一种基于节点相似性聚类的社区划分方法

王晓娟[1] 李晓晔[1]
WANG Xiaojuan LI Xiaoye

摘 要 现有的层次聚类社区划分通常使用公共邻居的数量来衡量节点对之间的相似性。由于忽略了公共邻居之间的差异性造成相似度值相同，许多节点对无法区分，导致社区划分结果不稳定、不准确。对此，提出了一种节点相似性聚类的社区划分方法 ISCM。在无权图上工作，在特征向量构建时考虑了节点的度、聚类系数、介数中心性、信息熵等综合指标，同时在相似性矩阵构建时采用综合指标的余弦相似性，实现了一个基于节点特征的谱聚类方法，而非传统基于图的方法。使用真实的数据集和评估算法，并在模块度、运行时间上进行了比较，结果显示，其在不同数据集上具有很高的模块度和准确性。

关键词 节点聚类；社区划分；余弦相似度；综合属性特征

doi：10.3969/j.issn.1672-9528.2024.06.024

0 引言

近年来，随着对网络性质的不断探索，人们发现在实际网络中除了小世界特性和无标度等特性外，还存在着社区结构。一般意义上的社区指的是网络节点的子集合，位于该集合内部的节点之间连接紧密，集合间的节点连接疏松。给定一个网络图，找出其社区结构的过程就是社区划分，它是网络分析中的一个基本问题。社区划分已经受到了国内外学者的广泛关注，但现有的社区划分算法忽略了不同节点之间的差异性，致使社区划分质量低下。因此，充分考虑节点之间的差异性以及同一社区间节点的相似性进行社区划分依然具有重要的意义。

现有社区划分的算法包括基于谱聚类、基于标签传播以及基于图神经网络的社区划分方法。谱聚类算法将网络结构划分的最终问题转化到求解矩阵的特征值和特征向量上。文献 [1] 针对谱聚类的社区发现方法以网络的邻接矩阵代替相似度矩阵造成效果受限的问题，通过信号传递原理衡量节点间相似度，并结合网络先验知识，通过引入半监督学习思想，在传统谱方法效果不佳时提供更可信的划分，并在现实和 LFR 人工网络上进行测试。文献 [2] 的改进是其先利用马尔可夫过程计算节点间的转移概率，并基于转移概率构建网络的概率矩阵，之后以均值概率矩阵重新构造相似图，最后通过优化归一化切割函数实现社区划分。文献 [3] 提出了二分网络社区发现方法，其应用了标准化的谱聚类于二分网络上，将分类效果较好的谱聚类算法与交互度集成，提高了二分网络聚类的精确度。在基于标签传播的社区划分方法中，LPA 是经典的社区检测算法之一。由于其随机更新导致的不稳定性，文献 [4] 运用 LDA 和 K-means 算法改进了网络标签传播方式。文献 [5] 在网络初始划分中，选择高节点度的种子节点，通过节点权重确定标签迭代顺序。Steve 对其进行延伸，提出了多标签传播算法 COPRA[6]，初始时每一个节点赋予一个标签并允许每个节点最多携带 v 个标签，但对于未知的网络，无法估计网络中节点所属的社区个数。研究者发现，图神经网络（graph neural networks，GNN）可以将网络的结构信息和节点属性信息结合，同时进行学习。传统的图神经网络在同质信息网络社区发现问题表现较好，但不能利用异质信息网络不同节点类型和边类型的特点，不能较好利用异质网络的语义信息。文献 [7] 基于图神经网络，并且根据异质信息网络的特点，提出基于异质图注意力网络的重叠社区发现方法。同时，文献 [8] 提出基于节点重要性和双重自编码器的社区发现方法，学习了网络的结构信息和属性信息，并实现了特征的加权求和，最终实现了社区的划分。

本文提出基于节点相似性聚类的社区划分算法 ISCM。首先计算成对节点间的相似性生成节点向量，再进行节点间相似性度量，选择每个节点的 k 邻居节点构成加权相似图，再重构相似图矩阵进行社区划分。

1 节点相似性聚类的社区划分

在基于相似性的社区划分算法中，公共邻居常用于计算节点相似性。然而，在复杂网络中，不同节点对之间的公共

1. 齐齐哈尔大学计算机与控制工程学院
黑龙江齐齐哈尔 161000

[基金项目] 黑龙江省省属高等学校基本科研业务费科研项目（145209124）；黑龙江省高等教育教学改革研究项目（SJGY20210962）

邻居的数量和度值通常是相同的。此时，当使用聚合策略形成初始社区时，节点选择的随机性和不确定性会降低社区划分的准确性。因此，考虑一些度量来计算成对节点之间的相似性，利用节点的聚类系数结合度的熵模型设计了一种新的基于节点聚类的度的熵的节点相似性度量，将其用于形成社区。在计算相似度矩阵时，综合考量节点的度、介数中心性、聚类系数和基于度的熵，以提高聚类的准确性。ISCM 算法由以下几个阶段组成：节点聚类、节点特征提取及相似度度量、社区划分。

1.1 节点聚类

节点聚类的目标是将图 G 中的节点划分为 T 个不相交的簇 $\{C_1, C_2, \cdots, C_T\}$，使得同一簇中的节点在图结构方面通常彼此靠近，而在其他方面则相距遥远。因此，节点聚类问题可以被视为找到图的一个分区，使得同一集群中的节点比不同集群之间的节点更相似。也就是说，集群内的边比集群之间的边具有更高的权重。对于给定的图 $G=(V, E)$，将 G 转换为相似图 G_S，以实现节点聚类。谱聚类是最常用的聚类方法之一，其性能优于传统的聚类方法。因此，使用 RatioCutgraph 类似谱聚类的划分方案将相似图划分为 T 个子图，使子图的大小近似相等。

1.2 节点特征提取及相似度度量

谱聚类的效果在很大程度上取决于用于构建相似性矩阵的相似性度量 [10]。为了构建相似图，考虑以下度量来计算每对节点之间的相似性：节点度、介数中心性、局部聚类系数和基于度的图熵。

定义 1（局部聚类系数）

$$Lc(v_i)=\mu_G(v_i)/\omega_G(v_i) \tag{1}$$

其中，分子分母分别表示封闭的三元组数目以及所有三元组数目。

定义 2（介数中心性）

图 G 中节点 v 的 $BC(v)$ 是图中通过 v 的所有节点对之间的最短路径的分数。

$$BC(v)=\sum_{s,t\in v,s\neq t\neq v}\frac{\sigma_{st}(v)}{\sigma(v)} \tag{2}$$

定义 3（基于度的图熵）[11]

$$I_f(G)=log(\sum_{i=1}^{n}d_i)-\sum_{i=1}^{n}\frac{d_i}{\sum_{j=1}^{n}d_j}logd_i=log(2m)-\frac{1}{2m}\sum_{i=1}^{n}d_ilogd_i \tag{3}$$

对于每个节点 $v_i\in G$，分别计算 $BC(v_i)$、$Lc(v_i)$ 和 $I_f(G(v_i))$。将向量 $x_i=<D(v_i),BC(v_i),Lc(v_i),I_f(G(v_i))>$ 称为节点向量。因此，每对节点 v_i、v_j 的相似性函数可以定义为：

$$f_{sim}(v_i,v_j)=e\text{^}((v_i\cdot v_j)/(\|v_i\|*\|v_j\|)) \tag{4}$$

通过上述对网络拓扑特性计算得到节点特征，为每个节点构建一个节点向量，其中包含节点的度、介数中心性、聚类系数和熵。

综合网络拓扑的节点特征构建特征向量，并通过节点对之间的公共邻居节点数来计算节点间相似度，选择节点的 k 近邻构成相似图，相似度矩阵被转换为加权邻接矩阵，其中权重表示节点间的相似度。

1.3 社区划分

传统的谱聚类算法通常使用节点之间的相似度直接构建相似度矩阵。此种方法可以根据节点的特征进行聚类，具有更好的聚类效果，具体算法如下。

算法 基于节点相似性聚类的社区划分方法

输入：原始社交网络图 G

输出：社区发现结果

1.for each v in V do

2. 根据公式（1）、（2）、（3）计算每个节点向量 $x_i=<D(v_i),BC(v_i),Lc(v_i),I_f(G(v_i))>$

3. 根据公式 4 计算节点间相似度 $f_{sim}(v_i,v_j)$

4.end for

5.for i = 1 to n do

6. 选择节点 v_i 的前 k 个近邻节点

7. 在节点 v_i 和它的邻居节点间加边构成相似图 G'

8.end for

9. 重构相似图 G' 的邻接矩阵和度矩阵分别记为 $\boldsymbol{W}$ 和 $\boldsymbol{D}$

10. 重构拉普拉斯矩阵 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$

11. 计算拉普拉斯矩阵 $\boldsymbol{L}$ 的特征值

12. 计算 x_1、$x2$、…、x_T 的前 T 个特征值的对应特征向量

13. $y_i=x_i/\sqrt{N_i}$，i = 1, 2, ..., T, $Y=(y_i)_{T\times n}$

14. 利用 k 均值算法将这 n 个节点划分为 T 个簇

15. 返回 $C_1, C_2, ..., C_T$

接下来，描述经典数据集 karate 中，指定社区数量时的社区划分的结果，如图 1 所示。

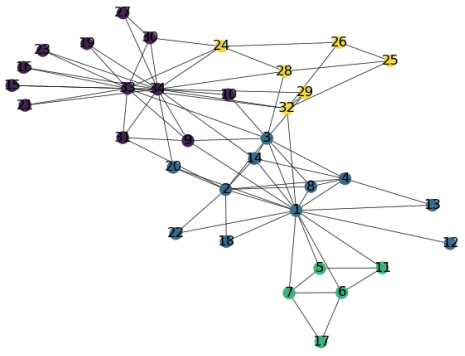

图 1 karate 数据集划分

用可视化工具来观察社区结构和节点之间的连接模式，以了解社区的基本特征和可能存在的子群。

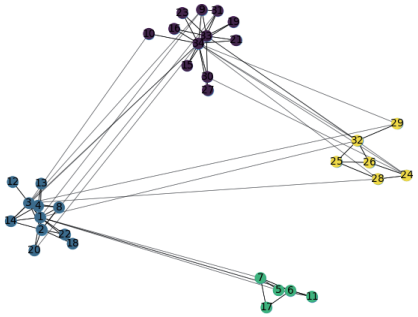

图 2 karate 数据划分的结果

2 实验结果与分析

2.1 实验数据

在 3 个真实的网络中进行实验，对于社区结构明显的网络，通过与网络的实际划分情况对比得出结论，以验证本文提出的基于节点相似性聚的谱方法划分社区的有效性。具体的网络信息如表 1 所示。

表 1 数据集参数

真实网络	节点数	边数
dolphin	62	159
polblog	1224	16 718
facebook_combined	4039	88 234

2.2 评价指标

（1）模块度

模块度是一种用来衡量网络社区结构的指标之一，它衡量了社区内部连接的紧密程度相对于社区之间连接的稀疏程度，可以用来作为评估社区划分算法的性能指标。模块度的计算基于节点之间的连接性和社区内外连接的差异性。模块度是基于 Newman 等人提出的模块度计算公式实现的，其公式为：

$$Q=\frac{1}{2m}\sum_{ij}[A_{ij}-\frac{k_ik_j}{2m}]\delta(c_i,c_j) \tag{5}$$

式中：Q 表示模块度，m 表示图中边的数量，A_{ij} 表示节点 i 和节点 j 之间的边的权重，k_i 和 k_j 分别表示节点 i 和节点 j 的度数（关联的边数），ci 和 cj 分别表示节点 i 和节点 j 所属的社区编号，$\delta(c_i, c_j)$ 是一个指示函数，当节点 i 和节点 j 属于同一个社区时，取值为 1，否则取值为 0。

（2）Jaccard 相似度

Jaccard 相似度：对于每一对社区（一个来自算法划分，另一个来自真实划分），计算它们之间的 Jaccard 相似度。它提供了一个度量，显示两个社区划分之间在节点层面的相似性，其公式为：

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|} \tag{6}$$

（3）调整兰德指数

调整兰德指数（ARI）是一种用于衡量两个社区划分的相似性的指标，使用 ARI 来评估提出的算法划分与真实划分之间的一致性，其公式为：

$$\text{ARI}=\frac{\text{（观察到的一致性）-（预期一致性）}}{\text{（观察到的一致性）+（预期一致性）}} \tag{7}$$

2.3 实验结果分析

Karate 网络反映了俱乐部成员之间的关系，由于管理员和教练之间的纠纷，俱乐部成员之间的关系被分为两部分。节点“1”和节点“34”分别代表俱乐部的管理员和教练。Karate 网络的真实社区如表 2 所示。

表 2 Karate 网络的真实社区

社区 ID	节点 ID
1	1 2 3 4 5 6 7 8 11 12 13 14 17 18 20 22
2	9 10 15 16 19 21 23 24 25 26 27 28 29 30 31 32 33 34

可以看出，最紧密相似的节点是节点“1”和节点“34”，它们是两个网络的核心节点。本文提出的 ISCM 方法由于考虑了节点的综合特征，能够准确地识别网络中的核心节点，如图 3 所示。

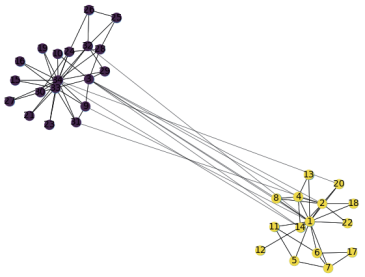

图 3 ISCM 划分 Karate 数据集结果

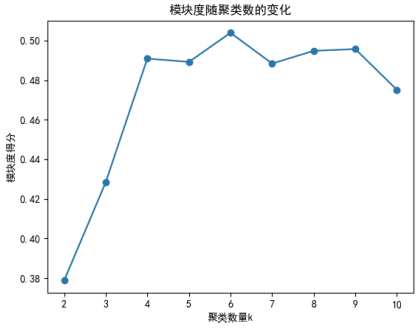

图 4 海豚网络社区划分模块度

Dolphin 数据集中，开始时，随着聚类数量的增加，模块度往往会显著上升。这是因为初始的少量聚类节点会形成一些相对独立的社区，社区内部连接紧密，且社区之间的连接相对稀疏。在这个凸起点附近的聚类数量范围，模块度可能会出现小幅度的反弹。这是因为在一些特定的聚类数量上，社区的划分方式可能刚好捕捉到了网络的某些结构，使得社区内部连接更紧密。

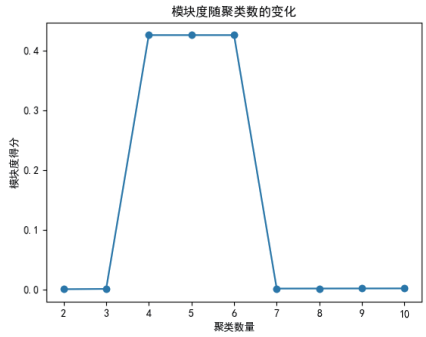

图 5 博客数据集模块度

Polblog 数据集中，因为原本图就是不联通的，而提出的方法又以谱聚类为基础，谱聚类对于子图的依赖性较强，所以不同参数变化会对社区的划分衡量产生较大影响。

Facebook_combined 数据集中，由于节点数量较多，在进行社区划分时在原来基础上多取了一部分点，发现社区划分数量在 10 个左右时，模块度最高且比较稳定。由此可以看出，提出的算法更适用于在较大型网络中进行社区划分。

由于只有 dolphin 数据集描述了新西兰 Dolphin 群体之间的关联，表 3 给出了真实的社区划分结果。

表 3 dolphin 网络的真实社区

社区 ID	节点 ID
1	1 3 4 5 9 11 12 13 15 16 17 19 21 22 24 25 29 30 31 34 35 36 37 38 39 40 41 43 44 45 46 47 48 50 51 52 53 54 56 59 60 62
2	2 6 7 8 10 14 18 20 23 26 27 28 32 33 42 49 55 57 58 61

将提出的方法和已有的方法在具有真实社区的 dolphin 网络上进行比较。提出的方法获得了很好的 ARI 值，划分的社区也更接近真实社区，如表 4 所示。

表 4 dolphin 数据集上 3 种算法的数值比较

网络	Jaccard 相似度	ARI	模块度
ISCM	1.0	0.934	0.504
SLPA	1.0	1.0	0.399
CORPA	1.0	1.0	0.379

在运行时间上，将所提出的算法与其他算法在小型和较大型数据集上进行测试，证明在较大型数据集上，本文提出算法比较有效，如图 6 所示。

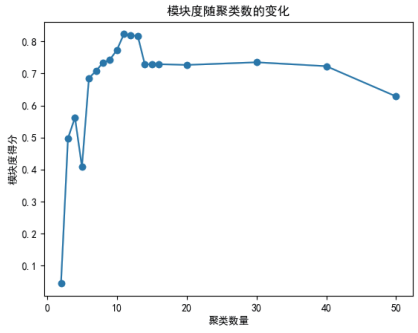

图 6 Facebook 数据集模块度

模块度比较，本文采用稳定时的平均值，如图 7 所示。

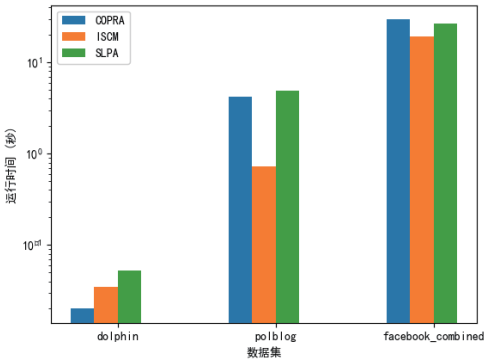

图 7 运行时间比较

可以明显看出，与之前两种算法比较，提出的算法在处理较大型数据时运行时间明显要低于其余两者。由于 dolphin 数据集较小，各个算法的运行时间都比较小，比较接近 0，故在柱状图中显示不是非常明显。从图 8 中可以清楚看到各个算法在不同数据集上划分的模块度得分变化趋势。

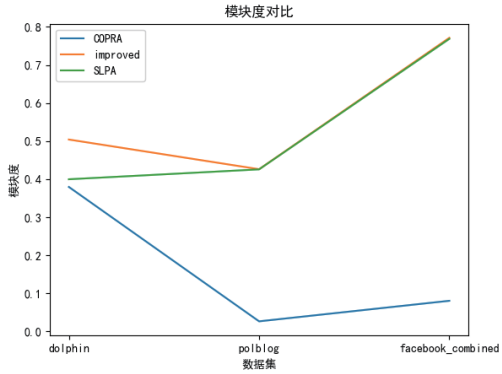

图 8 模块度对比

从图 8 中可以看出，本文提出的算法较其余两种算法相比模块度更高，即社区划分的结果更为优秀。

综上可以看出所提出的算法 ISCM 在较大型密集联通的数据集 Facebook_combined 上运行时间与 COPRA、SLPA 算法更低，模块度划分更高；实现了较大型连通网络中缩短了运行时间同时获得了更好的模块度得分。

3 总结

大多数网络都包含结点的特征信息和结构信息，将两者关联在一起，通常可以提高社区发现的精度。文章从图中的每个节点提取多种特性（度数、介数中心性、聚类系数和信息熵），不只是计算概率分布，还考虑了节点的特征信息。在考虑节点相似性特征的同时在谱聚类基础之上提出的社区发现算法，可以适用于较大型数据集，并缩短运行时间。由于网络演化的复杂性，动态网络社区划分得到的关注和工作量都不如静态网络多。未来，将在现有工作的基础上，捕捉动态网络中的社区结构，并尝试在没有任何先验知识的情况下自动确定社区的数量。

基于 IndRNN 的机场起飞航班延误预测模型研究

司毅洋[1] 吕 娜[1]
SI Yiyang LYU Na

摘 要 由于没有综合考虑天气等突发状况，导致航班延误预测结果存在一定的偏差，降低用户飞行体验，对此，提出基于 IndRNN 的机场起飞航班延误预测模型研究方法。起飞航班延误划分为 4 个等级，利用 ReLU 激活函数代替 IndRNN 网络中的 sigmod、tanh 激活函数，使得每个神经元都有其独立的时空特征；分离所有神经元，避免梯度出现消失爆炸的情况；经过数据读取、数据预处理、数据融合等一系列操作后，完成航班延误预测模型的构建。通过开展对比仿真实验，在 4 项评判指标下，所提方法均展现出了优秀的预测性能，且预测延误航班数、延误时间与实际值非常接近。

关键词 IndRNN 网络；起飞航班延误预测；ReLU 激活函数；传播梯度；数据预处理

doi：10.3969/j.issn.1672-9528.2024.06.025

0 引言

近年来，我国航班数量出现了大幅度增长，航班运输网络也越来越庞大、繁杂，因天气、飞机故障等原因导致的航班延误现象屡见不鲜，由此产生了巨大的经济损失。因此，展开对机场离港航班的延误预测是非常有必要的，可以在一定程度上帮助航空公司、机场以及相关单位制定延误解决方案，降低延误出现的次数，降低因延误产生的经济损失。

1. 新乡工程学院信息工程学院 河南新乡 453000

对此，吴仁彪等人[1]提出利用 CBAM-CondenseNet 实现对航班的延误预测。首先，分析航空系统中因航班延误产生的波及现象，得到受影响的航班链；然后，对航班链中的数据进行清洗，将其中的机场数据与航班数据进行融合；最后，对融合后的结果通过 CBAM-CondenseNet 算法完成特征提取，并结合 Softmax 分类器划分航班延误等级，完成预测。该方法仅对受影响的航班进行分析，没有考虑其他航班可能遇到的突发状况，适用范围较为狭隘。张成伟等人[2]通过分析离港航班计划，确定某航班出现延误的情况，完成预测。

参考文献：

[1] 崔宇童，牛强，王志晓 . 基于信号传递的半监督谱聚类社区发现算法 [J]. 计算机工程与设计 ,2018,39(5):1201-1205+1213.

[2] 张书博，任淑霞，吴涛 . 结合概率矩阵的改进谱聚类社区发现算法 [J]. 西安电子科技大学学报 , 2019,46(3):167-172.

[3] 张晓琴，安晓丹，曹付元 . 基于谱聚类的二分网络社区发现算法 [J]. 计算机科学 ,2019,46(4):216-221.

[4] 赵承志 . 融合节点信息的 LPA 社区检测算法的改进研究 [D]. 沈阳 : 东北财经大学 ,2022.

[5] 蔡威林，葛斌 . 基于影响度的标签传播算法 [J]. 佳木斯大学学报 (自然科学版), 2022,40(1):38-40+160.

[6]GREGORY S.Finding overlapping communities in networks by label propagation[J].New journal of physics, 2010, 12: 103018-103043.

[7] 孙悦 . 基于 GNN 的异质网络重叠社区发现算法研究 [D]. 包头 : 内蒙古科技大学 ,2023.

[8] 李昕泽 . 基于图神经网络的社区发现方法研究 [D]. 北京 : 北方工业大学 ,2023.

[9]YE X, SAKURAI T.Robust similarity measure for spectral clustering based on shared neighbors[J].ETRI journal, 2016, 38(3): 540-550.

[10]LIU Q, WANG G, LI F, et al.Preserving privacy with probabilistic indistinguishability in weighted social networks[J]. IEEE transactions on parallel and distributed systems, 2017, 28(5): 1417-1429.

[11]CAO S, DEHMER M, SHI Y.Extremality of degree-based graph entropies[J].Information sciences,2014,278(10):22-33.

【作者简介】

王晓娟（1998—），女，山东德州人，硕士研究生，研究方向：社交网络隐私保护。

（收稿日期：2024-03-22）