• 复杂性科学 •

基于多模体特征的科学家合作预测

曹红艳，许小可，许 爽*

(大连民族大学信息与通信工程学院 辽宁 大连 116600)

【摘要】科学学随着科学本身的发展已成为近年来国内外研究的热点，科研组织与知识传播的重要结构基础——科学家合作网络因此受到学者们的广泛关注。在此情况下，科学家合作网络中的合作形成及合作权重强弱成为很有意义的研究问题。该文提出了基于多模体特征和机器学习框架的链路预测和权重预测方法，将实验结果与几种经典方法进行对比，发现该方法可以有效提高预测的准确率，链路预测最高可提高 8.9%，而权重预测最高可提高 59.6%。该研究有助于预测科研网络中科学家合作的可能性及其合作权重，进而挖掘科学家合作网络的结构特性对学者科研产出和团队合作的深刻影响。

关 键 词 链路预测； 多模体特征； 科学家合作； 科学学； 权重预测

中图分类号 TP391 文献标志码 A doi:10.12178/1001-0548.2019173

Predicting Scientist Cooperation Based on Multiple motif Features

CAO Hong-yan, XU Xiao-ke, and XU Shuang*

(School of Information and Communication Engineering, Dalian Minzu University Dalian Liaoning 116600)

Abstract With the development of science itself, science of science has become research in recent years. The scientific cooperation network which is an important structural foundation of scientific research organizations and knowledge dissemination has attracted wide attention from scholars. Under this circumstance, the formation of cooperation and the weight of cooperation in the scientific cooperation network have become very meaningful research issues. This paper proposes a link prediction and weight prediction methods based on multiple motif features and machine learning framework, and compares the experimental results with several classical methods. It is found that the proposed methods can effectively improve the accuracy prediction: up to 8.9% in the link prediction and 59.6% in the weight prediction. This paper helps to predict the possibility of scientist collaboration in the scientific research network and their cooperation weight, and then to explore the profound impact of the structural characteristics of the scientific cooperation network on the scientific research output and teamwork of scholars.

Key words link prediction; multiple motif; scientist cooperation; science of science; weight prediction

随着科学研究的迅猛发展和数据分析技术的应用，“科学学”已经成为近年来国内外研究的热点[1-2]。其中，由于科学家合作网络是科研活动组织与科学信息传播的重要结构基础，因此受到科研学者的广泛关注[3]。在此情况下，科学家合作网络中的合作形成以及合作权重预测就成为很有意义的研究课题，对应网络科学中的科学问题为链路预测[4] 和权重预测[5]。通过对科学家合作网络的定量分析，可以辨识科学家在合作网络中的角色、了解他们之间的合作模式并预测他们学术合作的可持续性和合作强度。本文基于多个实证科学家合作网络的多模体特征对科学家合作进行预测，旨在预测网络中的科学家之间未来合作的可能性，及其合作的强度。

链路预测是通过网络中部分节点以及它们之间的结构信息，预测网络中任意两个节点之间存在连接的可能性[6]。近年来，基于网络结构相似性的链路预测方法引起了学者们的广泛关注。文献 [7] 提出了基于网络拓扑结构的相似性预测方法，并发现在科学家合作网中使用节点的共同邻居 (common

收稿日期：2019－08－04；修回日期：2019－11－15
基金项目：国家自然科学基金 (61773091)；辽宁省高等学校创新人才支持计划 (LR2016070)；辽宁省重点研发计划指导计划 (2018104016)
作者简介：曹红艳 (1994－)，女，主要从事社交网络数据挖掘和链路预测方面的研究.
通信作者：许爽，E-mail：xushuangcong@163.com

neighbors, CN) 和 Adamic-Adar(AA) 指标进行预测的准确性最好[8]。文献 [9] 提出了使预测准确性更高的资源分配 (resource allocation, RA) 指标和局部路径 (local path, LP) 指标。文献 [10] 将局部随机游走应用于链路预测，进一步提高了预测的准确性。近年来，人们将机器学习方法应用于链路预测问题，综合性使用多种特征大大提高了链路预测的准确性[11]。以上方法大都仅能应用于无权网络的链路预测中，在加权网络链路预测中，文献 [12] 提出了 3 个基于无权网络的经典相似性 CN、AA 和 RA 的加权形式指标，分别是 WCN、WAA 和 WRA。

权重预测是预测两个节点之间连接的权重。在现实世界中的许多网络都是加权网络，在不同的网络中权重通常代表不同的含义。如在航空网络中，连边权重代表航班数量；在社交网络中，连边权重代表朋友间的亲密程度；在科学家合作网中，连边权重代表作者与作者之间的合作强度。连边的权重预测是一个较新的研究主题，科研人员提出了一些有效的预测算法。文献 [13] 提出了一种基于局部网络结构 (分析节点的邻居集合结构) 的权重预测方法，此方法在链接存在或不存在的情况下都可以使用，且可以比线性相关方法更准确地预测权重。文献 [5] 提出了“可靠路线”策略来预测网络中的连边和权重，将未加权的局部相似性指数扩展到加权的网络中，称为 rWCN、rWAA 和 rWRA，使用这些指标计算出相似性得分，取得了较好的预测效果。

目前科学家合作网络中的链路预测和权重预测算法中，基于拓扑结构的相似性指标往往仅关注了科学合作的传递性结构 (三角形关系)，将这类指标应用于科学家合作的预测中，只对应了一种科学家之间的合作模式，而忽略了科学家之间存在的其他合作模式及多种合作模式的组合，这些合作模式可以表达为模体或子图。模体的概念最早是由文献 [14] 提出，定义为实际网络中出现的频次远远高于其对应随机网络的子图[15]。模体基于微观结构刻画了真实网络中局部相互作用的合作模式，并自下而上自组织构成网络整体结构[3]。

综上所述，本文在研究通用三角关系一种模体结构特征预测基础上，又利用了科学家合作网中其他 7 种合作模式，即 8 种模体结构进行合作预测。研究中通过提取科学家合作网中的单模体特征和多模体特征统计量，并采用机器学习算法对特征进行分析计算，实现网络中科学家之间合作的可能性及其合作强度的链路预测和权重预测，取得了较好的预测性能。

1　问题描述及评价指标

1.1　问题描述

本文使用的多个科学家合作网络为加权无向网络，形式为$G(V,E,W)$，其中V、E 和W分别是网络中的节点集合、连边集合和权重集合。网络中每条连边由(x,y)表示，且$(x,y)\in E$，每条连边的权重由w_{xy}表示。由于本文使用的都是无向网络，所以$w_{xy}=w_{yx}$。数据集被随机划分成训练集E_T和测试集E_V两部分，其中$E_T\cup E_V=E$，$E_T\cap E_V=\emptyset$。

1.2　评价指标

1) 链路预测评价指标 AUC

衡量链路预测算法性能的指标有 3 种，分别是 AUC、精确度 (precision) 和排序分 (ranking score)，它们的侧重点各不相同。其中，AUC 可以从整体上衡量算法的精确度而得到最广泛的使用[6]，因此本文采用该指标衡量不同算法链路预测的准确性。在链路预测算法中，计算出所有测试集两两节点间的相似度得分之后，AUC 指标可以描述为如下形式：每次从测试集中随机选取一条存在的边(x,y)，然后随机选取一条不存在的边(x_1,y_1)，比较这两条边的相似度得分，如果边(x,y)的分数大于边(x_1,y_1)的分数，则加 1 分；如果两条边的分数相等，则加 0.5 分。独立比较 n 次，如果有n'次边(x,y)的分数值大于边(x_1,y_1)的分数值，有n''次两条边的分数值相等，则 AUC 值可以定义为：

$$\mathrm{AUC}=\frac{n'+0.5n''}{n}$$

通常，上述评分算法计算出的 AUC 值应该至少大于 0.5。AUC 的值越高，算法的精确度越高，但 AUC 的值最高不会超过 1。

2) 合作权重预测评价指标 RMSE

科学家合作网络是被用于研究科学学的主要途径[1]，合作权重是科学家网络中的重要特征之一。本文使用均方根误差作为合作权重预测评价指标。它亦被称为标准误差，是真实值与预测值之间差值的平方与样本数 n 比值的平方根，具体定义为：

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{(x,y)=1}^{n}(w_{xy}-\hat{w}_{xy})^2}{n}}$$

式中，(x,y)为测试集中连边的集合；w_{xy}表示测试集中连边的真实值；$\hat{w}_{xy}$表示某种预测方法所给出

的预测值。RMSE 反映了预测值偏离真实值的程度，RMSE 越小，则表示预测的精度越高。

2 预测方法

2.1 基于共同邻居加权特征的预测方法

1) 加权 CN 指标 (WCN)：

$$s_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{w_{xz}+w_{yz}}{2}$$

式中，z为x和y的共同邻居；w_{xz}表示连接节点x和z之间连边的权重值；w_{yz}表示连接节点y和z的边的权重值。如果所有边权重都等于 1，那么上述指标都等价于无权的 CN 指标。

2) 加权 AA 指标 (WAA)：

$$s_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{w_{xz}+w_{yz}}{2\lg(1+s_z)}$$

式中，$s_z=\sum_{j\in\Gamma(z)}w_{zj}$表示节点$z$的强度。WAA 可以认为是 WCN 的一种变形，根据每个共同邻居强度值进行了加权。

3) 加权 RA 指标 (WRA)：

$$s_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{w_{xz}+w_{yz}}{2s_z}$$

WRA 是 WCN 的另外一种加权形式。

基于共同邻居加权特征的科学家合作预测主要提取测试集数据的上述指标，将指标得分视为数据集中可能存在连边的相似度得分，通过相关的评价指标衡量预测的准确性。

2.2 基于可靠路线加权特征的预测方法

在先前研究中，只有少数的预测算法被扩展到了加权网络，而且大多数只考虑了网络的拓扑结构，很少充分利用连边的权重信息。受通信网络中可靠路线问题的解决方案的启发，文献 [5] 提出了可靠路线策略方法来预测网络中的连边和权重，将未加权的局部相似性指数扩展到加权网络中，并使用这些指标计算出相似性得分，取得了较高的准确性。

1) 可靠路线加权 CN 指标 (rWCN)：

$$s_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}w_{xz}w_{yz}$$

2) 可靠路线加权 AA 指标 (rWAA)：

$$s_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{w_{xz}w_{yz}}{\lg(1+s_z)}$$

3) 可靠路线加权 RA 指标 (rWRA)：

$$s_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{w_{xz}w_{yz}}{s_z}$$

基于可靠路线加权特征的预测方法是从已知的训练集E_T和测试集E_V得到它们的加权邻接矩阵$\boldsymbol{W}_T$和$\boldsymbol{W}_V$，提取测试集中的上述指标得到连边相似度分数，由S_V表示。然后考虑到线性相关性，引入一个自由参数λ，定义预测函数为$F(\boldsymbol{W}_T)=\lambda\cdot S_V$，通过最小化预测函数与$\boldsymbol{W}_V$之间的差来确定$\lambda$：

$$\min_{\lambda}\|\lambda\cdot S_V-\boldsymbol{W}_V\|_F$$

式中，$\|\cdot\|$为 Frobenius 范数。最后通过相关的评价指标来衡量预测的准确性。

2.3 基于模体特征的预测方法

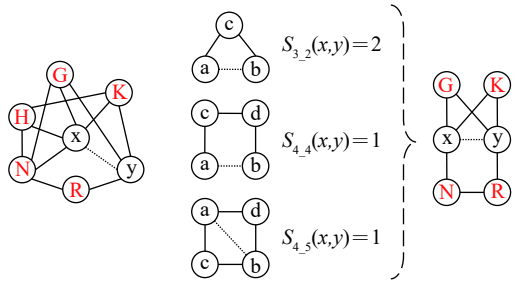
基于共同邻居和可靠路线的加权网络链路预测方法都是基于网络的传递特性 (分析三角形模体)。在基于模体特征的预测方法中，共涉及 8 个模体特征，分别为 2 个三节点模体和 6 个四节点模体，它们代表了科学家合作网中的 8 种合作模式。所有的模体编号、图示和合作模式如表 1 所示[3]，其中(a,b)为待预测连边。

表 1 模体对应的科学家合作模式

模体编号	图示	合作模式
3_1		一位科学家与两位不合作的科学家其中的一位合作，则可能与另一位科学家合作
3_2		一位科学家与两位不合作的科学家合作，则另两位科学家可能会合作
4_1		一位科学家与三位不合作的科学家中的两位合作，则可能与另一位科学家合作
4_2		两位合作的科学家与另两位合作的科学家之间可能会有合作
4_3		三位科学家两两合作，则第四位科学家可能与其中一位科学家合作
4_4		两位合作的科学家分别与两位无合作的科学家合作，则另两位科学家可能合作
4_5		四位科学家中，每位与且仅与其中两位科学家合作，则与第三位科学家可能会合作
4_6		除某两位科学家不合作外，四位科学家两两相互合作，则不合作的两位科学家可能合作

基于模体特征的科学家合作预测主要是提取训练集和测试集的模体特征，将每种模体的数量作为特征值，科学家之间是否合作作为机器学习的分类

标签、科学家之间合作的强度作为回归的预测值，得到预测结果后使用相关评价指标衡量预测的准确性。图 1 为基于模体特征的科学家合作预测的具体过程。



a. 7节点小网络图 b. 科学家合作预测的过程范例 c. 组合结构

图 1　基于模体特征的科学家合作预测

如图 1 所示，图 1a 为一个 7 节点的小网络图，边(x,y)为待预测连边。图 1b 以模体特征 3_2、4_4 和 4_5 为例说明科学家合作预测的主要过程。如图 1b 所示，分别计算模体特征 3_2、4_4 和 4_5 在图 1a 中的个数。模体特征 3_2 的计算方法为寻找节点x和y的共同邻居数。模体特征 4_4 的计算方法为寻找节点x和y的除去共同邻居节点的各自邻居节点，将节点x和y的各自邻居节点进行组合，其中邻居节点之间在网络中存在连边的记为 1，最终将所有连边记为 1 的累加求和。其他特征的计算方法依次类推。通过计算得出模体特征 3_2 的个数为 2，模体特征 4_4 的个数为 1，模体特征 4_5 的个数为 1。

在进行科学家合作预测时，可以将这些模体中的每一种模体的数量值单独作为机器学习方法的输入。也可以将图 1b 中的 3 种模体在拓扑结构上进行组合，即 $S_{3\_2}+S_{4\_4}+S_{4\_5}$，形成如图 1c 所示的结构，计算图 1c 中所涉及的 3 种模体在图 1a 小网络中的数量，并将得到的模体 3_2、4_4 和 4_5 的这 3 种模体的数量作为机器学习方法的输入。还可以将所有 8 种模体的数量都作为特征值作为机器学习方法的输入，从而得到连边的相似度得分或连边权重。

3　科学家合作预测

3.1　数据说明

本文使用了常用的 4 个科学家合作网络进行链路预测与合作权重预测。

1) netscience 是一个从事网络理论和实验的科学家之间的加权合作网络，包括 1 461 个节点，2 742 条连边。其中，节点代表网络中的科学家，连边代表两位科学家有合作关系。

2) geom 是一个计算几何领域的科学家之间的加权合作网络，包括 6 158 个节点，11 898 条连边，权重代表两位科学家合作的次数[16]。

3) hepth 是 1995 年 1 月 1 日−1999 年 12 月 31 日期间在高能物理领域上发布预印本论文的科学家之间的加权网络，包括 7 610 个节点，15 751 条连边。

4) condmat 是 1995 年−1999 年在凝聚态物理领域上发布预印本论文的科学家之间的加权合作网络，包括 16 264 个节点，47 594 条连边。

数据 netscience、hepth 和 condmat 来自于参考文献 [17]，其中连接的权重代表科学家合作的强度[18]。其权重的具体计算公式为：

$$w_{ij}=\sum_k \frac{\delta_i^k \delta_j^k}{n_k-1}$$

式中，δ_i^k表示科学家i是否是论文k的共同作者，如果是记为 1，否则记为 0；n_k是论文k的共同作者数量。在此计算方法中，由于单个作者论文没有为合作网络做出贡献，所以将其排除。

为了比较权重预测问题时不同网络的预测结果，所有链路权重均在区间 [0,1] 上进行归一化处理。具体的归一化方法为：

$$w^*=\mathrm{e}^{-\frac{1}{w}}$$

式中，w为原始权重值；w^*为归一化后的权重值。

3.2　XGBoost 机器学习方法

XGBoost 是 Extreme Gradient Boosting 的简称，它是 Gradient Boosting Machine 的一个 C++实现。XGBoost 最大的特点，它能够自动利用 CPU 的多线程进行并行，同时在算法上加以改进提高了精度[19]。XGBoost 是一种基于 GBDT 原理的改进算法，与普通的 GBDT 仅使用一阶导数信息不同，XGBoost 对损失函数做了二阶泰勒展开，并在目标函数中加入了正则项，减少过度拟合。除了与 GBDT 模型的理论差异外，XGBoost 还具有以下优势：速度快、可移植、少写代码、可容错。

本文利用 XGBoost 机器学习算法实现特征统计量的计算，实验中首先读取训练集和测试集的数据，将训练集的数据按 9∶1 的比例划分为训练集和验证集，并读取划分后的训练集、验证集和测试集的特征值；然后利用 XGBoost 机器学习方法建立基于训练集的模型，最终通过建立的模型实现测

试集中的链路和权重预测。

3.3 科学家合作链路预测

本文使用上述 4 个科学家合作网络进行实验验证。对于每个网络，选取 90% 的数据作为训练集 E_T，选取 10% 的存在边作为测试集 E_V 中的正样本，从不存在的边中去构建负样本，最终让测试集满足正负样本 1:1 的比例。然后基于单个模体特征和多模特特征 (所有 8 个模体) 进行链路预测，得到评价指标 AUC 的值，如表 2 所示，单个模体特征的最好预测性能和多模体特征的预测效果加粗标出。

表 2 基于模体特征的链路预测结果 (AUC)

模体编号	netscience	geom	hepth	condmat
3_1	0.867	0.785	0.595	0.841
3_2	**0.927**	**0.868**	**0.896**	**0.945**
4_1	0.736	0.831	0.579	0.609
4_2	0.893	0.790	0.715	0.853
4_3	0.890	0.762	0.542	0.877
4_4	0.557	0.483	0.606	0.606
4_5	0.661	0.637	0.617	0.697
4_6	0.818	0.738	0.710	0.831
多模体	**0.981**	**0.962**	**0.951**	**0.992**

由表 2 可以发现，使用单模体特征进行链路预测时，模体特征 3_2 的预测准确率最高。说明在科学家合作网络中，如果两位科学家同时与第三位科学家有合作，那么这两位科学家有合作的概率比较大。模体特征 3_2 从结构上看可以表示为计算节点的共同邻居数量，也是聚类系数的一种表达，与 WCN、WAA、WRA 和 rWCN、rWAA、rWRA 在拓扑结构上一样，说明上述方法仅仅是一种模体的加权形式。这类方法最大的缺陷是没有考虑到其他模体，即科学家合作的其他方式，本文综合多个模体特征进行预测，在表 2 中发现多模体特征的预测效果比单模体特征的最好预测效果高 5.0%～10.8%，说明综合科学家之间的多种合作模式进行链路预测效果更好。

在链路预测中，还将多模体特征链路预测的结果与基于共同邻居加权特征的预测方法 (WCN、WAA 和 WRA)，和基于可靠路线加权特征的预测方法 (rWCN、rWAA 和 rWRA) 进行了比较，其结果如表 3 所示，其中预测效果最好的方法加粗标出。

表 3 3 类方法的链路预测结果 (AUC)

数据名称	WCN	WAA	WRA	rWCN	rWAA	rWRA	多模体
netscience	0.933	0.917	0.915	0.933	0.933	0.933	**0.981**
geom	0.871	0.883	0.883	0.851	0.880	0.882	**0.962**
hepth	0.895	0.896	0.896	0.895	0.895	0.895	**0.951**
condmat	0.944	0.945	0.946	0.943	0.945	0.945	**0.992**

表 3 为使用 3 类不同方法进行链路预测的 AUC 结果对比，分析表 3 中的数据可以发现，多模体特征的预测准确率比共同邻居和可靠路线策略的最高预测准确率提高了 4.9%～8.9%。在与传统方法对比的基础上，以 netscience 网络为例，在 8 种模体特征中选取部分特征进行组合，然后进行链路预测，其结果如图 2 所示。通过图 2 可以发现预测效果最好的为多模体特征，说明在科学家合作网络中使用多模体特征 (即结合科学家之间多种合作模式)，进行链路预测能够有效提高预测的准确性。

在链路预测问题中，还对 8 种模体特征进行了皮尔逊相关性分析，结果如图 3 所示。从图 3 可以发现特征被分为两个不同的集合，第一个集合包括特征 3_1、4_3、4_2 和 4_1，它们之间有较强的相关性，是因为它们只关注待预测连边中两个节点的各自邻居节点之间的结构。第二个集合包括特征 3_2、4_6、4_5 和 4_4，它们之间同样具有比较强的相关性，是因为它们大多数关注待预测连边中节点的共同邻居之间的关系。

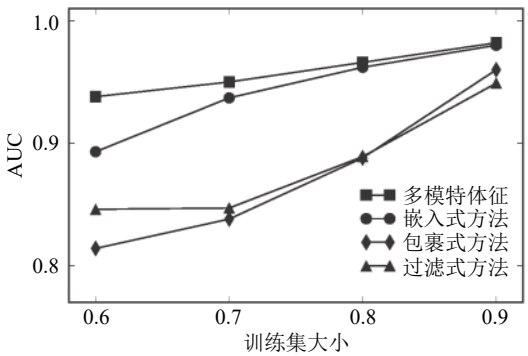


图 2 链路预测的特征选择方法性能比较

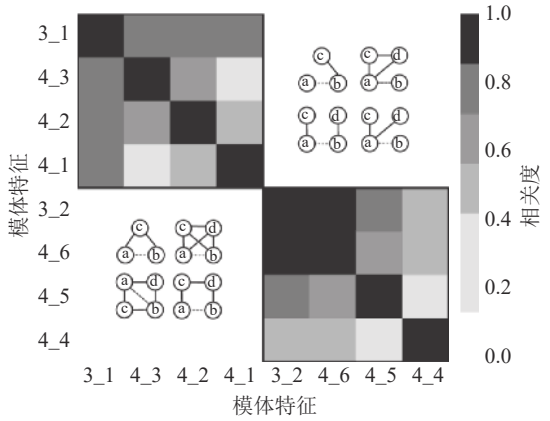


图 3　链路预测模体特征的相关性分析

3.4　科学家合作权重预测

除了使用上述 4 个网络进行链路预测算法比较，本文也使用这些数据进行合作权重预测。权重预测可以抽象为机器学习中的回归问题，在本文已知两位科学家有合作关系的前提下进行。对于每个网络，将数据集随机按 9:1 的比例分为训练集E_T和测试集E_V。然后基于单模体特征和多模体特征进行权重预测，得到评价指标 RMSE 的值，如表 4 所示。单模体特征的最好预测效果和多模体特征的最好预测结果已加粗标出。

通过表 4 可以发现，使用单模体特征进行合作权重预测时，其预测效果最好的单个模体特征在不同的科学家合作网络中是不一样的，在 netscience 和 hepth 两组数据中预测效果最好的单模体特征为特征 3_2，这与在链路预测中的结果是相同的。另两组数据 geom 和 condmat 中预测效果最好的单模体特征分别为特征 4_5 和特征 4_6。通过对模体特征拓扑结构的分析发现，这两个特征在拓扑结构上可以看作为特征 3_2 的组合，所以在权重预测问题中，预测效果最好的模体特征可以视为特征 3_2 以及该特征的组合。在每个网络中多模体特征的预测准确率比单模体特征的最好预测性能高 7.1%～25.2%，说明结合科学家多种合作模式进行权重预测效果更好。

表 4　基于模体特征的合作权重预测结果 (RMSE)

模体编号	netscience	geom	hepth	condmat
3_1	0.150	0.149	0.226	0.194
3_2	**0.111**	0.150	**0.203**	0.171
4_1	0.149	0.148	0.226	0.194
4_2	0.152	0.151	0.226	0.195
4_3	0.148	0.150	0.226	0.192
4_4	0.153	0.153	0.226	0.195
4_5	0.138	**0.141**	0.217	0.186
4_6	0.115	0.146	0.210	**0.160**
多模体	**0.083**	**0.131**	**0.184**	**0.132**

在合作权重预测中，本文同样将多模体特征的预测结果与基于共同邻居加权特征的预测方法 (WCN、WAA 和 WRA)，和基于可靠路线加权特征的预测方法 (rWCN，rWAA 和 rWRA) 进行了对比，结果如表 5 所示，其中最好的预测效果已加粗标出。

表 5　3 类方法的权重预测结果 (RMSE)

数据名称	WCN	WAA	WRA	rWCN	rWAA	rWRA	多模体
netscience	0.299	0.711	0.627	0.170	0.152	0.148	**0.083**
geom	2.458	0.923	0.339	0.336	0.338	0.324	**0.131**
hepth	0.521	0.526	0.763	0.307	0.272	0.271	**0.184**
condmat	0.470	0.691	0.535	0.213	0.180	0.178	**0.132**

表 5 表示的是使用 3 类不同方法进行权重预测的 RMSE 结果对比。分析表 5 中的结果可发现，多模体特征的合作权重预测准确率比其他两类预测方法的最高准确率提高了 25.8%～59.6%。

在权重预测中，同样以 netscience 网络为例，在 8 种模体特征中选取部分特征进行组合，然后进行权重预测，其结果如图 4 所示。通过图 4 可以发现预测效果最好的同样为多模体特征，说明使用多模体特征的权重预测方法可大幅提高其预测准确率。

在权重预测问题中，同样使用皮尔逊相关性分析法对 8 种模体特征进行了相关性分析，结果如图 5 所示。从图 5 可以看出模体特征 3_1、4_1、4_3 和 4_2 同样具有强相关性；此外，模体特征 3_2 和 4_6 具有强相关性，主要是因为是模体特征 4_6 是 3_2 拓扑结构特征的组合，模体特征 4_4 和 4_5 与其他特征之间都是相互独立的，几乎没有相关性。

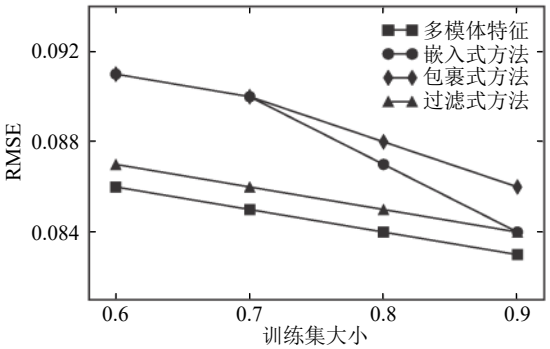


图 4 权重预测的特征选择方法性能比较

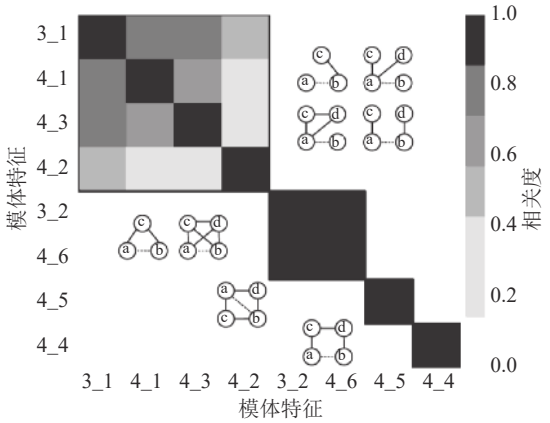


图 5 权重预测模体特征的相关性分析

3.5 链路预测与权重预测结果对比分析

分析表 3 和表 5 的数据发现，在合作链路预测和权重预测中，多模体特征预测方法的准确率比其他预测方法的准确率要高，说明结合多种科学家之间的合作模式进行链路和权重预测可有效提高预测的准确率。对比链路预测和权重预测两项预测的单模体特征预测结果 (表 2 和表 4)，可以发现在链路预测中，预测效果最好的单模体特征是模体特征 3_2。在权重预测中，预测效果最好的单模体特征分别为特征 3_2、4_5 和 4_6，其中模体特征 4_5 和 4_6 在拓扑结构上可以看作多个模体特征 3_2 的组合，所以综合以上结果，在使用单模体特征进行链路预测和权重预测时，预测效果最好为模体特征 3_2。从结构上看，模体特征 3_2 可以表示为计算节点的共同邻居数量，也是聚类系数的一种表达。基于共同邻居加权特征的预测方法和基于可靠路线加权特征的预测方法也是基于模体特征 3_2，说明所依赖的拓扑结构这两种方法是一致的。基于多模体的预测相对于上述两种方法，最大的优势是综合利用了多个模体特征。

4 结 束 语

本文提出了使用多模体特征进行科学家合作链路和权重预测的方法，涉及了 8 种不同的科学家之间的合作模式，用来解决经典预测方法在拓扑结构上仅仅利用的单一的科学家合作模式的问题，并在不同的网络上进行相关实验验证。实验结果表明，结合多个模体特征进行科学家合作预测可以有效提高预测的准确率，并可有效分析不同合作模式对于预测结果的影响。本文研究有助于预测出科学家合作的可能性及其合作权重，进而挖掘科研合作网络的结构特性对科研产出和科研影响力的影响。在后续研究中，将在模体特征的基础上引入权重信息，即利用模体的结构特征和权重信息进行更准确的预测。

参 考 文 献

[1] ZENG An, SHEN Zhe-si, ZHOU Jian-lin, et al. The science of science: From the perspective of complex systems[J]. Physics Reports, 2017(714-715): 1-73.

[2] FORTUNATO S, BERGSTROM C T, BÖRNER K, et al. Science of science[J]. Science, 2018, 359(6379): eaao0185.

[3] 刘岩, 刘亮, 罗天, 等. 基于子图的科学家合作网络家族辨识[J]. 科技管理研究, 2019, 39(7): 249-255.
LIU Yan, LIU Liang, LUO Tian, et al. Family identification of cooperative network of scientists based on subgraph[J]. Sciense and Technology Management Research, 2019, 39(7): 249-255.

[4] LÜ Lin-yuan, ZHOU Tao. Link prediction in complex networks: A survey[J]. Physica A Statistical Mechanics & Its Applications, 2011, 390(6): 1150-1170.

[5] ZHAO J, MIAO L, YANG J, et al. Prediction of links and weights in networks by reliable routes[J]. Scientific Reports, 2015, 5(1): 12261.

[6] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39(5): 651-661.
LÜ Lin-yuan. Link prediction on complex networks[J]. Journal of University of Electronic Science and Technology of China, 2010, 39(5): 651-661.

[7] LIBEN-NOWELL D, KLEINBERG J. The link-prediction problem for social networks[J]. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019-1031.

[8] ADANIC L A, ADAR E. Friends and neighbors on the web[J]. Social Networks, 2003, 25(3): 211-230.

[9] ZHOU T, LÜ L, ZHANG Y C. Predicting missing links via local information[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2009, 71(4): 623-630.

[10] LIU W, LÜ L. Link prediction based on local random walk[J]. EPL, 2010, 89(5): 58007.

[11] SA H R D, PRUDENCIO R B C. Supervised link prediction in weighted networks[C]// The 2011 International Joint Conference on Neural Networks. Piscataway, NJ: IEEE, 2011: 2281-2288.

[12] LÜ L, ZHOU T. Link prediction in weighted networks:

The role of weak ties[J]. EPL, 2010, 89(1): 18001.
[13] ZHU B, XIA Y, ZHANG X J. Weight prediction in complex networks based on neighbor set[J]. Scientific Reports, 2016, 6(1): 38080.
[14] MILO R, SHEN-ORR S, ITZKOVITZ S, et al. Network motifs: Simple building blocks of complex networks[J]. Science, 2002, 298(5594): 824-827.
[15] 刘亮. 复杂网络基元研究方法及应用[M]. 上海: 上海交通大学出版社, 2018.
LIU Liang. Complex network building blocks methods and applications[M]. Shanghai: Shanghai Jiao Tong University Press, 2018.
[16] BATAGELJ V, MRVAR A. Pajek datasets[EB/OL]. (2016-01-24). http://vlado.fmf.uni-lj.si/pub/networks/data/.
[17] NEWMAN M E J. The structure of scientific collaboration networks[J]. Proceedings of the National Academy of Sciences, 2001, 98(2): 404-409.
[18] NEWMAN M E J. Scientific collaboration networks. Ⅱ. Shortest paths, weighted networks, and centrality[J]. Physical Review E, 2001, 64(1): 016132.
[19] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2016: 785-794.

编　辑　蒋　晓