基于多模体边度的科学家合作关系预测

柳 娟 刘亚芳 许 爽 许小可
(大连民族大学信息与通信工程学院 辽宁 大连 116600)

摘 要 科学家合作关系预测近年来成为科学领域的热点研究方向,对于理解科学家之间的合作机制和科研网络的演化机理具有重要意义.但现有方法中对科学家合作关系的预测研究较少,且都是基于无向合作网络的预测.因此,本文构建了科学家合作有向网络,并在此基础上提出利用模体边度特征预测科学家合作关系.首先,针对传统方法中四节点模体无法使用朴素贝叶斯模型的难题,提出了一种单模体和双模体边度链路预测方法,为模体边度模型可进行链路预测的原因提供了理论解释.然后,提出了一种基于机器学习框架的多模体边度链路预测方法,并与现有预测方法的结果进行比较,该方法的预测性能提升了 5%~19%.最后,研究了 12 种模体边度特征之间的相关性,揭示了结构越相似的模体之间的预测结果相关性越强的规律.本文研究拓展了模体理论的应用场景,有助于进一步理解科学家有向合作网络的演化机制.

关键词 科学家合作关系;有向网络;模体边度;朴素贝叶斯;链路预测
中图法分类号 TP399 **DOI 号** 10.11897/SP.J.1016.2020.02372

《计算机学报》

Predicting Scientific Collaboration by Edge Degree of Multiple Motifs

LIU Juan LIU Ya-Fang XU Shuang XU Xiao-Ke
(*College of Information and Communication Engineering, Dalian Minzu University, Dalian, Liaoning* 116600)

Abstract In recent years, the prediction of scientific collaboration relationship has become a hot research topic in the field of science, because it is significant for understanding the cooperation mechanism among scientists and the evolution mechanism of scientific research networks. However, there are few researches on the prediction of scientific collaboration relationship using the methods based on complex networks, and they are almost based on the structures of undirected networks. In an undirected cooperative network, the direction of scientific collaboration is not considered, and this way assumes that each author's contribution to a paper is the same and the status of each author is equal. However, the collaboration of scientists is often not mutual and individual status is not equal in the actual collaboration, and this unbalanced relationship can be expressed by the direction of an edge. In the prediction of scientific collaboration relationship, it is impossible to distinguish scientists as the first author, corresponding author or ordinary author by the topology of undirected networks, which will lose key information such as scientific influence and future potential, and cause the deviation to scientific ranking and analysis of their influence in scientific fields. In addition, studying unequal cooperation in undirected networks will have a certain impact on the understanding of scientific research cooperation. In this study, we construct directed collaboration networks considering the direction of the edge between scientists, and then we try to predict the collaboration relationship in these directed networks.

收稿日期:2019-11-08;在线发布日期:2020-05-02.本课题得到国家自然科学基金(61773091,61603073)、辽宁省重点研发计划指导计划项目(2018104016)、辽宁省"兴辽英才"计划项目(XLYC1807106)、辽宁省高等学校创新人才支持计划(LR2016070)资助.**柳 娟**,硕士研究生,主要研究方向为社交网络分析和链路预测.E-mail:2426852886@qq.com.**刘亚芳**,硕士,主要研究方向为社交网络分析和数据可视化.**许 爽**,博士,副教授,主要研究方向为大数据分析与处理.**许小可**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为复杂网络社团检测、链路预测和数据挖掘.E-mail:xuxiaoke@foxmail.com.

The prediction of scientific collaboration relationship can be abstracted into a link prediction problem of directed networks in the field of network science. Link prediction based on network structure information can be divided into two kinds of methods: global information and local structure similarity. At present, the most widely used local structure algorithm is the motif-based method for link prediction in directed networks. However, some existing methods only consider the influence of one or two specific motifs on link prediction instead of using a lot of information of multiple motifs for link prediction. At the same time, few researchers do consider the contribution difference of nodes. Actually, the contribution of each node is often different in a real-life social network. In view of the limitations of the existing methods of link prediction in directed networks, we attempt to predict scientific collaboration relationship based on edge degrees of motif in directed scientific collaboration networks, and the aim is to predict the possibility of cooperation between scientists. Firstly, a link prediction method based on edge degrees of single and dual motifs is proposed in order to solve the problem that the Naïve Bayesian model cannot be used in the traditional four-node motif. Meanwhile, this study uncovers the intrinsic mechanism using edge-dependent motif for link prediction. Then, the link prediction method of multiple motifs using a machine learning framework is proposed. Compared with existing prediction methods, the prediction performance of the proposed method is improved 5%～19%. Finally, the correlation between the twelve kinds of motifs has been studied, uncovering that the more similar the structure, the stronger the correlation between motifs. This study expands the application scenarios of motif theory and can help us to further understand the evolution mechanism of directed collaboration networks of scientists.

Keywords　scientific collaboration relationship; directed network; motif-edge degree; Naïve Bayes; link prediction

《计算机学报》

1　引　言

近年来，随着科学研究的迅速发展以及数据分析技术的广泛应用，基于数据分析的“科学学”研究成为国内外的重要研究方向[1-2]. 由于科学家合作网络是科研活动组织与科学信息传播的结构基础，对知识的创造和传播具有重要意义，因此受到学者们的广泛关注[3]，分析和预测科学家社交网络中的合作关系也成为一项颇具价值的研究课题.

科学家合作网络是一种特殊的社交网络，在网络中将科学家视作网络的节点，科学家之间的合作关系视作连边[4-5]. 如果两位科学家至少合作过一次，则认为这两位科学家之间存在连边. 目前对科学家合作网的研究几乎都是基于无向网络的. 然而，在无向网络中，无法区分科学家谁是第一作者、通讯作者或普通作者，这样就会损失掉科学家的影响力和未来潜力等关键信息，对科学家的排名和分析其学科影响力造成偏差. 此外，构建无向网络的方式是假设每个作者对文章的贡献是相同的，而每篇文章中作者所起的作用往往是不平等的，这种不平等的关系会对理解科研合作方式产生一定的影响.

针对构建科学家合作无向网络存在的上述问题，曾安等假设第一作者是文章的主要贡献者，并在此基础上构建了有向科学家合作网络，基于节点重要性理论分析了科学家在该类型网络中所起的作用[6]. 鉴于该方法取得了很好的效果且受到网络科学领域学者的关注，本研究中假设第一作者是文章的主要贡献者，通过对 6 种世界著名期刊近 20 年的合作数据构建了有向科学家合作网络，并提出了利用模体边度特征进行科学家合作关系预测，旨在预测科学家之间合作的可能性.

对科学家合作关系的预测研究可以抽象为网络科学领域中的链路预测问题. 链路预测作为复杂网络的一个重要研究方向，近年来受到较多关注，可应用于网络重构、网络演化模型评价、推荐系统、社团发现等实际场景[7]. 链路预测是指通过已知的部分网络结构信息，预测网络中任意两个节点之间存在

连接的可能性.这种预测既包含了对静态链接的预测,即网络中存在但尚未被发现的链接的预测;也包含了对未来链接的预测,即目前网络中不存在但未来可能存在链接的预测[8-9].

近年来,学者们提出了很多种链路预测方法.Liben-Nowell 和 Kleinberg 发现基于节点共同邻居相似性是预测性能最好的局域结构方法之一,并分析了其在若干社交网络中链路预测的效果[10].周涛等使用 9 种基于局部信息的指标对多种实证网络的预测准确性进行比较,在此基础上提出了准确性更高的资源分配指标(Resource Allocation,RA)和局部路径指标(Local Path,LP)[11].Cannistraci 等人认为相连的两个节点的共同邻居会倾向于形成局部社团结构,可利用该特征来刻画共同邻居的连接紧密程度,从而提高链路预测性能[12].Grover 等人提出了一种基于图表示学习的链路预测方法(node2vec),与基于网络结构的链路预测基准算法相比,该方法可有效提升预测准确性[13].Kovács 等人在蛋白质相互作用网络中进行链路预测,提出充分利用长度为 3 的路径信息新算法,其性能明显优于现有链路预测方法[14].

以上方法均简单地认为,某一指标相同的节点的链路预测贡献相同,但是在实际的社交网络中,节点的贡献值往往是有差别的.Liu 等人提出了一种基于共同邻居的朴素贝叶斯模型预测方法,考虑了每个节点贡献值对预测结果的影响,该方法在一定程度上可以提高预测准确度[15].Wu 等人提出了加权局部朴素贝叶斯概率模型,将预测节点对的共同邻居的权值作为角色函数考虑到链路预测方法中,发现可提高加权网络中链路预测的准确度[16].

针对节点贡献不同,学者们在无向网络研究中使用了朴素贝叶斯模型,并在链路预测中增加了角色函数,取得了较好的预测结果.有向网络是一种连边具有方向性的网络,上述方法只利用了无向连边的信息而没有考虑链路的方向性,因此不能直接拓展到有向网络中.在很多实证研究中,均发现个体之间的作用往往不是相互的,个体地位也是不平等的,这种不均衡关系可以通过有向网络中连边的方向性来表示.因此,将这种类型网络简化成无向网络会损失其中的有用信息,有向网络链路预测的关键点是要考虑节点之间连边的方向性,而现有的链路预测方法中仅仅有少量方法考虑了这一点.

目前针对有向网络的链路预测,应用最多和较为有效的方法就是以势理论为基础的模体方法[17-19].Milo 等人首先提出了网络模体的概念[20-21],模体是网络的微观结构,即真实网络中频繁出现的由少数个体组成的小规模同构子图,其在真实网络中的出现频率远高于在具有相同节点和边数的随机网络中的出现频率[3-4].韩华等人在传统的顶点度和边聚类系数基础上,提出了基于模体的顶点度和边度来衡量网络中顶点和边的重要性[22].张千明提出了势理论,发现满足势理论的模体结构具有更好的链路预测效果[23].Hu 等人考虑了模体的局部信息,即节点的出入度,提出了基于四节点模体的 QMI 方法,与常用的预测方法相比,该方法能够提升预测精度[17].但是以上方法都只考虑了一两种特定模体的链路预测效果,而没有综合多个模体的多种信息进行链路预测.这类方法是一种直觉性的方法,没有相关理论去解释这类方法预测准确性的原因,而且它们也没有考虑不同节点贡献的差异性.

鉴于现有研究的局限性,本文通过构建 6 种世界著名期刊有向科学家合作网络,首先考虑节点贡献对有向网络链路预测的影响,提出基于朴素贝叶斯模型的单模体、双模体边度链路预测方法.朴素贝叶斯模型主要由两部分组成:预测边形成的模体的数量(模体边度)和与预测边构成预测器的节点的总贡献.当节点贡献对预测效果的影响不大时,可以忽略该部分影响而只使用模体边度,这样就为应用模体数量进行链路预测提供了理论依据.同时,本文提出了基于机器学习框架的多模体边度预测方法,该方法在考虑连边的方向性的同时,综合考虑多个模体特征对预测效果的影响,取得了更好的预测效果.最后,采用最大信息系数分析了多模体边度特征之间的相关性,揭示与科学家之间的合作模式相对应的每种模体之间的内在联系.本文的研究能够促进对科学家合作网络演化机制的理解,也可应用于其它类型有向网络的研究.

本文的创新点和主要贡献如下:

(1)将朴素贝叶斯模型和模体边度理论应用到有向网络的链路预测当中,不仅考虑了模体的数量,还考虑了与预测边构成预测器的节点的角色函数对预测的影响,尤其采用一种新方法解决了四节点模体无法使用朴素贝叶斯模型的问题.

(2)基于朴素贝叶斯理论为使用模体边度可进行链路预测的原因提供理论解释.我们的理论推导发现此类方法不仅对于单模体预测器有效,对于双模体预测器也有效.通过实验证明节点的角色函数在六种有向科学家实证网络中作用不大,因此在这

《计算机学报》

类网络中可以仅使用模体边度而忽略节点的角色函数.

(3)本文提出基于机器学习的多模体有向网络链路预测方法,与已有链路预测方法进行比较,实验表明该方法的预测精度最高,约提升了 5%～19%.此外,用最大信息系数分析了多模体边度特征之间的相关性和模体预测器的解释性,揭示与科学家之间的合作模式相对应的每种模体之间的内在联系.

2　实验数据与理论基础

2.1　网络数据说明

本文下载了 Web of Science 中六本著名期刊数据,包括《科学》(Science)、《自然》(Nature)、《新英格兰医学杂志》(NEJM)、《柳叶刀》(Lancet)、《美国医学会杂志》(JAMA)和《英国医学期刊》(BMJ).其中,Nature 包含了 1998 年～2018 年的全部数据,其他 5 本期刊包含了 1984 年～2019 年的全部数据.在这些期刊中设有各种不同的栏目,考虑到科学家合作主要是在研究性学术论文中进行的,因此我们仅考虑 Article 和 Review 这 2 种文献类型.

研究中将每一本期刊构建成该期刊作者间的有向科学家合作网络,将第一作者作为源节点,其他作者作为目标节点,由此形成从第一作者指向普通作者的有向边.在每一个网络中,如果科学家之间至少合著过一篇论文,则认为科学家之间存在连边.在该网络中,节点表示科学家,连边表示科学家之间的论文合著关系.上述网络均可以表达为 $G=(V,E)$,其中 V 代表网络中节点的集合,E 表示网络中连边的集合.网络中的链接关系用邻接矩阵来描述,矩阵元素表示节点之间连接与否,其记作 a_{ij},若 $a_{ij}=1$ 表示节点之间存在连边,$a_{ij}=0$ 表示节点之间不存在连边.本文基于以上方法构建了 6 个有向科学家合作网络,对于每个网络仅考虑其最大连通集团,所得网络的节点数和连边数统计数据如表 1 所示.

表 1　6 种有向科学家合作网数据说明

网络	节点数	连边数
Science	1773	5566
Nature	1701	5874
NEJM	497	1568
Lancet	1179	5117
JAMA	597	1748
BMJ	643	1589

2.2　链路预测的评价标准

本文使用的评价指标是 *AUC*、*Precision*、*Recall* 和 *F*1-score.链路预测方法对于四种评价指标的值越高,就说明该方法的性能越好.*AUC* 可以理解为在测试集中随机选择一条存在的边的分数值比随机选择一条不存在的边的分数值高的概率.也就是说,每次从测试集中随机选取一条存在的边,然后随机选取一条不存在的边,如果存在的边的分数值大于不存在的边的分数,就加 1 分;如果两个分数值相等,就加 0.5 分.这样独立比较 n 次,如果有 n' 次测试集中存在的边的分数值大于不存在的边分数,有 n'' 次两个分数值相等[24],那么 *AUC* 可以定义为

$$AUC=\frac{n'+0.5n''}{n} \tag{1}$$

AUC 能够从整体上衡量链路预测的准确性,而 *Precision* 可以从局部衡量预测准确性.该指标定义为在预测值排序在前 L 个预测边中预测准确(真实存在边)的比例[25].将特征得分从大到小排序,如果排序在前 L 的预测边中有 m 条真实边存在,那么 *Precision* 可以定义为

$$Precision=\frac{m}{L} \tag{2}$$

在本研究中,由于 L 取值并不影响本文的实验结论,因此根据 6 种真实网络的测试集连边数量不相同,统一选择了测试集连边数量的 20%作为 L 的值.

除了 *AUC* 和 *Precision* 外,*Recall* 和 *F*1-score 也是非常重要的性能评价指标.*Recall* 用以衡量所有存在的连边中被预测为存在的连边的比例,可以定义为

$$Recall=\frac{TP}{TP+FN} \tag{3}$$

*F*1-score 又称调和平均数,用以调和 *Recall* 和 *Precision*.需要注意的是,这里的 *Precision* 与上文的评价指标 *Precision* 不同,这里的 *Precision* 即精确率,用以衡量所有预测为存在的连边中真正存在的连边的比例,因此 *F*1-score 可以定义为

$$\begin{aligned}F1\text{-score}&=\frac{2\cdot Precision\cdot Recall}{Precision+Recall}\\&=\frac{2TP}{2TP+FP+FN}\end{aligned} \tag{4}$$

其中,TP 表示真正例,指存在的连边被预测为存在连边的数量.FP 表示假正例,指不存在的边被预测为存在边的数量.FN 表示假反例,指存在的连边被预测为不存在的连边的数量.

2.3　有向网络的势理论

在有向网络中,当且仅当网络中的每个节点都

《计算机学报》

能被分配势能，且该势能是唯一可确定的，那么这个网络就是可定义势的. 也就是说，对网络中的任意一对节点 i 和 j 来说，如果存在一条边从节点 i 指向节点 j，即 $i\rightarrow j$，那么节点 i 的势能大于节点 j 的势能. 因此，一条有向边显然是可定义势的，而包含互惠边的结构一定是不可定义势的[25]. 本文仅考虑回路较小的模体，即考虑有向网络中 3 节点和 4 节点的情况下得到的 6 种模体结构[23,26]，如图 1 所示. 我们选择包含节点较少的模体有两方面原因：一是模体中含的节点数越少，那么它的数量就越容易计算；二是高阶模体的数量依赖于低阶模体的数量[27]，因此 5 阶和 5 阶以上的模体在链路预测的应用中已经不起主要作用[28]. 对于在这 6 种模体结构中，只有双风扇结构(Bi-fan)和双平行结构(Bi-parallel)是可定义势的.

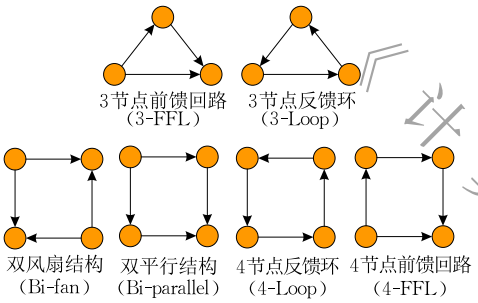


图 1 6 种含有回路的最小模体

在图 1 的 6 种模体中各取一条边，就能够得到 12 种预测器 $S_1\sim S_{12}$，如图 2 所示. 预测器中的虚线表示从原来的子图中移除的连边，即待预测的连边，其中，只有 S_5、S_6 和 S_7 是可定义势的. 此时对于虚线的待预测边，$S_1\sim S_{12}$ 每种模体的数量可以看成是包含该边的模体的边度[22]，因此本文所提的单模体链路预测方法可以理解为模体边度的一个实际应用. 韩华等人针对复杂网络拓扑结构中模体的存在性，在传统的顶点度和边聚类系数基础上，提出了基于模体的顶点度和边度来衡量网络中顶点和边的重要性，从网络复杂性度量的角度刻画了顶点和边在网络中的重要性，而目前还没有相关研究将链路预测与模体的顶点度和边度等概念联系起来.

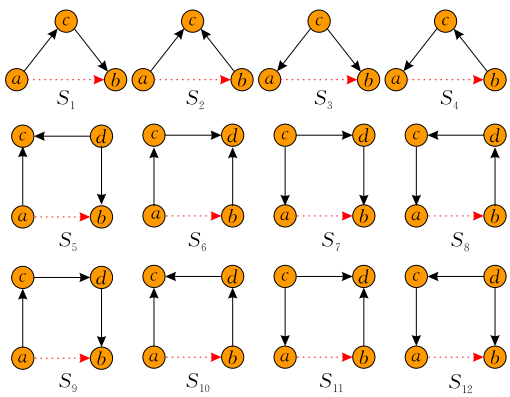


图 2 6 种模体对应 12 种模体边度预测器的图示

3 基于单模体边度的科研合作预测

3.1 单模体朴素贝叶斯模型

朴素贝叶斯的预测方法是将待预测边生成指定预测器的数量和预测器中每个节点对于连边形成的贡献综合在一起进行计算. 给定一个网络 $G=(V,E)$，其中 V 代表网络中节点的集合，E 表示网络中连边的集合，$|V|$ 表示网络中节点总数，$|E|$ 表示网络中连边总数. $\boldsymbol{E}^{\mathrm{T}}$ 表示训练集，$|\boldsymbol{E}^{\mathrm{T}}|$ 则表示训练集中的连边数量，E^P 表示测试集. 用变量 A_1 和 A_0 分别表示一对节点之间连接和不连接两种情况，根据训练集 $\boldsymbol{E}^{\mathrm{T}}$ 可以得到 A_1 和 A_0 的先验概率

$$P(A_1)=\frac{|\boldsymbol{E}^{\mathrm{T}}|}{|U|},\ P(A_0)=\frac{|U-\boldsymbol{E}^{\mathrm{T}}|}{|U|} \tag{5}$$

其中，$|U|=|V|(|V|-1)$ 表示网络中所有可能的连边的数量. 对每个节点 w，可以对它赋予两个条件概率 $P(w|A_1)$ 和 $P(w|A_0)$，其中，$P(w|A_1)$ 表示一对相连的节点与节点 w 生成指定模体预测器的概率，$P(w|A_0)$ 表示一对不相连的节点与节点 w 生成指定模体预测器的概率. 根据贝叶斯定理，分别计算这两个概率，表示为

$$P(w|A_1)=\frac{P(w)\cdot P(A_1|w)}{P(A_1)} \tag{6}$$

$$P(w|A_0)=\frac{P(w)\cdot P(A_0|w)}{P(A_0)} \tag{7}$$

其中，$P(w)$ 表示节点 w 和某节点对生成指定预测器的概率. $P(A_1|w)$ 表示与节点 w 生成指定预测器的节点对之间相连的概率，$P(A_0|w)$ 表示与节点 w 生成指定预测器的节点对之间不相连的概率，于是有

$$P(A_1|w)=\frac{N_{\Delta w}}{N_{\Delta w}+N_{\Lambda w}} \tag{8}$$

其中，$N_{\Delta w}$ 和 $N_{\Lambda w}$ 分别表示与节点 w 构成指定预测器节点对中有连接节点对数量和未连接节点对数量. 由 $P(A_1|w)+P(A_0|w)=1$，得

$$P(A_0|w)=1-P(A_1|w)=\frac{N_{\Lambda w}}{N_{\Delta w}+N_{\Lambda w}} \tag{9}$$

对于两个未知链接的节点 x 和 y，将与节点对 $\{x,y\}$ 生成指定预测器的所有节点的集合定义为

《计算机学报》

O_{xy}. 以图 2 中三节点模体 S_1 为例进行详细说明，当已知待预测边 ab 时，首先找到与节点 a 相连的所有目标节点，构成集合 A，再找到与节点 b 相连的所有源节点，构成集合 B，那么集合 A、B 的交集 C 就是与 ab 边构成该预测器的所有节点，也就是 O_{xy}. 要获取这个集合的目的有两点：一是要知道预测边所构成的指定预测器的数量，而预测边构成预测器的数量就是集合中节点或连边的数量；二是要计算与预测边构成预测器的节点或连边的角色函数，所以仅知道集合中节点和连边数量还不足以进行预测，重要的是需要知道该集合中的每一个节点或连边，以及它们对构成指定预测器的贡献. 假设集合中每一个节点对于节点 x 和 y 之间是否产生链接的贡献是相互独立的，根据朴素贝叶斯理论，得

$$P(A_1|O_{xy})=\frac{P(A_1)}{P(O_{xy})}\prod_{w\in O_{xy}}P(w|A_1) \tag{10}$$

$$P(A_0|O_{xy})=\frac{P(A_0)}{P(O_{xy})}\prod_{w\in O_{xy}}P(w|A_0) \tag{11}$$

其中，$P(A_1|O_{xy})$ 和 $P(A_0|O_{xy})$ 分别表示节点 x 和 y 之间相互连接和不连接的后验概率. 此时，给定一对节点，比较两个节点连接的概率 $P(A_1|O_{xy})$ 和不连接的概率 $P(A_0|O_{xy})$，就可以判断这两个节点之间产生连边的可能性. 为了更好地比较哪些连边更可能出现，可以定义节点对 x 和 y 的似然值为

$$\begin{aligned}r_{xy}&=\frac{P(A_1|O_{xy})}{P(A_0|O_{xy})}\\&=\frac{P(A_1)}{P(A_0)}\prod_{w\in O_{xy}}\frac{P(A_0)\cdot P(A_1|w)}{P(A_1)\cdot P(A_0|w)}\end{aligned} \tag{12}$$

在式(12)中，设 $s=\dfrac{P(A_0)}{P(A_1)}=\dfrac{|U|}{|\boldsymbol{E}^{\mathrm{T}}|}-1$，对于给定的网络和测试集，$s$ 可视为一个常数；设 $R_w=\dfrac{P(A_1|w)}{P(A_0|w)}=\dfrac{N_{\Delta w}}{N_{\Lambda w}}$ 为节点 w 的角色函数，用来刻画节点 w 对于两节点产生连接和不产生连接的贡献比. 这里需要注意的是，当与节点 w 构成预测器的节点对中未连接的节点对数量为 0，即 $N_{\Lambda w}=0$，那么 R_w 的分母就会为 0 而导致计算没有意义，于是，将角色函数 R_w 中的分子分母都做加 1 处理，即 $\widetilde{R_w}=\dfrac{N_{\Delta w}+1}{N_{\Lambda w}+1}$. 于是节点对 x 和 y 的似然值为

$$r_{xy}=s^{-1}\prod_{w\in O_{xy}}s\,\widetilde{R_w} \tag{13}$$

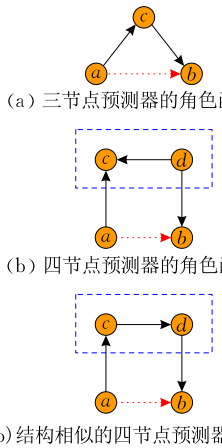
此模型为单模体朴素贝叶斯模型. 由于 s^{-1} 为一个常数，所以可以不考虑它的作用，于是式(13)取对数后得

$$r'_{xy}=|O_{xy}|\log s+\sum_{w\in O_{xy}}\log\widetilde{R_w} \tag{14}$$

3.2　基于单模体边度的链路预测

在基于单模体进行链路预测时，用 r'_{xy} 表示节点对 x 和 y 的特征分数值，根据此分数值得到链路预测的评价指标. 对于未知链接 $\{x,y\}$ 来说，式(14)中第一部分 $|O_{xy}|\log s$ 正比于该链接能够生成的对应模体的数量(模体边度)，$\sum_{w\in O_{xy}}\log\widetilde{R_w}$ 表示与 x 和 y 一起能够生成对应模体的所有节点的角色函数的总贡献. 如果不区分每个节点的贡献，则仅需要保留公式的第一部分，该部分是一种简化的基于朴素贝叶斯的单模体链路预测方法，也就是基于单模体边度[22]的预测方法，它是能够利用模体进行链路预测的理论基础.

在以前基于朴素贝叶斯模型的链路预测器中，都仅仅考虑了三节点模体中节点的角色，四节点模体中除了预测边外含有两个节点，比较复杂而无法使用朴素贝叶斯模型. 图 3 是三节点和四节点预测器角色函数的计算示意图. 对于三节点预测器，需要考虑的是除了预测边 ab 之外的节点 c 对于该预测器的影响. 对于四节点预测器，还需要考虑除了预测边 ab 之外的节点 c 和节点 d 对该预测器的影响，我们提出可以按照如下三种情况进行考虑：(1) 仅考虑节点 c 对该预测器的影响；(2) 仅考虑节点 d 对该预测器的影响；(3) 考虑节点 c 和节点 d 及 cd 之间的连边所构成的整体对该预测器的影响. 由于前两种方式只考虑了预测边之外的部分结构，忽略了整体结构的影响，所以对于四节点预测器，本文使用第三种方法计算角色函数，考虑预测边之外的所有结构对该预测器的影响，即四节点预测器中虚线框内的结构. 如果将(b)、(c)中两个四节点预测器的计



(a) 三节点预测器的角色函数计算

(b) 四节点预测器的角色函数计算

(c) 与(b)结构相似的四节点预测器的角色函数计算

图 3　三节点和四节点预测器的角色函数计算

《计算机学报》

算角色函数部分看成一个整体，那么两个预测器的结构与(a)中三节点预测器的结构相似. 同时，这种方式意味着式(14)这种直接对三节点模体有效的计算方法可以直接应用到更高阶(4 阶)模体的链路预测上.

针对 BMJ 网络的单模体链路预测的结果如图 4 所示：图中三角形表示直接使用式(14)，基于朴素贝叶斯的单模体链路预测结果；而圆形表示只使用式(14)中第一部分，不考虑节点角色函数的简化单模体链路预测结果，即基于单模体边度的预测结果. 对比基于朴素贝叶斯模型的单模体链路预测和简化版的结果，发现基于两种预测方法的 12 种模体预测器的性能基本是相同的，说明对于科学家合作网络而言，节点角色函数对其合作关系的预测影响不大，于是式(14)可以近似简化为

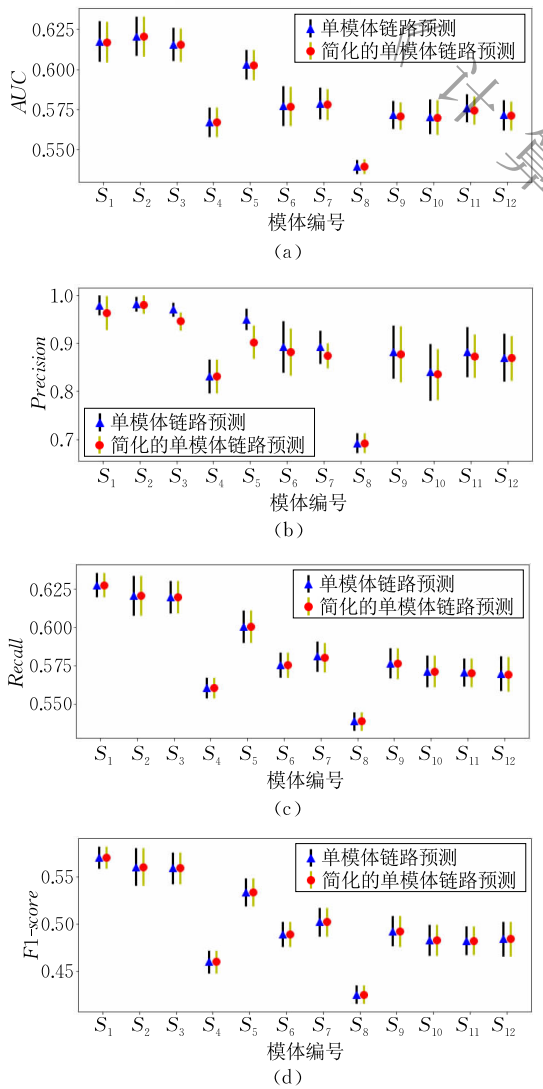


图 4 基于朴素贝叶斯的单模体链路预测与简化版预测器针对 BMJ 网络的结果比较

$$r'_{xy} \approx |O_{xy}| \log s \tag{15}$$

因此，我们在后续的科学家合作关系预测研究中将不再研究节点角色函数的影响.

4 基于多模体边度的科研合作关系预测

4.1 双模体朴素贝叶斯模型

对于未知链接$\{x,y\}$，用 e_{xy} 表示两个节点之间存在连边，$\overline{e_{xy}}$ 表示两个节点之间不存在连边. 定义 $O_1(x,y)$ 为与两个节点生成第一种预测器的所有节点集合，$O_2(x,y)$ 为与两个节点生成第二种预测器的所有节点集合，两个集合的获取方式与第 3 节中 O_{xy} 的获取方式相同，那么两个节点之间相互连接和不连接的后验概率为

$$\begin{aligned} &P(e_{xy}|O_1(x,y),O_2(x,y)) \\ &=\frac{P(e_{xy})\cdot P(O_1(x,y),O_2(x,y)|e_{xy})}{P(O_1(x,y),O_2(x,y))} \\ &=\frac{P(e_{xy})\cdot P(O_1(x,y)|e_{xy})\cdot P(O_2(x,y)|e_{xy})}{(O_1(x,y),O_2(x,y))} \end{aligned} \tag{16}$$

$$\begin{aligned} &P(\overline{e_{xy}}|O_1(x,y),O_2(x,y)) \\ &=\frac{P(\overline{e_{xy}})\cdot P(O_1(x,y),O_2(x,y)|\overline{e_{xy}})}{P(O_1(x,y),O_2(x,y))} \\ &=\frac{P(\overline{e_{xy}})\cdot P(O_1(x,y)|\overline{e_{xy}})\cdot P(O_2(x,y)|\overline{e_{xy}})}{(O_1(x,y),O_2(x,y))} \end{aligned} \tag{17}$$

此时，对于节点 x 和 y 来说，比较它们之间相互连接的概率和不连接的概率，就可以判断两个节点之间是否有更大的可能性产生连边. 为了更好地比较哪些连边更可能出现，可以用这两个概率的比值为每个节点对计算一个数值

$$\begin{aligned} r_{xy} &= \frac{P(e_{xy}|O_1(x,y),O_2(x,y))}{P(\overline{e_{xy}}|O_1(x,y),O_2(x,y))} \\ &= \frac{P(e_{xy})}{P(\overline{e_{xy}})}\cdot\frac{P(O_1(x,y)|e_{xy})}{P(O_1(x,y)|\overline{e_{xy}})}\cdot\frac{P(O_2(x,y)|e_{xy})}{P(O_2(x,y)|\overline{e_{xy}})} \end{aligned} \tag{18}$$

假设集合 $O_1(x,y)$ 和 $O_2(x,y)$ 中的每一个节点对于节点对之间产生链接与否的贡献是相互独立的，那么 r_{xy} 可以化简为

$$\begin{aligned} r_{xy} = &\frac{P(e_{xy})}{P(\overline{e_{xy}})}\prod_{w\in O_1(x,y)}\frac{P(\overline{e_{xy}})\cdot P(e_{xy}|w)}{P(e_{xy})\cdot P(\overline{e_{xy}}|w)} \\ &\prod_{v\in O_2(x,y)}\frac{P(\overline{e_{xy}})\cdot P(e_{xy}|v)}{P(e_{xy})\cdot P(\overline{e_{xy}}|v)} \end{aligned} \tag{19}$$

其中，$P(e_{xy}|w)$ 表示在构成的第一种预测器时，与节点 w 构成预测器的节点对之间相互连接的概率，

$P(e_{xy}|v)$表示在构成的第二种预测器时，与节点 v 构成预测器的节点对之间相互连接的概率. 相应地，$P(\overline{e_{xy}}|w)$和 $P(\overline{e_{xy}}|v)$则表示节点对之间不连接的概率，因此有

$$P(e_{xy}|w)=\frac{N_{\Delta w}}{N_{\Delta w}+N_{\Lambda w}} \tag{20}$$

$$P(\overline{e_{xy}}|w)=1-P(e_{xy}|w)=\frac{N_{\Lambda w}}{N_{\Delta w}+N_{\Lambda w}} \tag{21}$$

$$P(e_{xy}|v)=\frac{N_{\Delta v}}{N_{\Delta v}+N_{\Lambda v}} \tag{22}$$

$$P(\overline{e_{xy}}|v)=1-P(e_{xy}|v)=\frac{N_{\Lambda v}}{N_{\Delta v}+N_{\Lambda v}} \tag{23}$$

那么，可以得到两个预测器的节点角色函数为

$$R_w=\frac{N_{\Delta w}}{N_{\Lambda w}},\ R_v=\frac{N_{\Delta v}}{N_{\Lambda v}} \tag{24}$$

为了防止分母为 0 没有意义，将分子分母都加 1，于是

$$\widetilde{R}_w=\frac{N_{\Delta w}+1}{N_{\Lambda w}+1},\ \widetilde{R}_v=\frac{N_{\Delta v}+1}{N_{\Lambda v}+1} \tag{25}$$

定义 $s=\dfrac{P(\overline{e_{xy}})}{P(e_{xy})}=\dfrac{|U|}{|\boldsymbol{E}^{\mathrm{T}}|}-1$，那么对于节点对$\{x,y\}$来说

$$r_{xy}=s^{-1}\prod_{w\in O_1(x,y)}s\widetilde{R}_w\prod_{v\in O_2(x,y)}s\widetilde{R}_v \tag{26}$$

其中，s^{-1}为一个常数，所以不考虑，于是将式(26)剩余部分取对数得

$$r'_{xy}=(|O_1(x,y)|+|O_2(x,y)|)\log s+\sum_{w\in O_1(x,y)}\log\widetilde{R}_w+\sum_{v\in O_2(x,y)}\log\widetilde{R}_v \tag{27}$$

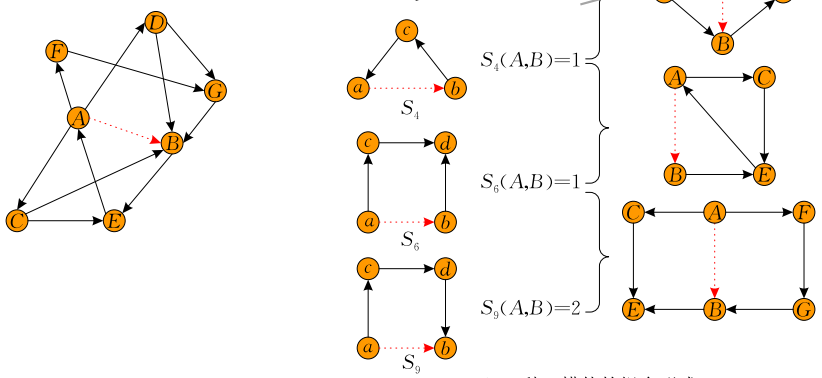
该式中第一部分是两种模体的数量之和，第二部分是构成第一种模体的所有节点的角色函数影响力之和，第三部分是构成第二种模体的所有节点的角色函数影响力之和. 由于节点角色函数对科学家合作有向网络的链路预测准确性影响不大，因此式(27)可以近似为

$$r'_{xy}=(|O_1(x,y)|+|O_2(x,y)|)\log s \tag{28}$$

在该式中仅需要计算节点 x 和 y 及其邻居节点构成的两种模体的数量之和.

4.2　基于双模体边度的链路预测

基于双模体边度的链路预测就是将两个单模体的模体个数相加. 如图 5 所示，(a)表示一个小型的网络示例，AB 表示待预测的连边. (b)表示在(a)的小型网络中，预测边 AB 可以生成 2 个预测器 S_1，分别为 ABC 和 ABD；可以生成 1 个预测器 S_4，为 ABE；可以生成 1 个预测器 S_6，为 $ABCE$；也可以生成 2 个预测器 S_9，分别为 $ABFG$ 和 $ABDG$. 同时图 5(b)也列出了三种双模体组合形式，分别为两个三节点模体组合 S_1+S_4、三节点和四节点模体组合 S_4+S_6 和两个四节点模体组合 S_6+S_9.



(a) 小型网络的拓扑结构图　　(b) 3种双模体的组合形式

图 5　双模体链路预测图示

科学家合作关系网络基于图 5 中双模体链路预测的结果如表 2 所示，双模体边度链路预测的准确性比单模体边度的预测准确性在一定程度上有所提高. 但是这种简单地将两种单模体的个数相加的预测方法也会造成误差的叠加，因此表 2 中并不是任意两个模体组合形成的双模体预测器相对于单个模体都有更好的预测效果. 如果将这种方法直接应用到基于更多模体的链路预测中，会导致更大误差的叠加效应，导致最终的预测结果不准确，所以将这种方法直接应用到多模体链路预测中有一定困难.

计算机学报

表 2 双模体边度的链路预测结果

网络及评价指标		S_1	S_4	S_6	S_9	S_1+S_4	S_4+S_6	S_6+S_9
Science	*AUC*	0.695	0.602	0.680	0.682	**0.726**	0.702	0.722
	Precision	0.959	0.973	0.968	0.977	**0.986**	0.982	0.982
	Recall	0.672	0.606	0.668	0.658	**0.674**	0.666	0.665
	*F*1-score	0.636	0.535	0.632	0.615	**0.636**	0.629	0.628
Nature	*AUC*	0.667	0.583	0.661	0.667	0.700	0.691	**0.702**
	Precision	0.983	0.936	0.970	0.983	**0.996**	0.983	0.991
	Recall	0.657	0.586	0.653	0.660	**0.662**	0.644	0.640
	*F*1-score	0.617	0.505	0.615	**0.624**	0.615	0.604	0.601
NEJM	*AUC*	0.711	0.641	0.698	0.701	**0.784**	0.760	0.741
	Precision	0.903	0.887	0.935	0.935	0.935	0.903	**0.952**
	Recall	0.685	0.637	0.673	**0.687**	0.682	0.674	0.668
	*F*1-score	0.654	0.587	0.642	**0.661**	0.652	0.645	0.639
Lancet	*AUC*	0.750	0.635	0.753	0.750	**0.786**	0.780	0.777
	Precision	0.995	0.946	0.985	0.990	**0.995**	0.990	0.990
	Recall	0.660	0.638	0.632	0.627	**0.660**	0.646	0.642
	*F*1-score	0.617	0.586	0.578	0.568	**0.617**	0.606	0.602
JAMA	*AUC*	0.670	0.596	0.642	0.660	**0.708**	0.682	0.693
	Precision	0.957	0.942	0.942	0.913	**0.986**	0.942	0.942
	Recall	**0.673**	0.605	0.655	0.650	0.670	0.665	0.660
	*F*1-score	0.634	0.535	0.615	0.608	**0.637**	0.629	0.626
BMJ	*AUC*	0.614	0.574	0.558	0.569	**0.670**	0.610	0.598
	Precision	0.937	0.825	0.889	0.857	**0.937**	0.905	0.921
	Recall	**0.634**	0.571	0.580	0.588	0.632	0.629	0.628
	*F*1-score	0.580	0.476	0.497	0.509	**0.580**	0.576	0.575

4.3 基于多模体边度的链路预测

多模体边度的链路预测就是综合利用多个模体特征，并基于 XGBoost 机器学习框架实现的. XGBoost 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[29]. 不同于传统梯度提升决策树(Gradient Boosted Decision Trees，简记为 GBDT)在优化时仅用一阶导数信息，XGBoost 对损失函数进行二阶泰勒展开，并在目标函数中加入了正则项，整体求最优解，用以权衡目标函数和模型的复杂程度，防止过拟合[30]. 除理论与传统的 GBDT 存在差别外，XGBoost 具有速度快、可移植、代码较少、可容错的优点.

本文研究中划分训练集与测试集的比例为 8∶2，通过将 12 种预测器在单模体链路预测过程中得到的训练集中连边的分数值(即模体的边度)作为特征，利用 XGBoost 进行训练学习，得到测试集中连边的相似度得分，根据此分数求得四种评价指标的值，得到预测结果. 使用 XGBoost 方法将 12 种预测器综合起来进行预测，实验结果如表 3 所示，在 6 种有向科学家合作网络中，基于所有模体特征进行机器学习的链路预测能力与单个模体的预测能力相比都有较大提升.

表 3 单模体和多模体的链路预测结果

网络及评价指标		S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	All
Science	*AUC*	0.678	0.681	0.654	0.601	**0.711**	0.664	0.645	0.599	0.670	0.668	0.646	0.645	**0.816**
	Precision	0.968	0.955	0.986	0.981	**0.986**	0.968	0.981	0.932	0.977	0.977	0.972	0.977	**0.990**
	Recall	**0.680**	0.683	0.668	0.606	0.664	0.673	0.650	0.611	0.673	0.666	0.656	0.648	**0.792**
	*F*1-score	0.644	0.648	0.629	0.535	0.621	**0.656**	0.605	0.548	0.636	0.629	0.613	0.603	**0.786**
Nature	*AUC*	0.662	0.653	0.649	0.581	**0.720**	0.664	0.645	0.589	0.652	0.661	0.638	0.644	**0.824**
	Precision	0.978	0.965	0.927	0.935	**0.983**	0.952	0.961	0.901	0.965	0.944	0.935	0.957	**0.987**
	Recall	0.671	0.660	0.640	0.578	0.647	0.663	0.643	0.576	0.654	0.644	0.635	**0.678**	**0.790**
	*F*1-score	**0.632**	0.619	0.588	0.490	0.596	0.628	0.601	0.511	0.612	0.605	0.592	0.592	**0.783**
NEJM	*AUC*	0.690	0.679	0.698	0.610	**0.743**	0.688	0.700	0.612	0.704	0.682	0.684	0.676	**0.851**
	Precision	0.887	0.854	0.854	0.855	**0.974**	0.870	0.935	0.919	0.951	0.951	0.887	0.935	**0.983**
	Recall	0.681	0.686	0.671	0.625	0.684	0.693	0.677	0.649	0.692	**0.703**	0.681	0.687	**0.827**
	*F*1-score	0.646	0.654	0.633	0.570	0.655	0.669	0.647	0.609	0.668	**0.680**	0.651	0.657	**0.824**
Lancet	*AUC*	0.768	0.754	0.679	0.624	**0.791**	0.767	0.697	0.640	0.770	0.761	0.684	0.687	**0.894**
	Precision	0.995	0.975	0.980	0.936	**0.995**	0.975	0.975	0.951	0.995	0.985	0.950	0.980	**0.996**
	Recall	**0.680**	0.676	0.605	0.630	0.635	0.642	0.578	0.577	0.644	0.645	0.572	0.576	**0.857**
	*F*1-score	**0.644**	0.640	0.533	0.574	0.581	0.592	0.488	0.487	0.593	0.596	0.479	0.483	**0.855**

《计算机学报》

（续　表）

网络及评价指标		S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	All
JAMA	*AUC*	0.680	0.666	0.668	0.592	**0.699**	0.644	0.648	0.592	0.651	0.635	0.630	0.643	**0.830**
	Precision	0.913	0.957	0.928	0.942	**0.966**	0.957	0.957	0.928	0.957	0.942	0.942	0.942	**0.983**
	Recall	0.666	0.642	**0.673**	0.583	0.643	0.599	0.647	0.577	0.642	0.599	0.616	0.632	**0.812**
	*F*1-score	0.627	0.594	**0.637**	0.501	0.593	0.534	0.601	0.493	0.594	0.533	0.554	0.585	**0.807**
BMJ	*AUC*	0.617	**0.620**	0.615	0.567	0.607	0.577	0.578	0.539	0.571	0.570	0.575	0.571	**0.751**
	Precision	**0.937**	0.921	0.937	0.825	0.889	0.889	0.905	0.698	0.937	0.889	0.825	0.873	**0.940**
	Recall	0.621	0.615	**0.628**	0.569	0.599	0.579	0.588	0.543	0.563	0.579	0.582	0.580	**0.744**
	*F*1-score	0.562	0.554	**0.570**	0.473	0.531	0.504	0.510	0.432	0.471	0.499	0.507	0.505	**0.735**

同时，本文还研究了不同训练集和测试集比例下的链路预测效果，并将本文提出的多模体预测效果与可定义势模体特征 S_5 和 S_7 以及已有依据模体的链路预测方法 QMI[17] 和图表示学习的经典方法 node2vec[31-32] 进行比较，结果如图 6 和表 4 所示. 图 6 以 BMJ 网络数据的实验结果为例进行说明，其他 5 种网络数据集的结果都是类似的. 图 6 实验结果表明，融合所有模体特征的多模体特征链路预测准确性最高，其预测精度与另外四种方法相比提升了约 5%～19%，说明该方法的预测效果最好. 表 4 中选择训练集和测试集的比例为 8∶2，利用四种评价指标对 5 种预测方法的实验结果进行比较，发现融合所有模体特征的预测结果仍然是最好的. 由于多模体特征不仅考虑了网络连边的方向性，还融合了网络的多个微观结构（模体）特征，相比其他使用单一特征的方法更充分地利用了网络结构信息，因此预测性能更高.

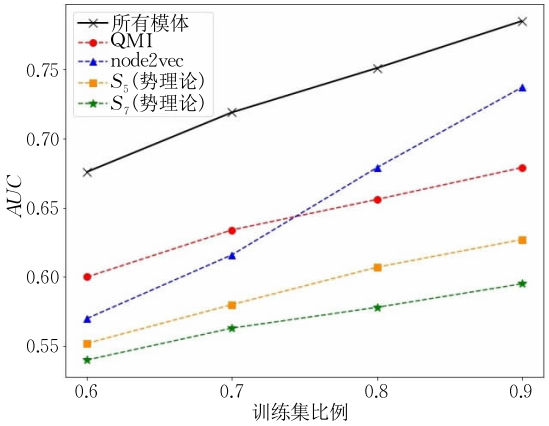


图 6　多模体特征与现有方法的对比结果

表 4　5 种链路预测方法的预测结果

网络及评价指标		QMI	node2vec	S_5	S_7	所有模体
Science	*AUC*	0.749	0.756	0.711	0.645	**0.816**
	Precision	0.981	0.986	0.986	0.981	**0.990**
	Recall	0.653	0.719	0.664	0.650	**0.792**
	*F*1-score	0.608	0.718	0.621	0.605	**0.786**
Nature	*AUC*	0.745	0.682	0.720	0.645	**0.824**
	Precision	0.944	0.978	0.983	0.961	**0.987**
	Recall	0.647	0.648	0.647	0.643	**0.790**
	*F*1-score	0.601	0.647	0.596	0.601	**0.783**
NEJM	*AUC*	0.819	0.781	0.743	0.700	**0.851**
	Precision	0.967	0.968	0.974	0.935	**0.983**
	Recall	0.688	0.714	0.684	0.677	**0.827**
	*F*1-score	0.660	0.712	0.655	0.647	**0.824**
Lancet	*AUC*	0.820	0.758	0.791	0.697	**0.894**
	Precision	0.980	0.985	0.995	0.975	**0.996**
	Recall	0.613	0.702	0.635	0.578	**0.857**
	*F*1-score	0.547	0.701	0.581	0.488	**0.855**
JAMA	*AUC*	0.742	0.749	0.699	0.648	**0.830**
	Precision	0.956	0.971	0.966	0.957	**0.983**
	Recall	0.651	0.685	0.643	0.647	**0.812**
	*F*1-score	0.608	0.683	0.593	0.601	**0.807**
BMJ	*AUC*	0.642	0.674	0.607	0.578	**0.751**
	Precision	0.857	0.936	0.889	0.905	**0.940**
	Recall	0.602	0.623	0.599	0.588	**0.744**
	*F*1-score	0.536	0.619	0.531	0.510	**0.735**

《计算机学报》

4.4 不同模体之间的相关性分析

传统上，一般使用 *Pearson* 相关系数求两个变量之间的相关性[33]. *Pearson* 相关系数只能求线性相关性，不能度量线性关系的斜率和非线性关系，而且容易受噪声的影响. 因此，本文采用最大信息系数 *MIC*(Maximal Information Coefficient)[34] 进行多模体特征的相关性分析. 该方法优于 *Pearson* 相关系数，可以判定变量间的函数关系或者非函数关系，进而得出该变量在数据集中的影响力. *MIC* 计算分为三个步骤：给定 i 和 j，对变量 X、Y 构成的散点图进行 i 列 j 行网格化，并求出最大的互信息值；然后对最大的互信息值进行归一化；最后选择不同尺度下互信息的最大值作为 *MIC* 值. 其公式定义如下：

$$MIC[X;Y]=\max_{|X||Y|<B}\frac{I[X;Y]}{\log_2(\min(|X|,|Y|))} \tag{29}$$

其中，$I[X;Y]$ 表示变量 X 和 Y 之间的互信息，$|X|$、$|Y|$ 表示在散点图网格中，分别在 X 和 Y 方向共被分成了多少段，$|X||Y|<B$ 表示所有的方格总数不能大于 B，B 取数据总量的 0.6 或 0.55 次方，该值是一个经验值. 实验中对任意两个模体 f_i 和 f_j 之间的冗余性（也是一种相关性）定义为 $MIC=(f_i,f_j)$. $MIC=(f_i,f_j)$ 值越大，说明模体 f_i 和 f_j 间的可替代性越强，即冗余性越强. $MIC=(f_i,f_j)$ 的值为 0，说明 f_i 和 f_j 之间相互独立.

本文以 NEJM 网络数据为例，对 12 种预测器进行相关性分析，结果如图 7 所示. 从模体之间的相关性可以看出，相关性较大的模体之间的结构是相似的，因此根据相关性大小将所有预测器分成了四类，即图中的四个实线方框. 每一类都包含三种模体，其中有两个四节点模体和一个三节点模体，如果将两个四节点模体中计算角色函数部分（节点 c、d 及其连边构成的整体）看成是一个节点，那么这两个模体的结构与三节点模体结构一致，说明图 3 中基于三节点模体来构建四节点模体预测器并将其简化是合理的、可行的. 此外，通过分析模体之间的相关性，能更好地理解科学家合作网络结构形成的机理，也能为多模体链路预测的模体（特征）选择提供选择依据，在降低算法复杂度的同时不会大幅降低算法性能.

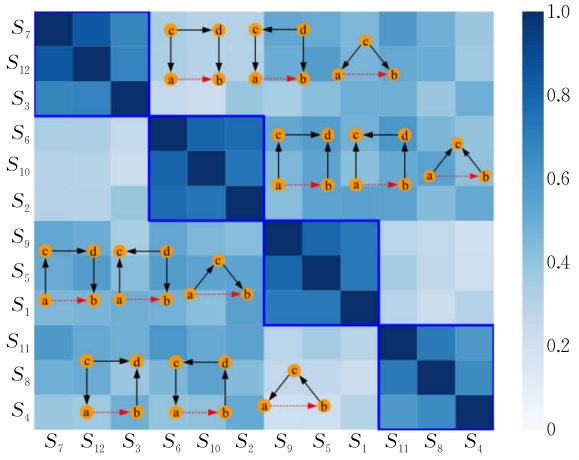


图 7 12 种模体预测器之间的相关性分析

5 结 论

本文针对有向科学家合作网络分别进行了单模体边度、双模体边度和多模体边度的科研合作关系预测. 首先利用朴素贝叶斯模型推导出模体边度模型进行链路预测，解决了传统方法中四节点模体无法使用贝叶斯模型的难题，也为模体边度模型可进行链路预测的原因提供理论解释. 理论推导发现此类方法不仅对于单模体预测器有效，对于双模体预测器也有效，并且与单模体边度链路预测结果相比，双模体的预测性能更好. 然后，研究了基于机器学习框架的多模体边度链路预测，通过与 QMI、node2vec 和满足势理论的模体预测结果进行比较发现，融合所有模体特征的预测结果更好，预测性能提升了约 5%～19%，证明了本文所提方法的有效性. 最后，应用最大信息系数方法分析了 12 种模体边度预测器之间的相关性，每一种模体形式都与科学家之间的合作模式相对应，发现结构越相似的模体之间的预测性能的相关性越强. 本研究拓展了模体理论的应用场景，提升了科研合作关系预测的准确性，有助于进一步理解有向网络的演化机制，也为有向网络上的其他应用提供了一些新的思路.

参 考 文 献

[1] Zeng A, Shen Z S, Zhou J L, et al. The science of science: From the perspective of complex systems. Physics Reports, 2017, 714(16): 1-73

[2] Fortunato S, Bergstrom C T, Börner K, et al. Science of science. Science, 2018, 359(6379): eaao0185

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

[4] Liu Liang, Luo Tian, Cao Ji-Ming. A study of the multi-scale scientific collaboration patterns based on complex networks. Science Research Management, 2019, 40(1): 191-198(in Chinese)
(刘亮，罗天，曹吉鸣. 基于复杂网络多尺度的科研合作模式研究方法. 科研管理，2019，40(1)：191-198)

[5] Li J J, Zhang J, Li H J, et al. Network and community structure in a scientific team with high creative performance. Physica A: Statistical Mechanics and Its Applications, 2018, 508(15): 702-709

[6] Zhou J L, Zeng A, Fan Y, et al. Identifying important scholars via directed scientific collaboration networks. Scientometrics, 2018, 114(3): 1327-1343

[7] Lü L, Zhou T. Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150-1170

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

[9] Zhang Bin, Ma Fei-Cheng. A review on link prediction of scientific knowledge network. Journal of Library Science in China, 2015, 41(3): 99-113(in Chinese)
(张斌，马费成. 科学知识网络中的链路预测研究述评. 中国图书馆学报，2015，41(3)：99-113)

[10] 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

[11] Zhou T, Lü L, Zhang Y C. Predicting missing links via local information. The European Physical Journal B-Condensed Matter and Complex Systems, 2009, 71(4): 623-630

[12] Cannistraci C V, Alanis-Lobato G, Ravasi T. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. Scientific Reports, 2013, 3(1613): 1-13

[13] Grover A, Leskovec J. node2vec: Scalable feature learning for networks//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 855-864

[14] Kovács I A, Luck K, Spirohn K, et al. Network-based prediction of protein interactions. Nature Communications, 2019, 10(1): 1-8

[15] Liu Z, Zhang Q M, Lü L, et al. Link prediction in complex networks: A local Naïve Bayes model. Europhysics Letters, 2011, 96(4): 48007

[16] Wu J H, Zhang G J, Ren Y Z, et al. Weighted local Naïve Bayes link prediction. Journal of Information Processing Systems, 2017, 13(4): 914-927

[17] Hu X X, Liu S X, Chang S, et al. A quad motifs index for directed link prediction. IEEE Access, 2019, PP(99): 1-1

[18] Jiao Z J, Wang H, Ma K, et al. Directed connectivity of brain default networks in resting state using GCA and motif. Frontiers in Bioscience, 2017, 22(10): 1634-1643

[19] Aghabozorgi F, Khayyambashi M R. A new similarity measure for link prediction based on local structures in social networks. Physica A: Statistical Mechanics and Its Applications, 2018, 501(1): 12-23

[20] Milo R, Shen-Orr S, Itzkovitz S, et al. Network motifs: Simple building blocks of complex networks. Science, 2002, 298(5594): 824-827

[21] Milo R, Itzkovitz S, Kashtan N, et al. Superfamilies of designed and evolved networks. Science, 2004, 303(5663): 1538-1542

[22] Han Hua, Liu Wan-Lu, Wu Ling-Yan. The measurement of complex network based on motif. Acta Physica Sinica, 2013, 62(16): 168904(in Chinese)
(韩华，刘婉璐，吴翎燕. 基于模体的复杂网络测度量研究. 物理学报，2013，62(16)：168904)

[23] Zhang Q M, Lü L, Wang W Q, et al. Potential theory for directed networks. PLoS One, 2013, 8(2): e55437

[24] Xu Xiao-Ke, Xu Shuang, Zhu Yu-Xiao, et al. Link predictability in complex networks. Complex Systems and Complexity Science, 2014, 11(1): 41-47(in Chinese)
(许小可，许爽，朱郁筱等. 复杂网络中链路的可预测性. 复杂系统与复杂科学，2014，11(1)：41-47)

[25] Lü Lin-Yuan, Zhou Tao. Link Prediction. Beijing: Higher Education Press, 2013(in Chinese)
(吕林媛，周涛. 链路预测. 北京：高等教育出版社，2013)

[26] Zhang Qian-Ming. Structure Analysis and Link Prediction in Complex Networks [Ph. D. dissertation]. University of Electronic Science and Technology of China, Chengdu, 2016 (in Chinese)
(张千明. 复杂网络结构分析与链路预测[博士学位论文]. 电子科技大学，成都，2016)

[27] Chiang K-Y, Natarajan N, Tewari A, et al. Exploiting longer cycles for link prediction in signed networks//Proceedings of the 20th ACM Conference on Information and Knowledge Management. Glasgow, UK, 2011: 1157-1162

[28] Vázquez A, Dobrin R, Sergi D, et al. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(52): 17940-17945

[29] Li Ye-Zi, Wang Zhen-You, Zhou Yi-Lu, et al. The improvement and application of XGBoost method based on the Bayesian optimization. Journal of Guangdong University of Technology, 2018, 35(1): 23-28(in Chinese)
(李叶紫，王振友，周怡璐等. 基于贝叶斯最优化的 XGBoost 算法的改进及应用. 广东工业大学学报，2018，35(1)：23-28)

[30] Li Zhan-Shan, Liu Zhao-Geng, Ding Guo-Xuan, et al. Feature selection algorithm based on XGBoost. Journal on Communications, 2019, 40(7): 1-8(in Chinese)

《计算机学报》

(李占山，刘兆赓，丁国轩等. 基于 XGBoost 的特征选择算法. 通信学报，2019，40(7)：1-8)

[31] Tu Cun-Chao，Yang Cheng，Liu Zhi-Yuan，et al. Network representation learning：An overview. SCIENTIA SINICA Informationis，2017，47(8)：980-996(in Chinese)
(涂存超，杨成，刘知远等. 网络表示学习综述. 中国科学：信息科学，2017，47(8)：980-996)

[32] Zhang Jin-Zhu，Yu Wen-Qian，Liu Jing-Jie，et al. Predicting research collaborations based on network embedding. Journal of the China Society for Science and Technical Information，2018，37(2)：132-139(in Chinese)
(张金柱，于文倩，刘菁婕等. 基于网络表示学习的科研合作预测研究. 情报学报，2018，37(2)：132-139)

[33] Mudelsee M. Estimating Pearson's correlation coefficient with bootstrap confidence interval from serially dependent time series. Mathematical Geology，2003，35(6)：651-665

[34] Hsu W H. Genetic wrappers for feature selection in decision tree induction and variable ordering in Bayesian network structure learning. Information Sciences，2004，163(1)：103-122

LIU Juan，M. S. candidate. Her research interests include social network analysis and link prediction.

LIU Ya-Fang，M. S. Her research interests include social network analysis and data visualization.

XU Shuang，Ph. D.，associate professor. Her current research interests include big data analysis and processing.

XU Xiao-Ke，Ph. D.，professor. His current research interests include community detection，link predicting，and data mining on complex networks.

《计算机学报》

Background

In recent years，the prediction of scientific collaboration relationship has become a hot research topic in the field of science，because it is significant for understanding the cooperation mechanism among scientists and the evolution mechanism of scientific research networks. Link prediction based on network structure information can be divided into two kinds of methods：global information and local structure similarity. Although the two kinds of methods have achieved efficient prediction results in the undirected network，they can only use the undirected edge information and consider that the individual status is equal. In the prediction of scientific collaboration relationship，it is impossible to distinguish scientists as the first author，corresponding author or ordinary author by the undirected network，which will lose key information such as scientific influence and future potential，and cause the deviation to scientific ranking and analysis of their influence in scientific fields. In addition，studying unequal cooperation in undirected networks will have a certain impact on the understanding of scientific research cooperation. Therefore，a prediction method based on potential theory is proposed for link prediction of directed networks. But this method does not extend the potential theory to the motif theory，and does not comprehensively consider comprehensive information of multiple motifs.

In this study we construct two predictors using the edge degree of single and dual motifs. Then，the link prediction method of edge degrees of multiple motifs using a machine learning framework is proposed. Compared with the prediction results of QMI，node2vec and the motifs satisfying potential theory，the new predictors have higher performance for link prediction，the prediction accuracy is increased 5%～19%. Our findings expand the application scenarios of motif theory，which can promote our understanding of the evolution mechanism of scientific collaboration networks.

The work was supported by the National Natural Science Foundation of China (61773091，61603073)，the Key Research and Development Plan of Liaoning province (2018104016)，the Liaoning Revitalization Talents Program (XLYC1807106)，and the Program for the Outstanding Innovative Talents of Higher Learning Institutions of Liaoning (LR2016070).