分矩阵 P.

$$F^{\mathrm{T}}F = (P^{\mathrm{T}}P)^{-\frac{1}{2}} P^{\mathrm{T}} P (P^{\mathrm{T}}P)^{-\frac{1}{2}} = I \tag{10}$$

上式中 I 是单位矩阵, 根据 F 矩阵可以很容易的得到非标记节点的划分 L 矩阵如公式 (11):

$$L = P^{\mathrm{T}}P = (Diag(FF^{\mathrm{T}})^{-\frac{1}{2}} F)^{\mathrm{T}} (Diag(FF^{\mathrm{T}})^{-\frac{1}{2}} F) \tag{11}$$

最终扩展将非标记和标记节点聚类结果进行整合, 得到扩展的$\tilde{L}$矩阵.

4 实验

4.1 实验环境和数据集

本文我们主要使用来自于 UCI 数据库的 Iris 和 Wine 数据集和 DBLP 数据集进行测试实验. Iris 数据集包含 150 个鸢尾花实例样本, 每个实例包含 4 个属性 (花瓣长度, 花瓣宽度, 叶子长度, 叶子宽度) 进行描述, Wine 数据集与 Iris 数据集基本类似, 包含 178 个葡萄酒样本, 每个实例有 13 个属性信息 (化学分析指标) 描述, 其中如表 1 所示, 实验中, 随机选取几种属性构成一个多视角结构, 节点为实例向量, 边的权值表示各实例间的相似度值, 聚类结果将数据集分配到各自类别中.

表 1 UCI 数据集

数据集	样本数	属性数	类别数
Iris	150	4	3
Wine	178	13	3

DBLP 数据集我们提取了与数据库方向相关的 4 个会议 (SIGMOD, VLDB, ICDE, SIGKDD) 的作者及论文, 作者总数为 9628, 论文总数为 10175. 每个会议形成一个社会网络, 节点为作者, 边表示作者间协作关系. 最终形成的每个社区将代表具有相似研究兴趣的课题小组. 如表 2 所示.

表 2 DBLP 数据集

会议	作者数	论文数
SIGMOD	2790	2684
VLDB	2024	2444
ICDE	2024	3120
SIGKDD	2790	1927

4.2 评价指标

规范化互信息 (Nomalized Mutual Information, NMI)[19]用来度量网络的真实结构与经过算法所得到的聚簇结构的相似性. 假定两个集合向量 X 和 Y, NMI 的定义如公式 (9) 所示, NMI 的取值为 0~1 之间, 取值越大 (接近 1 时), 表示聚类结果越精确.

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \tag{12}$$

模块度评价方法[20]是一种基于内部标准的聚类效果衡量指标, 一般用于没有已知的聚簇结果对比时衡量聚类效果的优劣. 模块化是指网络中连接社区结构内部节点的边所占的比例和另一个随机网络中连接社群结构内部节点的边所占比例的期望值之差.

模块度的计算如公式 (13) 所示, 取值范围在 0 到 1 之间, 并且值越接近 1 表示簇内越紧密而簇间越分离.

$$Q(C) = \sum_i (E_{ii} - a_i^2) = Tr(E) - \|E^2\| \tag{13}$$

4.3 性能评估

实验 1. 本实验比较 Wine 数据集和 Iris 数据集的

多视角聚类结果. Wine 数据集中随机选取 4 个属性表示 4 个视角, 同理, Iris 数据集中选取 3 个视角, 分别比较节点重合率为 100%、75%、50%、25% 条件下的协同聚类结果, 如图 3 至图 10 所示, 从图中可以看出, 随着各个视角中相同点个数的增加, 社区发现的准确率逐渐提升, 当重复率为 100% 时, 迭代过程就变成传统的多视角聚类算法. 同时, 对于属性条件区分能力差的视角, 通过其他视角的相互影响, 能够有效提高该视角社区发现的准确性. 并且, 随着迭代次数的增加, 每个视角中可识别节点的局部社区结构得到集成处理, 各个视角的聚类结果都呈现收敛的状态.

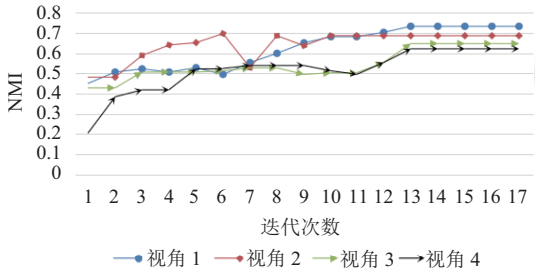

图 3 Wine 多视角协同选择聚类 (重合率 100%)

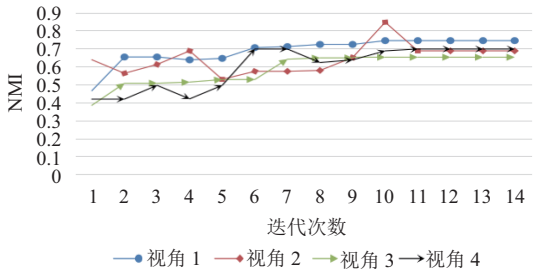

图 4 Wine 多视角协同选择聚类 (重合率 75%)

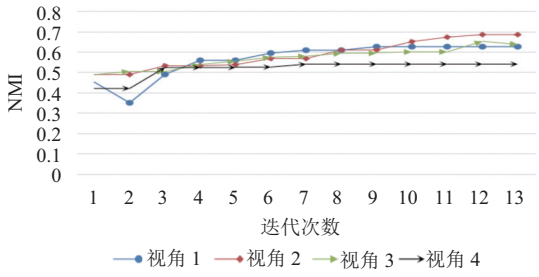

图 5 Wine 多视角协同选择聚类 (重合率 50%)

实验 2. 图 11 和图 12 是显示在同样覆盖率的条件下, 将 3 个视角的聚类精度与 2 个视角聚类精度进行比较, 可以看出 3 个视角的属性关系比 2 个视角的属性关系达到收敛状态后精度更高, 表明视角越多所包含的实例关系属性越完全, 聚类效果越好. 这里我们选取 100% 覆盖率, 是保证 1 视角和 2 视角的初始聚类精度相同.

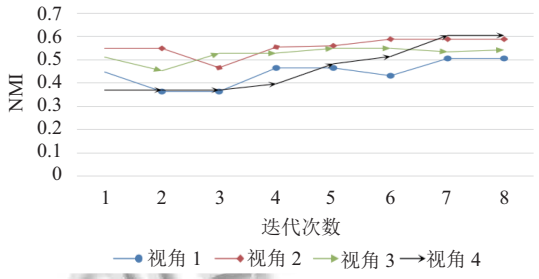

图 6 Wine 多视角协同选择聚类 (重合率 25%)

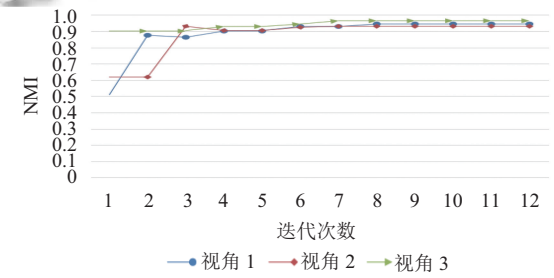

图 7 Iris 多视角协同选择聚类 (重合率 100%)

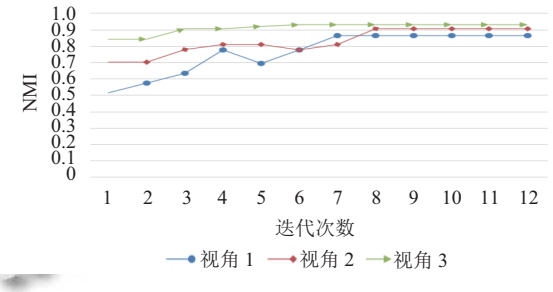

图 8 Iris 多视角协同选择聚类 (重合率 75%)

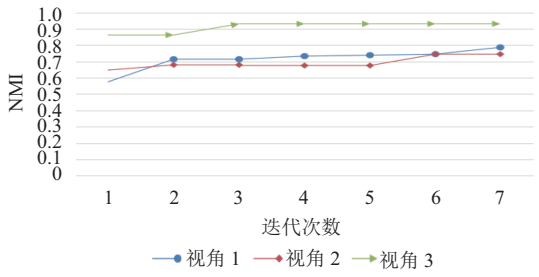

图 9 Iris 多视角协同选择聚类 (重合率 50%)

实验 3. 本实验将 co-MLSC 算法和 CSC 及 CS 算法算法进行比较, 各视角间节点集重合率为 75%, 从图 13 中可以看出, 当将谱聚类算法选取固定的最优聚

类中心的时候, 单视角的聚类结果呈持平状, 说明在单视角聚类中 CS 算法聚类结果已达到最优状态. 当重合率为 75% 的时, 不能直接使用 CSC 算法, 所以我们必须将各视角中的非识别节点添加到其他视角中并将相似度进行填 0 处理. CSC 的过度调整过程使低精度视角明显影响高精度视角的走势. 通过比较可以看出 co-MLSC 算法并没有过度调整的缺点.

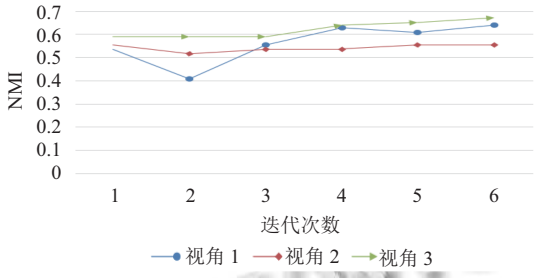

图 10 Iris 多视角协同选择聚类 (重合率 25%)

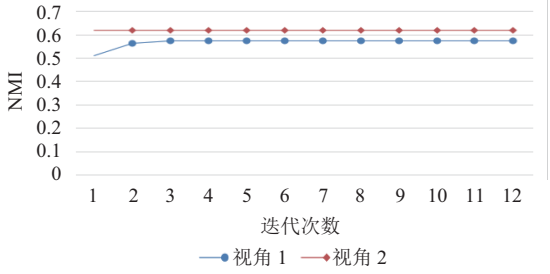

图 11 2 个视角的 Iris 数据聚类结果趋势图

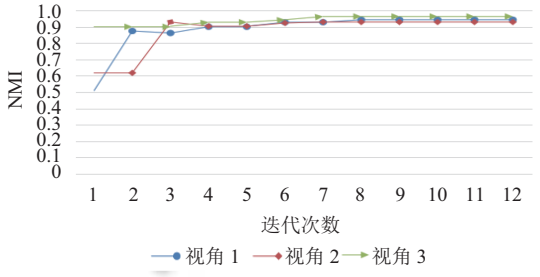

图 12 3 个视角的 Iris 数据聚类结果趋势图

实验 4. 使用 DBLP 数据集构建真实的基于作者协助关系的多维网络, 由于数据的重复率普遍较低, 所以我们进行了一定的人工调节, 删除合作者为 1 的论文, 并对作者统一 ID 值, 然而若将数据重复率控制太高的话每个领域内数据数目将非常少, 所以我们只能将数据集的重复率控制在 40% 以下进行变化, 具体实验结果如图 14 所示. 这里我们仅取 SIGMOD, VLDB, ICDE3 个视图关系结构, 该实验表现了本文提出的局部协同聚类算法在数据重复率较低的情况下有促进其他视角的作用, 并且也体现出该算法在真实构建的多维网络中有很好的适用性.

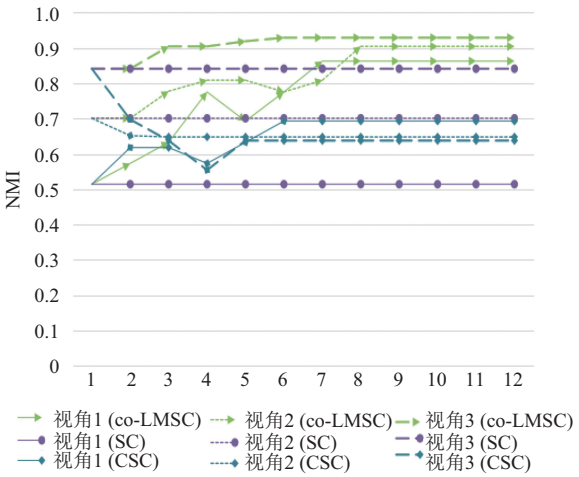

图 13 co-MLSC、CSC、CS 算法的比较结果图

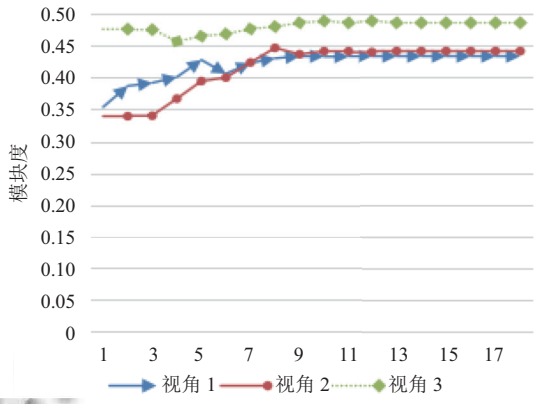

图 14 DBLP 数据集社团发现精度走势

5 总结

针对现有社区发现技术的不足, 本文提出一种基于局部协同选择聚类的多视角社区发现算法 co-MLSC, 该算法充分考虑了各视角间聚类的协同作用, 利用联合选择的方法相互促进来确定聚类中心结构, 同时也是对多个视角的聚类中心进行融合的过程, 以此来提升社区发现的准确性. 另外, 使用核岭回归估计的方法将每个视角下的不可识别节点进行划分, 最后通过实验验证了该算法的可行性和有效性.

下一步工作将对迭代式协同聚类算法的计算复杂度进行分析, 提出优化策略来降低社区发现算法的执行代价.

参考文献

1 Tang L, Wang XF, Liu H. Uncoverning groups via heterogeneous interaction analysis. Proceedings of the 9th IEEE International Conference on Data Mining. Miami Beach, FL, USA. 2009. 503–512.

2 Kanungo T, Mount DM, Netanyahu NS, *et al*. An efficient K-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 881–892. [doi: 10.1109/TPAMI.2002.1017616]

3 Ng AY, Jordan MI, Weiss Y. On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems 14. 2001.

4 Shi JB, Malik J. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888–905. [doi: 10.1109/34.868688]

5 Kuang D, Dingo C, Park H. Symmetric nonnegative matrix factorization for graph clustering. Proceedings of the 2012 SIAM International Conference on Data Mining. Anaheim, CA, USA. 2012. 106–117.

6 Xu W, Liu X, Gong YH. Document clustering based on non-negative matrix factorization. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, Canada. 2003. 267–273.

7 Strehl A, Ghosh J. Cluster ensembles-a knowledge reuse framework for combining multiple partitions. The Journal of Machine Learning Research, 2003, (3): 583–617.

8 Christoudias CM, Urtasun R, Darrell T. Multi-view learning in the presence of view disagreement. Proceedings of the 24thConferenceonUncertaintyinArtificialIntelligence.Helsinki, Finland. 2008.

9 Muslea I, Minton S, Knoblock CA. Active +semi-supervised learning=robust multi-view learning. Proceedings of the 19th International Conference on Machine Learning. San Francisco, CA, USA. 2002. 435–442.

10 Chaudhuri K, Kakade SM, Livescu K, *et al*. Multi-view clustering via canonical correlation analysis. Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Quebec, Canada. 2009. 129–136.

11 Kumar A, Daumé H. A co-training approach for multi-view spectral clustering. Proceedings of the 28th International Conference on Machine Learning. Bellevue, WA, USA. 2011. 393–400.

12 Kumar A, Rai P, Daumé H III. Co-regularized multi-view spectral clustering. Proceedings of the 24th International Conference on Neural Information Processing Systems. Granada, Spain. 2011. 1413–1421.

13 Niu DL, Dy JG, Jordan MI. Multiple non-redundant spectral clustering views. Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel. 2010.

14 Cheng W, Zhang X, Guo ZS, *et al*. Flexible and robust co-regularized multi-domain graph clustering. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, IL, USA. 2013. 320–328.

15 Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis. Cambridge, UK: Cambridge University Press, 2004.

16 Schlkopf B, Smola AJ. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge, MA: The MIT Press, 2001.

17 Chan PK, Schlag MDF, Zien JY. Spectral k-way ratio-cut partitioning and clustering. IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems, 1994, 13(9): 1088–1096. [doi: 10.1109/43.310898]

18 Yu SX, Shi JB. Multiclass spectral clustering. Proceedings of the 9th IEEE International Conference on Computer Vision. Nice, France. 2003.

19 Ng AY, Jordan MI, Weiss Y. On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems. 2002. 849–856.

20 Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004, 69(2): 026113. [doi: 10.1103/PhysRevE.69.026113]