情报学报　2026 年 3 月　第 45 卷　第 3 期
Journal of the China Society for Scientific and Technical Information, Mar. 2026, 45(3): 332-349

DOI: 10.3772/j.issn.1000-0135.2026.03.002

基于XGBoost-SHAP框架的科研合作频次归因分析

彭兆祺[1,2]，石　斌[1,2]，杨　杰[1,2]，邓三鸿[1,2]

（1. 南京大学信息管理学院，南京　210023；
2. 数据工程与知识服务省高校重点实验室（南京大学），南京　210023）

摘　要　科研合作频次的背后蕴含着复杂的合作特征，在揭示科研人员合作关系的强度与模式方面具有关键作用。本研究选取生物医学领域数据集PKG 2.0（PubMed knowledge graph 2.0）中的13220951条真实合作关系作为样本，旨在揭示不同合作频次背后的关键驱动因素和内在规律。首先，以合作频次作为表征合作关系强度的代理变量，并根据其数据分布特征将频次划分为低频、中频和高频合作；其次，选择研究相似度、成果生产模式、学术资本和个体属性等维度的10类差异性变量，作为描述作者间合作关系的特征体系，并运用XGBoost（extreme gradient boosting）算法捕捉高维特征间的复杂关联；最后，通过SHAP（shapley additive explanations）解释框架对预测模型进行归因分析，以评估特征对不同合作频次的影响程度和交互机制。本研究发现，作者间研究主题相似度在合作频次分类中发挥主要作用，其次为论文数量差异、总被引用次数差异和均被引次数差异。研究主题高相似是维持高频合作的核心要素，低频合作则更关注总量积累而非个体论文影响力。研究主题相似度、论文数量差异与H指数差异对合作频次的作用模式分别为双峰对称型、阈值稳定型及边际递减型。此外，高频合作对知识结构或影响力构成等异质性因素具有显著的包容性，其背后往往形成了以互补为导向的有效分工机制。

关键词　科研合作；合作频次；XGBoost-SHAP；算法归因

Attribution Analysis of Research Collaboration Frequency Based on XGBoost-SHAP Framework

Peng Zhaoqi[1,2], Shi Bin[1,2], Yang Alex Jie[1,2] and Deng Sanhong[1,2]

(1. *School of Information Management, Nanjing University, Nanjing*　210023; 2. *Key Laboratory of Data Engineering and Knowledge Services in Provincial Universities (Nanjing University), Nanjing*　210023)

Abstract: The frequency of scientific collaboration may exhibit complex collaborative characteristics, which plays a key role in comprehensively understanding the intensity and patterns of researchers' collaborative relationships. This study selected 13,220,951 authentic collaborative relationships from the PubMed knowledge graph 2.0 (PKG 2.0) biomedical dataset as research samples to reveal the key drivers and inherent patterns behind varying collaboration frequencies. First, the frequency of collaboration was treated as a proxy variable to measure the strength of the collaborative relationship, categorizing the frequency into low, medium, and high levels based on its distribution characteristics. Second, ten distinct variables across four dimensions were selected to form a feature system describing author collaboration intensity, which included

收稿日期：2025-04-15；**修回日期**：2025-10-03

基金项目：国家社会科学基金一般项目“基于科学数据关联的科研合作网络研究”（24BTQ041）。

作者简介：彭兆祺，1999年生，博士研究生，主要研究领域为科研合作、科学计量；石斌，1999年生，博士研究生，主要研究领域为多模态数据处理技术；杨杰，2000年生，博士研究生，主要研究领域为科学计量、复杂网络、创新测度；邓三鸿，通信作者，1975年生，博士，教授，博士生导师，主要研究领域为科学计量、领域知识图谱、信息分析与科学评价，E-mail：sanhong@nju.edu.cn。

research similarity, achievement production models, academic capital, and individual attributes. The eXtreme Gradient Boosting (XGBoost) algorithm was employed to capture the complex correlations among high-dimensional features. Finally, the SHapley Additive exPlanations (SHAP) framework was used to analyze the attribution of the prediction model. The influence degree and interaction mechanism of the features on different cooperation intensity were further evaluated. The study reveals that thematic similarity among authors plays a dominant role in the classification of collaboration frequency, followed by differences in the number of papers and citations (total and average). High thematic congruence serves as the core factor sustaining high frequency collaboration, whereas low frequency collaboration prioritizes cumulative output over individual paper impact. The effects of thematic similarity, paper count differences, and H-index variations on collaboration frequency followed distinct patterns: bimodal symmetry, threshold stability, and marginal diminishing effects, respectively. Furthermore, high-frequency collaborations demonstrated remarkable tolerance for heterogeneous factors such as knowledge structure or influence composition, often forming effective complementary division mechanisms through coordinated efforts.

Keywords: scientific collaboration; collaboration frequency; XGBoost-SHAP; algorithmic attribution

0　引　言

科学研究作为人类认知体系演进的核心驱动力，通过系统揭示客观世界的本质规律与内在机理，持续深化人类对自然存在与宇宙秩序的理性建构。在当代知识生产体系高度复杂化的背景下，科研合作已逐步演变为促进多学科交叉融合、有效应对重大科学挑战以及突破前沿领域发展瓶颈的核心范式。研究人员借助彼此的知识和专业技能，通过合作协同推进特定研究项目[1]，不仅有效提升了研究效率，促进创新成果产出[2]，更实现了隐性知识的跨主体流动与重组，进而成为驱动科学发现指数级增长的关键机制[3]。合著作为科研合作中最直观的表现形式，是深入探索科研合作特征的关键切入点。通过对科研合著关系与行为的研究，不仅可以明确作者的科研合作偏好与倾向，还能在一定程度上揭示科研合作的发展脉络及其背后的影响因素。信息资源管理领域尤为关注通过科学文献的共同署名来审视和研究科研人员之间的合作情况。已有研究表明，中国在成为全球科学强国的整个轨迹中，对国内外合作持开放态度[4]。近 30 年来，跨机构、跨学科合著论文比例持续攀升[5]，合作网络呈现小世界性与无标度性并存的复杂特征[6]，这揭示了科研合作形成机制的紧迫性。作为衡量科研合作强度与稳定性的核心指标，合作频次将抽象的合作关系转化为可量化分析的关键变量。它不仅可以深刻揭示合作网络的演化规律与知识流动路径，还在理解与推动科学研究的进程中扮演着关键角色，并为科研工作者识别潜在伙伴、管理者优化资源配置以及研究者探索领域前沿提供可靠的分析依据与决策支持。

目前，学术界对科研合作的研究主要聚焦于合作成果的后验性分析与合作关系的预测建模，形成了“结果导向”的研究范式。这为理解科研合作的宏观趋势和成效提供了重要视角，但相关研究在深入探讨合作频次本身的影响因素及其背后的驱动机制方面尚显不足。具体而言，学者间合作频次可能受到多层次特征因素的非线性交互影响，这些因素共同构成科研合作环境的复杂网络，其相互作用机制对于揭示科研合作动态及优化合作模式具有深远意义。然而，传统计量模型难以捕捉复杂非线性与交互效应，复杂网络分析则侧重结构拓扑而非个体特征归因。鉴于此，本研究聚焦于科研合作频次这一核心维度，构建基于 XGBoost-SHAP 框架的科研合作频次分类预测模型。选择该框架主要鉴于其与科研合作内在特征的高度适配性：第一，科研合作中的特征作用往往是非线性的，如学术资历差异对合作频次可能存在“倒 U 形”影响，而 XGBoost（extreme gradient boosting）基于树结构的分裂机制天然适合捕捉此类规律；第二，合作关系的形成可能存在因素间的组合效应，如“高主题相似度”可缓解“高地位差异”带来的合作阻力，模型能够自动学习特征间的高阶交互，避免人工构建的主观性；第三，针对 XGBoost 等集成模型“黑箱”难解释的问题，SHAP（shapley additive explanations）框架基于博弈论思想，能够将模型预测值无偏地分解为各特征的边际贡献，从而精准量化各因素对合作频次分类的重要性及作用方向。这种“高精度预测+量化归因”的路径，不仅能精准识别影响合作频次的关键要素，更能系统解构不同合作关系强度背后的特征匹配逻辑与协同作用机制，有望为深化合作模式认知和优化科研合作网络提供参考。

1 相关研究

科研合作作为推动知识生产与创新的重要机制，历来备受学界关注。现有研究主要围绕合作的驱动机制、形成模式及绩效产出等核心议题展开，形成多学科交叉、多方法融合的研究图景。不同学科背景及研究方向的学者，往往基于特定目标、理论框架与实践需求，选择不同的研究视角与路径。这种差异不仅体现在对合作本质的理解上，也影响着合作模式的选择、实施策略与评估。鉴于本研究聚焦于合作频次的预测及其驱动因素，下文主要从合作模式分类、合作演化特征及影响因素3个维度进行文献梳理。

在合作模式分类上，学者基于不同视角提出多种分类框架。Bozeman等[7]从合作主体关系出发，探讨基于指导关系的合作、与青年教师及女性合作、同团队同大学内部合作等类型。Leahey等则从知识结构角度分析领域互补性、同领域强化性以及开拓新主题的3种合作模式[8]，并认为科研合作主要涵盖同领域合作、知识互补与拓展合作、师生合作、跨国合作、跨校合作及跨学科合作等[9]。Liu等[10]基于共生理论的视角，将115对院士的科研合作模式分为引领增长、持续引领、相互追逐、平等站立4个类别。刘飞等[11]通过合作强度、合作强度相似度、合作不平等性等指标，探究国家之间科研合作的倾向性、不平等性及空间集聚模式。刘晓婷等[12]从内聚-耦合视角出发，识别科研团队的内部合作与外部合作模式，并对比分析不同合作模式的差异性。此外，单一国家的不同省、市、地区之间的合作模式也是研究热点之一[13-14]。

在合作演化特征方面，相关研究主要关注科研合作随时间变化的动态规律。欧桂燕等[15]基于化学领域杰青的SCI（Science Citation Index）合著论文，分析其学术职业生涯阶段3个关键时间节点合作角色的演变特征。刘晓娟等[16]分析学者在获评荣誉称号前后的科研合作演变情况，包括合作网络大小、合作强度、合作持续性、在合作中承担角色以及国际/国内合作情况的变化。曹仁猛等[17]则聚焦特殊时期科研合作的变化，分析了新冠疫情暴发前后国际合作的增长情况、生存概率、规模、强度和地位等维度的变化。

在影响因素方面，大量学者探讨了驱动科研合作形成与发展的关键因素。任妍等[18]基于H指数与网络拓扑指标相结合的科研合作角色计量指标，探讨学者的科研合作角色演化的影响因素。Petersen[19]、Bu等[20-21]分别从合作强度、合作持续性、合作稳定性等方面对合作伙伴关系进行分析。Wu等[22]发现生产、经济和信息因素在促进持续合作方面发挥着重要作用，研究课题的邻近度与持续合作之间存在反比的U形关系，地理距离、性别和年龄则对持久合作产生不利影响。也有研究发现，科研人员的知识背景多样性与其社会影响力[23]、论文总被引次数与作者数量[24]之间均存在显著的倒U形关系。还有学者强调其性别、年龄、职称和学科的差异，如基于性别差异探讨学者在合作网络中的角色与偏好[25-26]，不同年龄层学者合作行为的特征与趋势[27]，高级职称学者在合作中的引领作用[28]，关键联络人有助于学者发现国际上的重要合作伙伴[29]，以及学科特性是决定合作的重要影响因素[30]。

综上所述，科研合作相关研究在科学计量与计算科学领域取得显著进展，已形成较为完备的理论与方法体系。学者通过复杂网络分析、计量统计模型和机器学习等多元化方法，从合作模式分类、演化特征和影响因素等多个维度，揭示科研合作的驱动因素与特征。然而，科研合作网络结构特征研究主要依赖于网络密度和聚类系数、平均路径长度等静态指标[31-32]，这些指标虽能刻画宏观拓扑结构，但难以直接量化个体作者的属性特征对具体合作频次的微观影响。此外，合作频次与网络中心度存在强烈的双向因果关系，纳入模型会导致严重的内生性困扰，从而影响结论的有效性。既有影响因素分析则主要聚焦学科、性别、机构、职称等单一变量，未能系统整合关键因素并检验其交互机制，研究主题相似度、论文数量、引用次数等特征及其交互效应与合作频次的关系尚不清晰。近年来，随着可解释机器学习在提升复杂模型决策透明度方面取得显著进展，广泛应用于信息资源管理领域的虚假信息检测[33-34]、政务信息评估[35]、文献引用模式挖掘[36]等。在科研合作研究领域，可解释机器学习为破解合作频次影响因素的“黑箱”、量化特征贡献及其交互机制提供了强有力的方法论工具。因此，本研究基于XGBoost-SHAP框架构建影响合作频次的多维特征体系，利用XGBoost在处理高维数据与捕捉非线性关系上的算法优势，以及SHAP方法的特征贡献归因能力，将复杂的模型预测结果分解为各特征变量的边际贡献，从而实现对合作频次分类预测结果的归因和量化解释。

2　研究方法

2.1　XGBoost-SHAP 框架

XGBoost[37] 是一种基于梯度提升（gradient boosting）思想的集成学习算法，通过迭代地添加决策树来优化预测性能[38]，并引入正则化项控制模型复杂度以防止过拟合[39]。XGBoost 凭借其线性计算复杂度与分布式处理能力，在 Kaggle 等数据科学竞赛中展现了卓越性能[40]。为验证 XGBoost 模型性能，本研究同时采用逻辑回归（logistic regression，LR）、随机森林（random forest，RF）、支持向量机（support vector machine，SVM）、决策树（decision tree，DT）、梯度提升决策树（gradient boosting decision tree，GBDT）、光梯度提升（light gradient boosting machine，LightGBM）及深度神经网络（deep neural network，DNN）作为对比模型。这些模型分别代表线性模型、树集成方法及深度学习方法，涵盖不同的建模思路和能力特点。例如，LR 的系数可解释性直接反映特征对对数几率的线性影响，但无法自动捕获非线性关系，需依赖特征工程引入交互项或多项式项；RF、GBDT、LightGBM 等集成学习方法则能够处理不平衡数据集，具有强大的分类能力和鲁棒性；SVM 适用于高维数据，可以通过核技巧处理非线性关系；DT 通过 CART（classification and regression tree）算法生成可解释规则，其树结构可直接展示特征决策路径；DNN 则通过自动特征工程提取数据中的非线性模式，其深层的网络结构能够学习并表征数据中的复杂特征。此外，XGBoost 模型可解释性仍受限于树模型的全局特征交互遮蔽效应，即高阶交互关系的不可追溯，而 SHAP 值则可以作为相应的解决方案。

对于集成学习和深度学习等复杂度高、决策过程不透明的黑盒模型，Lundberg 等[41]提出了 SHAP 框架，作为模型中特征贡献的解释性指标。SHAP 框架源于合作博弈论中的 Shapley[42]概念，通过将每个特征都视为贡献者，计算其加入模型时的边际贡献，并将所有特征的贡献值求和得出模型的最终预测[43]。特征 i 的 Shapley 值计算方式为

$$\phi_i = \sum_{S \subseteq D \setminus \{i\}} \frac{|S|!(|D| - |S| - 1)!}{|D|!} [f(S \cup \{i\}) - f(S)] \tag{1}$$

其中，D 为全体特征集合；S 为不含 i 的特征子集；$f(S)$ 为仅使用特征集 S 时模型输出的期望值，$f(S \cup \{i\})$ 是加上 i 之后的期望值，$[f(S \cup \{i\}) - f(S)]$ 则量化 i 对 S 的价值增益，即 i 的边际贡献；$\frac{|S|!(|D| - |S| - 1)!}{|D|!}$ 表示排列组合权重系数，保证所有可能的加入顺序具有均等概率。树模型的预测结果依赖于所有分裂特征，对于未包含在 S 中的特征，如何“遮蔽”其影响需要取期望，且直接枚举所有子集 S 的复杂度是指数级的。因此，SHAP 的贡献在于通过在树的分支结构中递归计算期望输出，实现对 Shapley 值的精确高效求解，其核心解释公式为

$$f(x) = E[f(x)] + \sum_i \phi_i \tag{2}$$

其中，$E[f(x)]$ 为模型在训练集上的期望输出（全局基线）；ϕ_i 为特征 i 对该样本的个体化贡献；$\sum_i \phi_i$ 为样本相对于基线的预测偏移。这保证了解释的可加性、局部准确性与单调一致性。首先，模型解释的局部准确性在于，所有特征的 SHAP 值之和等于模型预测值与基线值（通常为全部样本预测均值）之差；其次，一致性表现为模型在某一特征上的依赖度增加（其他特征不变），则该特征的 SHAP 值不会相应减少，确保特征重要性的排序具有可靠性与可比性。在全局解释层面，通过聚合所有样本的 SHAP 值获得特征的整体重要性，以及显式捕捉特征之间的非线性与交互效应。相较于 LIME（local interpretable model-agnostic explanations，局部线性近似可能不稳定）、Permutation Importance（特征相关性高时估计有偏）等事后解释方法，SHAP 能更高效且可靠地量化特征在存在复杂交互场景下的贡献[44]。综上，SHAP 框架以其独特的解释能力，能够清晰地揭示各特征变量对科研合作频次的独立影响与交互机制，为深入理解学者合作行为的内在机理提供有力支持。

2.2　研究框架

科研合作可以视为研究人员出于共同研究目标和互补专业知识等目的，评估潜在合作对象的相关信息或特征，进而做出合作匹配的一种决策过程。在此过程中，各特征变量实质上构成了合作决策的多维代理指标。合作频次作为合作关系强度的显性表征，其形成受到多层次因素的交互驱动，作用强度与路径呈现显著异质性。例如，作者间的发文数量、学术年龄、研究主题、署名位置等差异性因素通过改变科研生产的边际收益以及降低合作成本，直接影响其合作是止步于“一次性尝试”还是演化为“高频合作关系”。若将上述因素对应到合作频

次的分类预测任务中，则表现为各特征因素在分类边界上的影响程度及作用方式差异。

因此，本研究通过 XGBoost-SHAP 框架量化各类特征对合作频次的影响程度，进而探索不同频次背后的关键驱动因素和作用方式，以期为科研合作频次的实证研究与合作策略优化提供参考依据。具体研究框架如图 1 所示。算法归因研究通常以多维度特征驱动机制解析为核心目标，主要包含以下 3 个阶段：第一，特征-结果关联建模，为深入理解学者合作行为的多样性和模式，以学者间的合作频次类别为核心预测目标，并基于 PKG 2.0（PubMed knowledge graph 2.0[①]）数据集，从研究相似度、成果生产模式、学术资本和个体属性等维度构建特征体系；第二，模型构建与评价，选取 XGBoost 算法作为科研合作频次分类预测的基准模型，其余模型用作对比验证，通过准确率、ROC-AUC（receiver operating characteristic - area under curve）、PR-AUC（precision, recall - area under curve）以及对数损失等指标评估模型性能；第三，归因解释体系构建，基于 Shapley 理论计算全部特征的 SHAP 值，通过重要性排序和个体效应分解进行全局解释与局部解释，以及借助相关性分析选取两两特征展开交互效应分析，进而识别影响合作频次的关键特征，并揭示特定合作模式背后的内在规律。

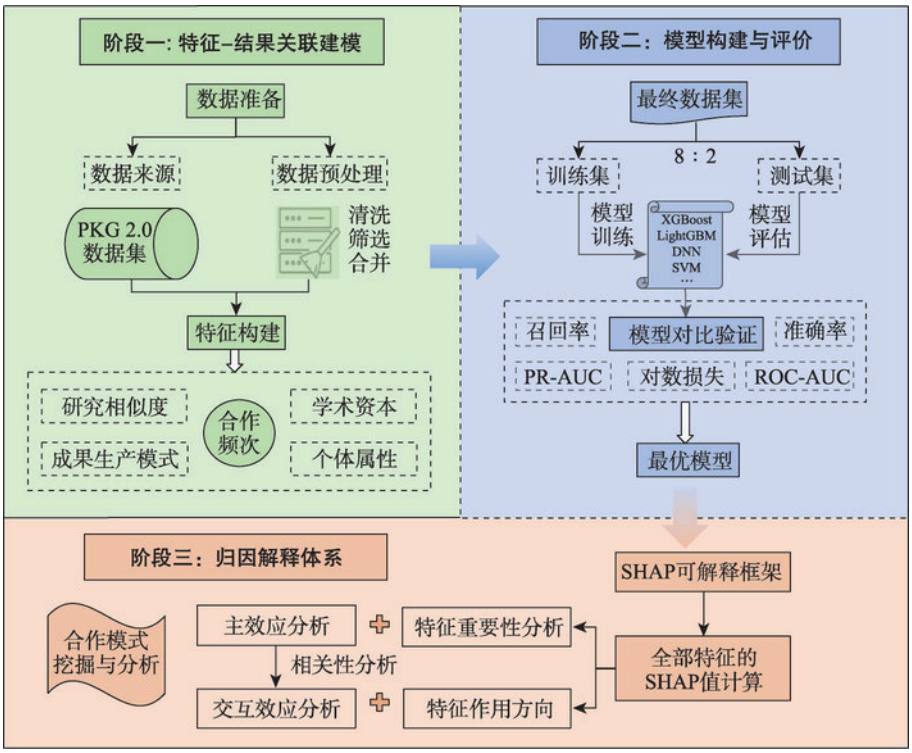


图 1 研究框架

3 数据准备与模型训练

3.1 数据来源与预处理

本研究采用的数据源自 PKG 2.0。作为一个大规模、结构化的生物医学科研知识图谱，PKG 2.0 系统整合了 PubMed[②]数据库中超过 3600 万篇文献及其关联的元数据，涵盖题名、作者、机构、关键词及引文等内容[45]。相较于原始 PubMed 数据，PKG 2.0 融合了 Author-ity 与 Semantic Scholar 等作者消歧数据，并基于深度神经网络对作者身份进行识别与归一化处理，以应对 PubMed 中作者信息稀疏、难以唯一标识的问题。在作者属性补充方面，PKG 2.0 借助 genderize[③]与 gender_guesser[④]等工具，依据

① https://figshare.com/s/05158da6355d6199d3b7

② PubMed 是由隶属于美国国立卫生研究院的国家医学图书馆于 1996 年创建的权威生物医学文献检索数据库，涵盖医学、生命科学、药学、护理学、公共卫生及其交叉学科的学术文献。

③ https://genderize.io/our-data

④ https://gender-guesser.com

作者姓名特征对其性别与种族进行概率推断，该方法在相关研究中已被广泛采用[46-47]，为本研究分析个体属性对合作行为的影响提供了可行路径。

为构建高质量的科研合作关系样本，本研究经过严格的数据预处理与清洗流程。首先，以 PKG 2.0 中的论文元数据、作者信息、关键词列表等作为核心数据表。通过构建跨表映射字典，以论文唯一标识符（PubMed unique identifier，PMID）为键，关联整合论文发表年份、被引频次以及作者的全名、唯一标识符、H 指数等关键字段。初步抽取涵盖 1786—2024 年的 36554965 篇论文与 26230657 位作者信息。其次，剔除噪声数据并确保样本的有效性，筛选操作如下：①剔除单一作者署名的论文记录，仅保留包含两位及以上作者的合著论文，作为科研合作关系的观测基础；②剔除关键属性（如发表年份、作者 ID、关键词）缺失的记录，确保后续特征计算的可行性；③基于消歧后的作者 ID，剔除由消歧算法误差导致的逻辑异常样本（如同一作者 ID 在同一篇论文中重复出现）。另外，为提升关键词表征质量、降低噪声干扰，本研究基于词频筛选出前 10 名具有代表性的名词型关键词，包括“COVID-19”“inflammation”“SARS-CoV-2”“cancer”“obesity”“epidemiology”等（未考虑“methods”等宽泛含义的单词），并通过正则表达式提取相应关键词记录，将其与论文及作者信息进行关联整合。这一步骤确保后续基于 Jaccard 系数计算的主题相似度能够真实反映科研内容的认知邻近性。基于清洗后的合著论文数据，采用全排列组合方法构建作者间的二元合作关系。即若一篇论文有 N 位作者，则生成 $N(N-1)/2$ 对合作关系。最终构建的数据集包含 13220953 对合作关系，覆盖多维度的科研合作特征变量。

需要注意的是，尽管合作发生时的瞬时状态可用于预测单次合作行为，但本研究聚焦于合作频次类别这一结构性指标，能够在一定程度上反映作者间长期、结构化的匹配关系与资源互补特征。论文数量、总被引次数、H 指数等全局特征能从宏观层面表征决定高频合作模式的基础性势能与持久性因素。相比之下，合作发生时的“瞬时状态”（如合作前 N 年的局部相似度）更适用于预测“某次具体合作是否会发生”，而非“合作关系的累积频次”。进一步地，SHAP 解释的是模型预测的“归因”，而非现实世界的“因果时序”，其直接回答了“哪些全局性的作者差异特征最能区分不同合作频次类别”这一核心研究问题。因此，鉴于研究目标与机器学习任务驱动特征选择、SHAP 解释原理与全局特征的高度适配性，本研究未设置时间窗口筛选条件。

3.2　特征变量构建

本研究旨在构建基于 XGBoost-SHAP 的科研合作频次预测模型，通过解析两两作者间异质性特征变量对合作频次的影响，挖掘潜在的合作模式及其作用机制，可突破现有研究对合作关系的二值化处理局限（合作/不合作），从而实现对科研合作频次的连续谱系解析。已有研究发现，合著次数通过简单统计能够直接反映作者间合作关系的浅层特征[48]。然而，合作频次类别作为本研究的核心因变量，其分类阈值的划分直接影响整个预测模型的精准度与泛化能力，并通过影响特征贡献度的计算基准，最终关系到 SHAP 解释结果的可靠性和可解释性。因此，本研究先对所有论文生成的合作对进行聚合统计，计算每对作者的累积合作频次，再对合作频次数据进行描述性统计，发现其分布呈现显著的长尾特征。具体而言，合作 1 次的作者对占全部样本的 81.23%，合作 2~5 次的占比依次为 8.7%、4.19%、1.93% 和 1.11%，其他频次的占比均小于 1%。这种偏态分布符合科研合作从“试错探索”到“稳定协同”的一般规律。结合数据统计的特征与科研合作模式的相关研究，本研究将合作频次划分为 3 个离散等级：低频合作（1 次），代表偶发或探索性合作；中频合作（2~4 次），代表过渡性合作；高频合作（5 次及以上），代表稳固的合作关系。

为全面捕捉影响合作频次的复杂因素，本研究从个体属性、学术资本、成果生产模式与研究匹配性 4 个维度构建 10 类差异化特征变量（表 1）。在计算逻辑上，除类别特征外，均采用“绝对差异值”或“相似度系数”进行量化，以反映合作双方的非对称依赖或认知邻近性。

首先，个体属性反映学者先天属性或身份差异对合作行为的潜在约束，包括性别与种族两个特征变量。科学研究中的性别不平等现象长期存在，女性学者往往处于弱势地位[49]。在学术合作偏好层面，Benenson 等[50]分析美国多所大学终身教授的科研成果数据，发现无论男性或女性终身教授，均更倾向与男性副教授合作。种族则凸显学术圈层的结构性壁垒，少数族裔学者常面临“文化资本贴现”

表 1 科研合作频次预测模型的特征变量

变量类型	所属维度	特征变量	数据类型	变量描述
因变量	合作关系强度	合作频次	分类变量	根据作者之间合作频次进行分类
自变量	个体属性	性别差异	二值变量	性别异质性
		种族差异		种族异质性
	学术资本	H指数差异	数值变量	双方H指数的绝对差值
		学术年龄差异		双方学术生涯年限的绝对差值
		总被引次数差异		双方论文总被引次数的绝对差值
	成果生产模式	论文数量差异	数值变量	双方发文总量的绝对差值
		均被引次数差异		双方篇均被引次数的绝对差值
		署名顺序差异		双方平均署名位次的绝对差值
		贡献度差异		基于调和计数的累积贡献度差值
	研究匹配性	研究主题相似度	数值变量	作者所有论文的关键词集合的Jaccard相似系数

现象，即使学术水平相当，其研究成果在主流学术体系中的可见度与接受度仍显著偏低。因此，学者的性别与种族差异可能会影响科研合作的频次，计算方式为数据集中对应字符串是否等同作为量化指标，等同则为0，反之则为1。

其次，学术资本体现学者在其学术领域内的累积性影响力，特征变量包括H指数、学术年龄、总被引次数，作者间差异的计算方式为合作双方数值的绝对差值。H指数和总被引次数不仅量化学者的学术产出数量，也反映其学术成就的深度与广度。高H指数与高被引学者往往因其显著的学术成就和影响力，更容易吸引合作伙伴，进而成为科研合作网络中的“枢纽”。由于数据集中包含作者截至2024年的最新H指数，能够反映作者的累积影响力[51]。尽管H指数会随时间增长，但在全局合作频次预测中，采用当前累积值是常见做法，故本研究直接计算合作双方之间H指数的差值。总被引次数则需要先对学者所有论文的被引次数进行求和，再计算合作双方的总被引次数差值。学术年龄代表学者从事研究的年限，通过计算作者第一篇和最后一篇论文发表年份之间的差值并加1，然后计算合作双方的学术年龄差异[52]，反映代际差异对合作模式的塑造作用。例如，资深学者凭借丰富的经验和深厚的学术积累倾向于跨机构和国际合作，青年学者往往依赖导师网络获取必要的资源和支持[53]。这种依赖关系在一定程度上塑造了合作的初期模式，并对学者间合作频次产生直接影响。

再其次，成果生产模式反映学者科研成果的数量、质量分配与贡献分配特征，其特征变量包括论文数量、均被引次数、署名顺序以及贡献度。论文数量是指学者发表论文的总数，可通过直接计算两两作者所有发表论文的数量差。均被引次数为学者单篇文章平均被引用的次数，需要先计算单个作者的论文平均被引次数，再求两者差值。两者通过“马太效应”加剧合作网络的中心-边缘结构[54]，使得高产且高质量成果的学者成为合作网络的中心节点，而相对边缘的学者则面临合作机会的限制。署名顺序是指作者在合著论文中的排列位次，差异值计算同样为绝对差值。贡献度则主要根据论文的作者总数和排名量化单一学者的贡献值，使用调和平均数（harmonic mean）[55]计算单个作者对论文的贡献度，其优势在于平等考虑每位作者的贡献，对极端贡献敏感，能够体现合作程度的不均衡性。计算方法为

$$\text{Credit}_i = \frac{1/r_i}{\sum_{j=1}^{k} 1/r_j} \tag{3}$$

其中，r_i为作者i在包含k位作者的论文中的署名位次。

在科研合作生态中，有影响力的作者往往有固定的合作网络结构，其论文署名顺序也较为固定[56]。因此，论文署名顺序与贡献度差异在一定程度上体现作者在合作中的角色定位（主导者-参与者）[57]。学术话语权和资源分配优势直接作用于单次合作的创新效率与成果层级，并通过学术声誉累积效应，持续拓宽后续合作网络的边界，而贡献度差异则促使学者基于角色预期调整合作策略与行为模式。上述差异不仅影响单次合作的效率和成果质量，还将长远影响其未来的合作机遇与发展路径，形成基于角色认同的合作偏好，进而对科研合作模式的塑造发挥作用。

最后，研究匹配性则聚焦于合作主体间研究主题的相似程度。学者往往倾向于与研究主题或兴趣相似的学者建立合作关系，这种相似性不仅能够降

低合作过程中的协调成本，促进知识与技能的快速融合，还能够在一定程度上减少学科文化壁垒对跨学科合作的阻碍[58]。Tuninetti 等[59]使用共被引关系和共同的关键词来预测两位作者之间的合作。因此，本研究主要采用 Jaccard 相似系数量化作者间的研究主题相似度。具体来说，令作者 A 和 B 发表的所有论文的关键词集合分别为 S_A 和 S_B，其相似度计算公式为

$$J(A,B)=\frac{|S_A \cap S_B|}{|S_A \cup S_B|} \tag{4}$$

3.3　模型构建与评估

本研究以合作频次类别为因变量，以作者之间 H 指数、性别、种族、学术年龄、论文数量、总被引次数、均被引次数、署名顺序、论文贡献度、研究主题相似度等差异值为特征变量，构建基于 XGBoost 的科研合作频次分类预测模型，并与随机森林、决策树、支持向量机、光梯度提升、深度神经网络等主流模型进行对比。为处理类别不平衡问题，本研究按照最小类（合作 5 次及以上）的数据量进行分层抽样，每类随机抽取 50 万条合作关系，并按 8∶2 将其划分为训练集和测试集。模型评估采用多指标综合策略，包括准确率（accuracy）、对数损失（log loss）、ROC（receiver operating characteristic）曲线与 P-R（precision-recall）曲线，并分别计算 ROC-AUC 与 PR-AUC 值。其中，P-R 曲线通过精确率和召回率的关系展示模型在不同阈值下的性能表现；ROC 曲线则以真正例率和假正例率为基础，反映模型对正例和负例的区分能力；准确率衡量模型预测正确的比例；ROC-AUC 和 PR-AUC 分别对 ROC 曲线和 P-R 曲线下的面积进行量化，数值越大，表明模型性能越好；对数损失则从概率角度评估模型预测的准确性，数值越小越好。所有模型的整体预测表现如图 2 所示。

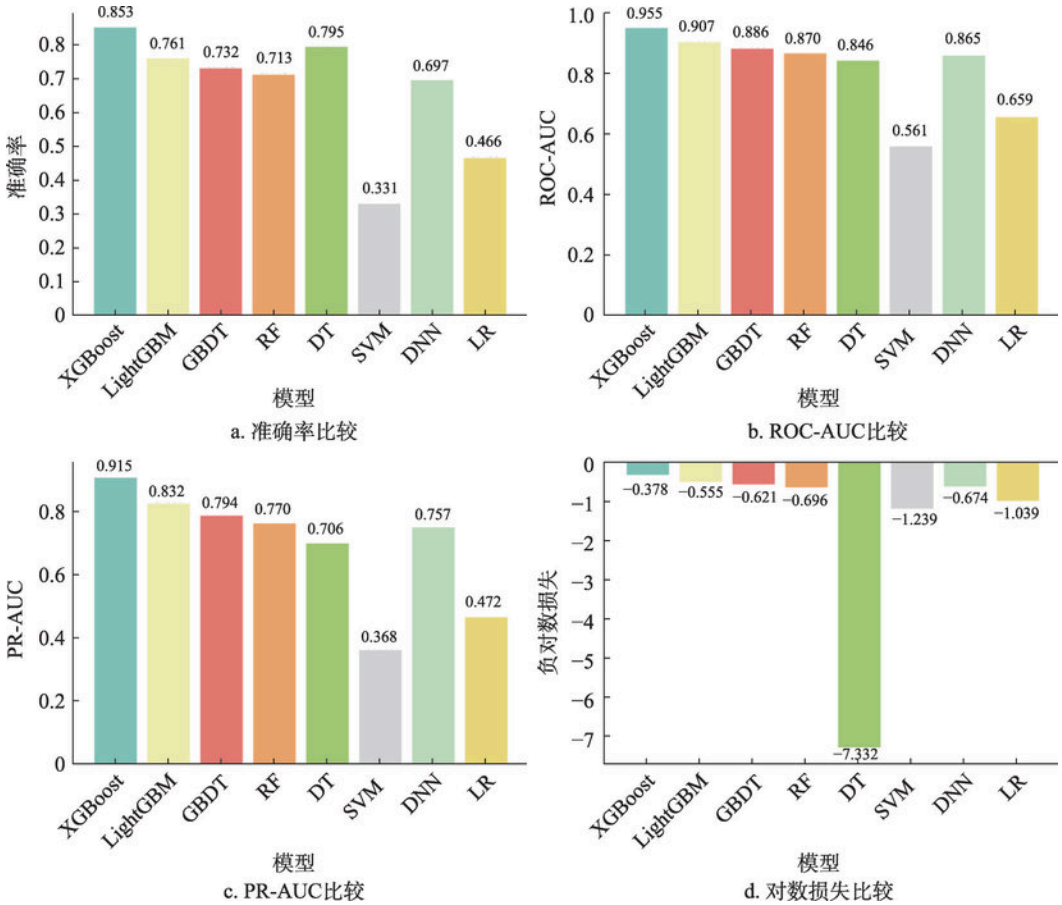


图 2　各模型的整体预测结果比较

显然，XGBoost 在各项指标上均表现最优，其在对数损失的控制上也优于其他对比模型。在 XGBoost 模型训练过程中，采用网格搜索与 5 折交叉验证进行超参数调优，以验证集上的加权 F1-score 作

为主要优化指标。max_depth 与 learning_rate 等关键参数的敏感性分析显示，在选定值附近小范围波动对 F1-score 影响较小，表明模型具有较好的鲁棒性。最终采用的超参数如下：评估指标为多分类对数损失（eval_metric='mlogloss'），树数量（n_estimators）为 500，提升方法为基于树模型（booster='gbtree'），学习率（learning_rate）为 0.8，Gamma 参数为 0.1，L1 和 L2 正则化参数分别为 0.7 和 0.6（reg_alpha=0.7, reg_lambda=0.6），最大树深（max_depth）为 18，最小子节点权重（min_child_weight）为 3，子采样比例和特征采样比例均为 1.0（subsample=1.0, colsample_bytree=1.0）。这些参数经优化调整，可平衡模型复杂度和泛化能力，确保高效建模和稳健性能。

由图 3 所有模型的 P-R 曲线与 ROC 曲线可知，在不同科研合作频次分类下，各模型表现有所差异。在预测科研合作频次为 1 次时，XGBoost 模型在较高召回率水平下仍能保持相对较高的精确率，其平均精度（average precision，AP）达到 0.902。随着预测合作频次的增加（2~4 次以及 5 次及以上），XGBoost 模型依然保持领先，在 2~4 次预测中 AP 为 0.863，5 次及以上预测中 AP 达到 0.979，其表现显著优于其他模型。ROC 曲线的分析结果与 P-R 曲线相呼应，进一步证实 XGBoost 模型具有卓越的分类性能。该模型在各个合作频次分类中的 ROC 曲线、AUC 值均较高，尤其在预测 5 次及以上合作频次时，AUC 值达到 0.990，表明模型能够精准地区分不同合作频次的样本。

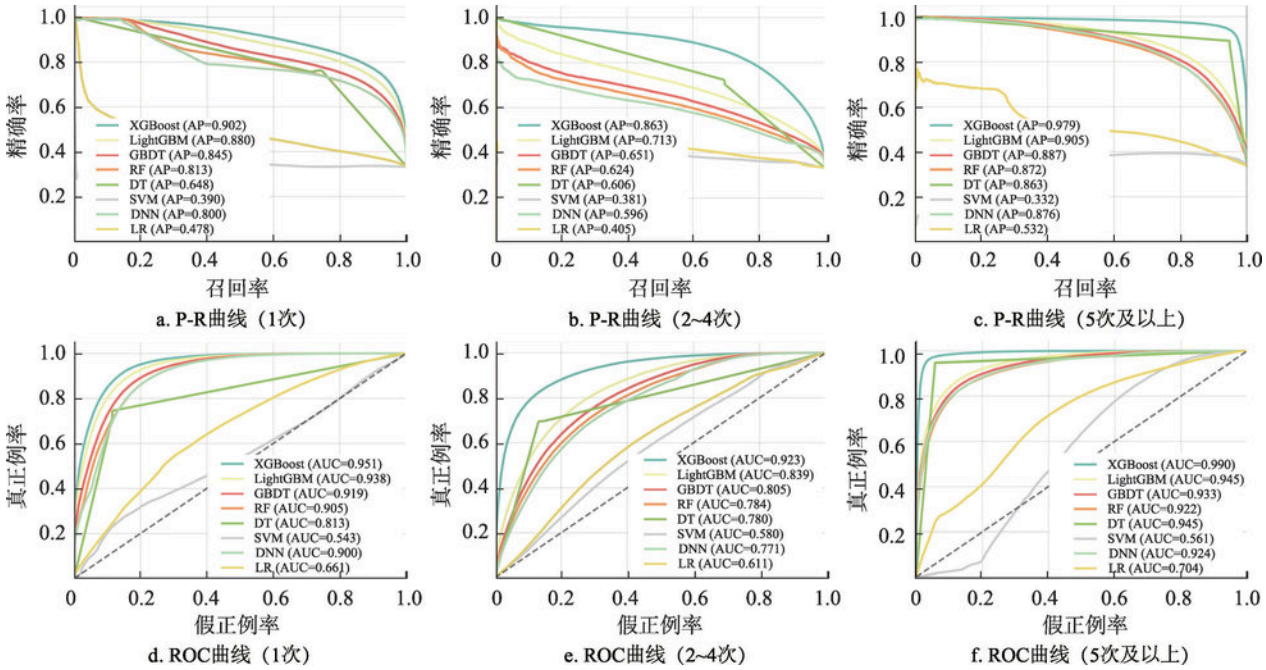


图 3 ROC 曲线与 P-R 曲线（彩图请见 https://qbxb.istic.ac.cn）

采用混淆矩阵对模型在不同合作频次上的分类效果进行可视化评估，结果如图 4 所示。混淆矩阵的对角线元素表示模型预测正确的样本数量。以 XGBoost 模型为例，其低频合作的正确预测数为 84415，其余位置的数据则反映了模型预测错误的样本数量。整体而言，XGBoost、LightGBM、RF、DT 等树模型的预测效果较好，其次为 DNN 模型，而 SVM 与 LR 等线性模型表现欠佳。进一步比较发现，除 LightGBM 在低频合作预测上与 XGBoost 表现相当之外，XGBoost 在中频与高频合作上的分类预测能力均显著优于其他所有模型。XGBoost 模型虽然在区分低频与高频合作样本上表现较好，但在低频与中频合作样本之间存在较多混淆。这一现象可能表明，高频合作具有更强且稳定的判别性特征，即“强模式特征”；而中频及低频合作则表现出更为稀疏、异质且非持续的特征模式，同属于“稀疏突发性特征”，从而导致两者间的区分界限模糊。

4 结果分析与讨论

基于优化后的 XGBoost 模型，下文使用 SHAP

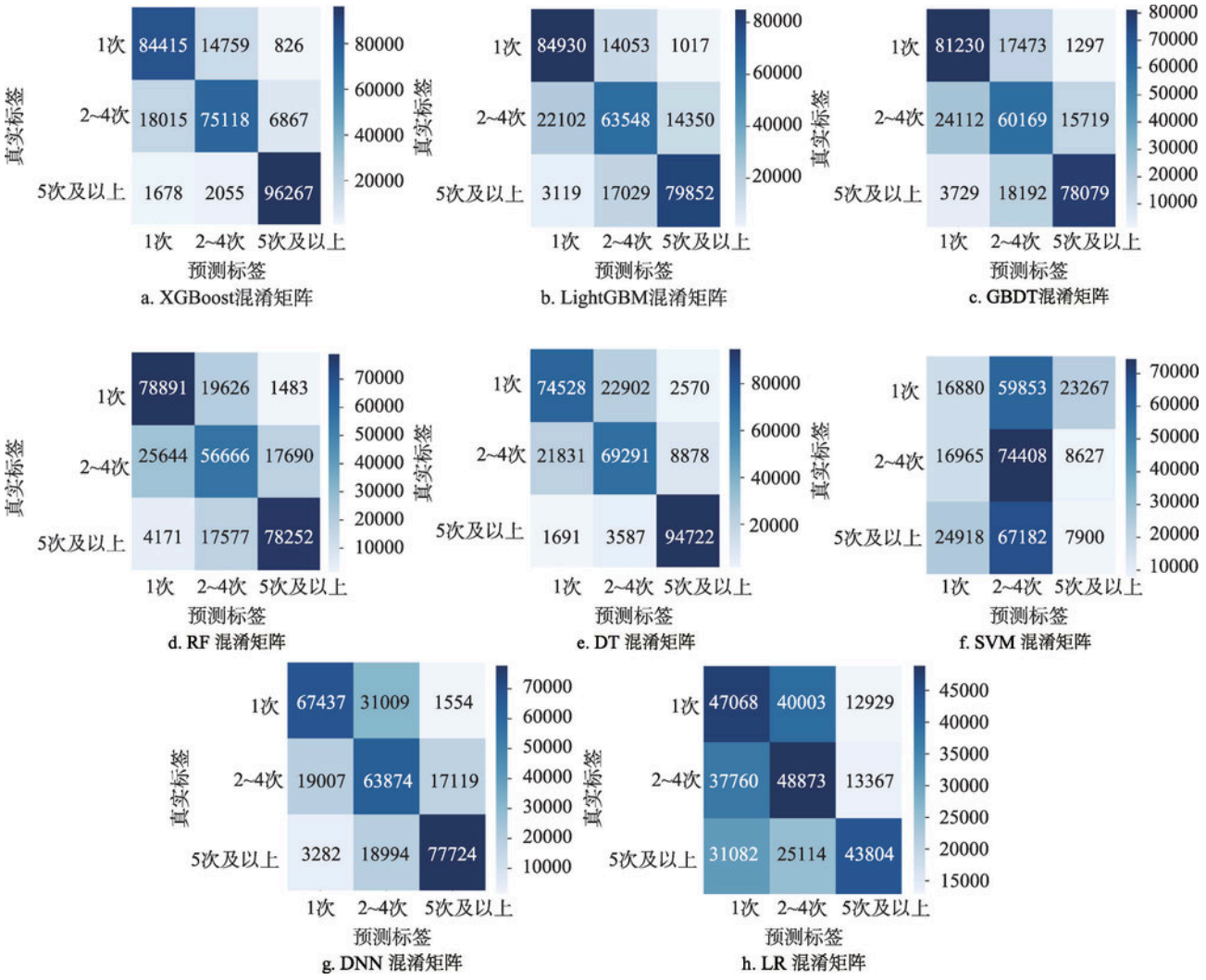


图 4　混淆矩阵

进行可解释分析，包括计算全部特征的 SHAP 值，进行特征的重要程度、作用方式、主效应以及交互效应分析，旨在挖掘出不同科研合作模式的内在逻辑和规律。此外，考虑 XGBoost 模型训练时的特征可比性，所有输入特征变量均已通过 StandardScaler 进行标准化处理（均值为 0，标准差为 1），图 6 和图 8 中横坐标的负值仅表征特征值低于均值的相对偏移量。换言之，标准化后的负值只是数据分布的一种表示方式，有助于算法更好地比较不同特征的重要性，以确保所有特征都在相同的尺度上。

4.1　特征重要性分析

鉴于 SHAP 方法能够同时提供全局与局部的解释能力，每个特征均可通过计算 SHAP 值来量化其影响强度。SHAP 特征概要图通过将特征的重要性与特征效应相结合，直观展示特征值与预测结果之间的正负相关性。在对特征进行全局重要性分析时，先计算每个样本的 SHAP 绝对值，再将所有绝对值求平均，以强调相应特征在整个数据集上对模型预测结果的实际贡献，数值越大，意味着该特征对模型的分类决策过程有较大的影响，计算结果如图 5 所示。其中的条形图（图 5a）和概要图（图 5b）均按照特征的 SHAP 绝对值排序，即不考虑正负影响作用，顶部为最重要的特征。图中每个点代表一个样本实例，横轴为 SHAP 值，纵轴对应具体的特征变量，点的颜色表示单一样本中某个特征变量的具体数值大小，红色表示特征数值较大，蓝色表示值较小。

根据图 5a 和图 5b，重要性较高的特征主要包括作者的研究主题相似度、论文数量差异、总被引用次数差异、均被引次数差异。研究主题相似度为影响作者合作频次的关键因素，其重要性尤为突出。

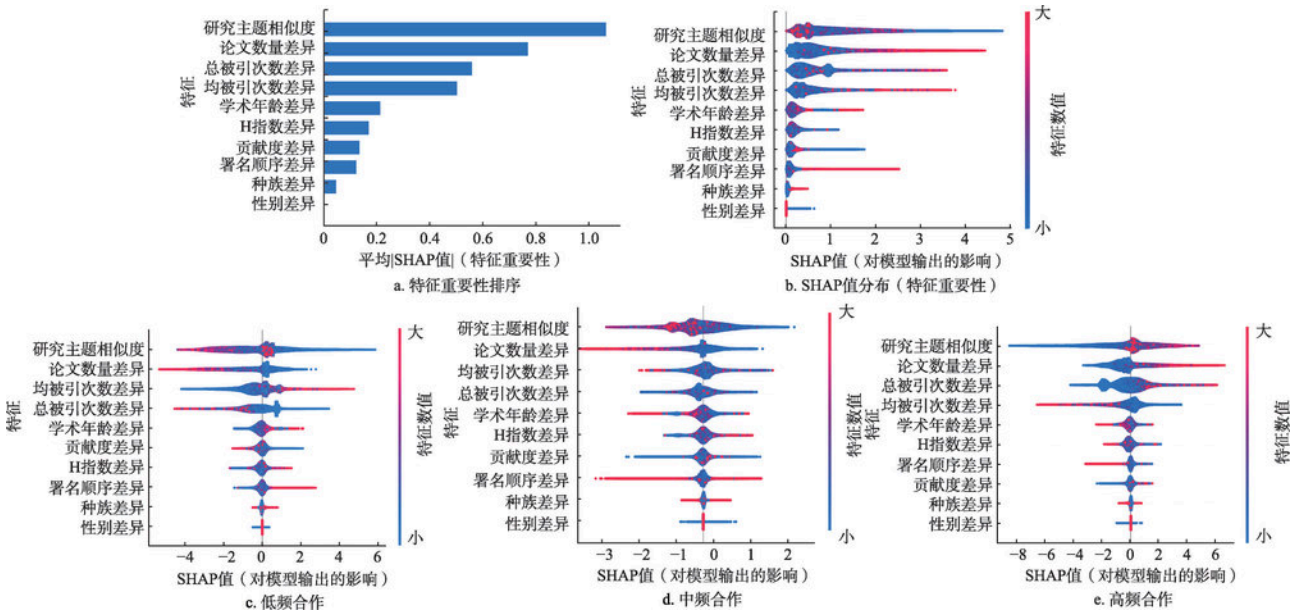


图 5 特征重要性与概要图（彩图请见 https://qbxb.istic.ac.cn）

重要性较低的特征有学术年龄差异、H 指数差异、贡献度差异和署名顺序差异，而种族和性别的差异则对模型影响非常小。这表明作者之间的合作频次主要受到研究匹配性、成果生产模式和学术资本等维度的显著影响，而个体属性维度的性别与种族差异特征则影响较小。

4.2 特征作用方向

由于图 5b 中的 SHAP 值为绝对值，仅体现单一样本中特征对全局的影响。因此，为进一步探究不同合作频次类别中相关特征的重要性和作用方向，下文按照低频、中频和高频 3 类进行特征作用方向分析。如图 5c~图 5e 所示，多数特征的 SHAP 值均呈现单峰的偏态分布，即数据点集中分布于 SHAP 值接近零的中轴线附近，仅作者研究主题相似度、全部被引用次数差异和均被引次数差异呈现双峰分布。单峰分布的峰值区间表征该特征对合作频次预测的主要影响范围，长尾部分则反映其极端值对预测结果的特殊影响。值得注意的是，同一特征在不同频次类别中的 SHAP 值分布区间及方向均存在显著异质性，表明特征对合作频次的影响具有模式依赖性。具体而言，低频与高频合作的特征重要性呈现显著层级分化，重要性较高的特征主要集中在前 4 项，而中频合作的特征重要性分布相对均衡，这种差异化可能是模型在该合作频次的预测效能较弱的重要原因。下文将结合特征重要性排序和具体类别上的差异，系统阐释关键特征的作用方向。

研究主题相似度在 3 种合作频次类别中均居关键特征首位，但其作用方向呈现显著差异。在低频合作中，高相似度（红色）同时分布于 SHAP 负值长尾区（抑制预测）和正值单峰区（促进预测），表征既存在研究领域相近的偶然合作，也存在领域差异较大的探索性合作。中频合作的高相似度集中于 SHAP 负值双峰区，表明主题趋同倾向于抑制中频合作的预测，可能源于此类合作对领域互补性的需求。高频合作的高相似度则稳定分布于 SHAP 正值单峰区，证明研究主题高相似是维持高频合作的核心要素，并与低频和中频合作形成显著区分。

论文数量差异的作用方向与研究主题相似度趋同，但作用强度存在异质性。论文数量差异小在低频和中频合作中呈现正向促进，体现两种合作类别下作者间的学术产出均衡性；在高频合作中则为负向影响：高频合作中频次数据分布比较分散，其中论文数量差异较小的样本占比偏低，从而导致模型在识别高频合作时产生负向作用，这说明高频合作中存在“强弱合作”（教授-学生）和“强强合作”（教授-教授）两种差异化模式。

总被引次数和均被引次数的差异在所有合作频次类中均存在峰值区域，而在低频和高频合作中形成镜像效应。在低频合作中，均被引差异小为抑制预测，而总引用差异小为促进预测；在高频合作中，均被引差异小为促进预测，总引用差异小为抑制预测，并且存在两个较为稳定的作用程度区间（双峰区域）。上述结果表明，初次合作更关注总量

积累而非个体论文影响力，高频合作则注重成果质量的趋同。在中频合作中，两类指标均未显现确定性作用，可能源于其合作动机的多元混杂性，导致模型难以建立稳定关联。其余特征变量对模型的贡献度较低且作用机制相似，在此不作赘述。此外，上述特征因素与合作频次之间并非简单的线性关系，其深层作用规律仍需进一步探究。

4.3　主效应分析

在科研合作频次的分类预测中，特征因素的影响不仅体现在作用强度和方向，更深刻表征于特征与模型预测结果的动态依赖关系中。SHAP 依赖图通过两重理论优势提升模型可解释性：其一是基于博弈论 Shapley 值的公平贡献分配机制，其二是采用条件期望估计规避特征共线性带来的偏差[60]。相较于传统的 PDP（partial dependence plots）部分依赖图，SHAP 依赖图主要分为单变量依赖图和交互依赖图。单变量依赖图通过固定其他特征取值，可视化目标特征与 SHAP 值之间的非线性关联，揭示其对合作频次的独立作用机制。每个特征在不同合作频次类别上的样本分布和作用方式存在差异，导致多数特征的依赖关系散点图呈现非线性或无显著规律分布，故仅选择具有明确变化趋势的典型特征进行主效应解析，结果如图 6 所示。

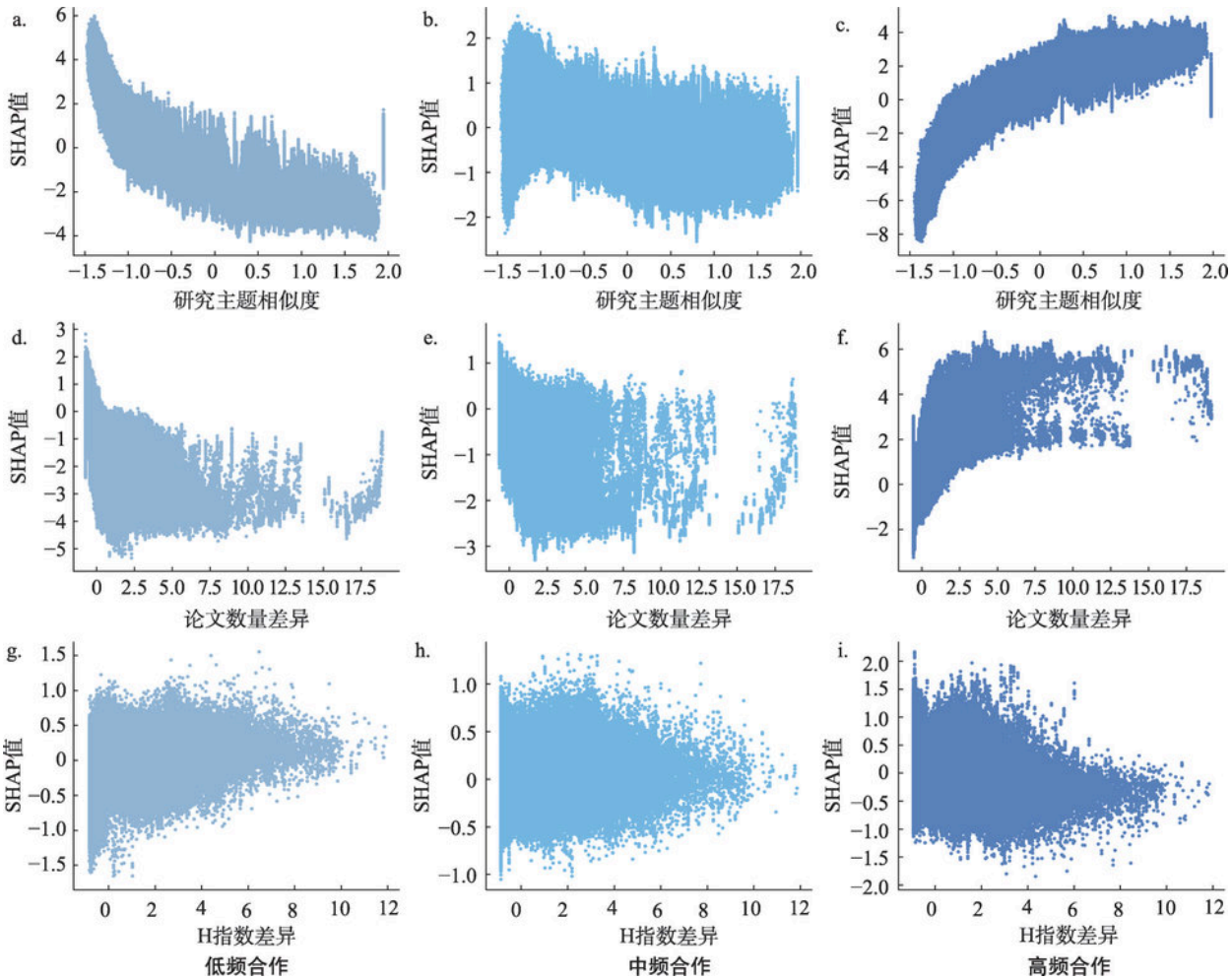


图 6　研究主题相似度、论文数量及 H 指数差异的依赖图

第一，研究主题相似度对合作频次的影响呈现典型的“双峰对称型”特征。在低频合作中，表现为“低相似度高贡献”趋势，说明多数低频合作者之间的主题相似度往往较小，其合作主要基于特定项目或问题的临时性结合。在中频合作中，主题相似度对模型的影响复杂且波动显著，暗示其合作频次兼具偶然性与稳定性，如长期项目与短期协作的交织、多学科方向的交叉探索等。高频合作则遵循“高相似度高贡献”规律，其数据分布与低频合作形成对称，说明高频合作者通常在高度相关的研究

领域内多次合作，形成稳定的研究团队或合作关系。这种合作模式有助于深入挖掘特定方向的问题，推动知识的持续积累和进步。

第二，论文数量差异的影响呈现“阈值稳定型”特征。该因素对低频与中频合作的作用随数值增大而逐步衰减，并在达到某一临界值后趋于平稳，表明较小范围内的产出差异对短期合作具有促进效应，但存在明确的阈值约束。高频合作则呈现“差异强化”特征，如较大论文数量差异可能对应“导师-学生”式的资源互补型合作。该特征能够在一定程度反映合作者之间的研究积累、影响力或资源差异，进而影响合作的稳定性和持续性。

第三，H指数差异则呈现“边际递减型”特征，其影响强度随数值减小而逐步减弱。当合作者间的H指数差异趋近于零时，模型识别效能显著提升，印证了科研合作中存在的“同质匹配”效应——学术影响力相当的研究者之间更易建立高效、稳定的协同关系。

综上，本研究通过主效应分析识别出科研合作中的3类作用机制：双峰对称型揭示研究主题趋同或异质化对合作稳定性的双向影响，阈值稳定型反映资源差异在合作过程中的阶段性作用规律，边际递减型则描述H指数差异匹配的边际效益变化。这3类机制共同形成了一个整合性的理论框架，为深入解析科研合作的形成与演化提供了新思路。

4.4 交互效应分析

交互依赖图通过双变量响应曲面揭示特征间的协同作用机制。主效应分析显示，论文数量差异等特征数值与SHAP值呈现非对称性映射关系。这种非线性源于SHAP值的计算过程捕捉了特征间的交互效应，特定特征的贡献度会受到协变量动态调节，以及不同特征可能通过差异化路径影响合作频次的预测分类。因此，本节通过对相关特征进行交互效应分析，进一步探究影响合作频次的关键特征组合及其潜在规律。首先，基于斯皮尔曼相关系数评估特征间关联性，筛选具有显著相关性的特征对，相关性热力矩阵如图7所示。其中，研究主题相似度与论文数量差异、总被引次数差异、均被引次数差异、H指数差异均存在显著负相关，总被引次数差异、均被引次数差异、论文数量差异与H指数差异两两之间均存在显著正相关，H指数差异还和学术年龄差异存在显著正相关。其次，根据特征

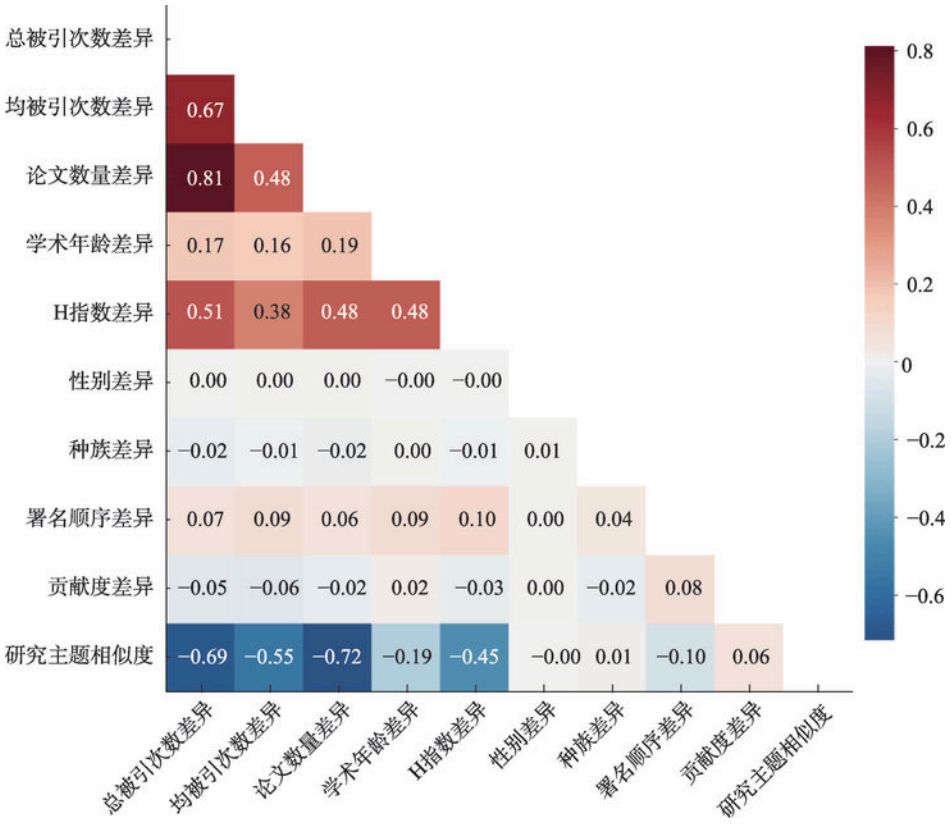


图7 特征相关性热力矩阵

之间的相关性计算结果，对两两特征的交互效应展开讨论。

通过筛选具有显著相关性的两两特征交互效应结果，发现其中存在分布结构相近的特征对，限于篇幅，这里仅列举部分具有代表性、结果显著的交互效应图，如图 8 所示。其中，第一列的子图均为低频合作下的特征交互图，第二、第三列依次为中频与高频合作，横轴与颜色条表示相应特征的数值，纵轴为 SHAP 值。整体来看，研究主题相似度、总被引次数差异、均被引次数差异、论文数量差异等特征之间的交互效应图结构比较相似，低频合作与高频合作的散点图以 SHAP 值 0 为中线对称分布，两者的拼接为中频合作的散点图。

由图 8a~图 8c 可知，随着论文数量差异的增大，研究主题相似度在多个数值区间对模型预测产生显著的正向或负向影响，呈现明显的非线性交互模式。

在低频合作中，论文数量差异较大与研究主题相似度小的特征组合可以显著促进模型预测。这一发现与 *Proceedings of the National Academy of Sciences of the United States of America*（*PNAS*）发表的一项大规模合作演化研究相呼应，该研究追踪 280 万名科学家的合作主题演化，发现高产学者倾向于拥有更高比例的单主题合作者[61]。而其单主题合作者群体中可能存在两类截然不同的作者，一方面，部分低生产力作者通过参与单主题合作逐步积累经验，如短暂师生合作；另一方面，高生产力作者可能通过选择性合作，将资源集中在高潜力项目或关键合作伙伴上，以完成跨学科攻关、大型临床试验等特定任务或项目。这呼应了资源依赖理论中“通过外部资源整合降低不确定性”的核心逻辑，更揭示学术合作网络中权力与资源分配的动态博弈。

在中频合作中，论文数量差异与研究主题相似度的交互效应呈现不确定性。即同研究主题相似度的情况下，论文数量差异较大对模型同时存在正负双向影响；但当主题相似度提升至一定区间，论文

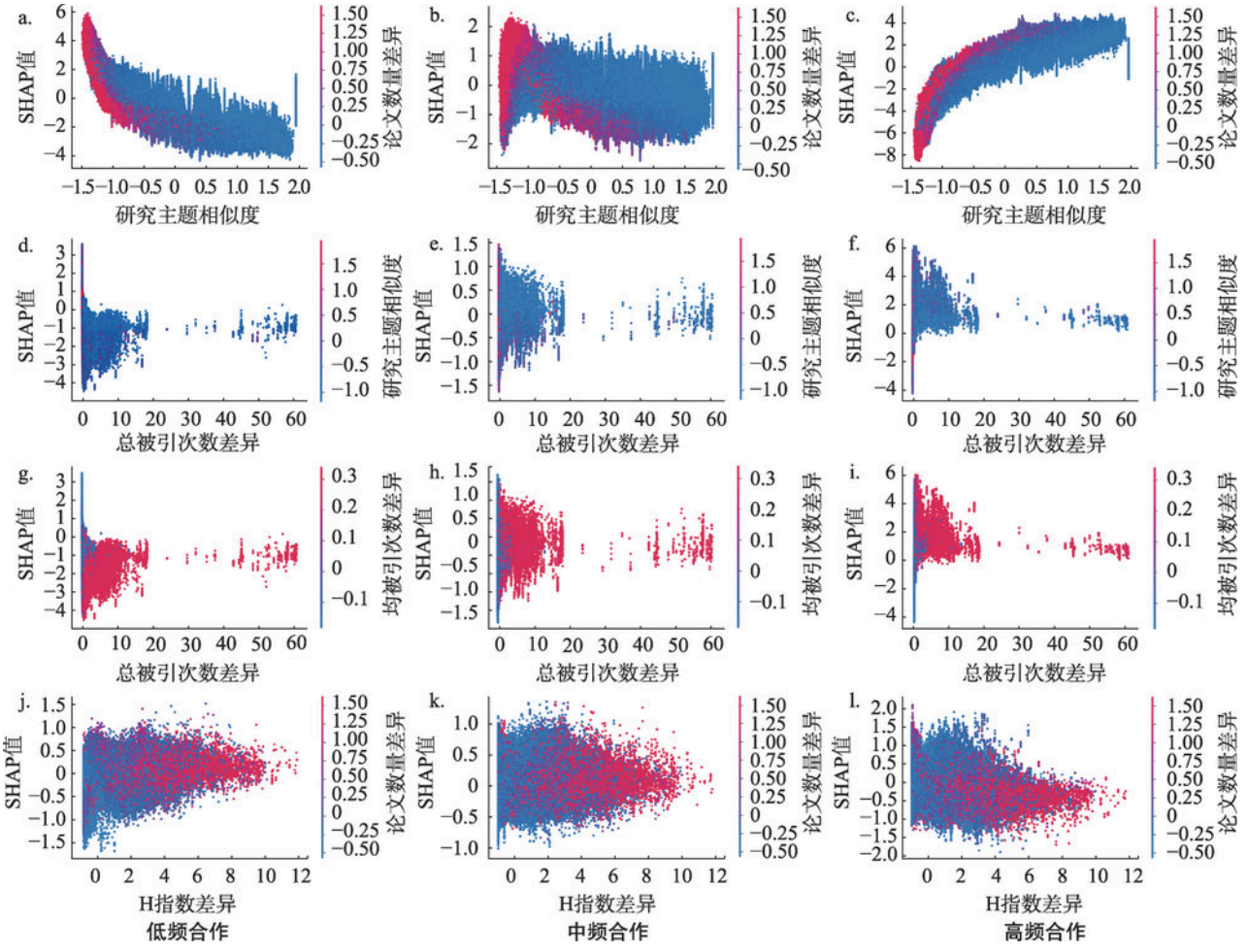


图 8　相关特征对的交互效应图（彩图请见 https://qbxb.istic.ac.cn）

数量差异较大则倾向于促进模型预测；当主题相似度继续升高时，论文数量差异增加反而削弱了模型预测。这种复杂的非线性关系印证了合作研究中常见的阈值效应现象，已有研究表明，当合作历史超过 3 次时，信任因素对合作频次提升的边际贡献显著衰减[62]。

在高频合作中，红色区域的分布类似于抛物线，论文数量差异大对模型预测主要为抑制作用，多数高频合作者之间往往具有高的研究主题相似度和较低的论文数量差异。主题相似度能够降低合作初期的沟通与理解成本，而相近的论文数量差异暗示着双方拥有可匹配的研究经验。例如，在生物医学领域，高频合作的团队通过共享实验范式与术语体系，从而有效缩短知识整合周期。

综上，低频合作者之间的主要特征为研究主题相似度低和论文数量差异大，典型场景为跨领域项目初期合作。高频合作者之间表现为研究主题高度契合和论文数量差异均衡，学术共同体特征显著，合作关系稳定。相比之下，中频合作的特征模式则呈现从低频向高频过渡的迹象，反映出合作关系可能存在的演化路径：合作初期往往体现出资源互补或探索性倾向，随着合作次数累积与信任机制逐步巩固，双方的研究主题可能逐渐趋于集中，合作模式逐步向稳定结构演进。中频合作呈现的波动特征也表明，该阶段可能融合了多种合作动机，如延续性项目协作、师生指导关系或初步建立的伙伴关系等，因而其稳定性通常低于高频合作类别。

由图 8d~图 8i 可知，横坐标均为总被引次数差异，不同合作频次类别下的研究主题相似度均偏低，而均被引次数差异则偏高。在低频合作中，总被引次数差异小结合低主题相似度或均被引次数差异较大时，对模型预测产生抑制作用。当两个影响力相当的学者跨领域合作时，这种合作往往具有其特定价值（如重大交叉问题），其合作可能并非最典型的偶然单次合作，或者暗示这种组合本身就更可能发展为多次合作，从而模型不倾向于将其预测为低频。在高频合作中，总被引次数差异小结合低主题相似度或均被引次数差异大时，则对模型预测产生促进作用。即使研究者之间存在领域专长分散或单篇成果影响力不均的情况，但其整体学术影响力相当，由多次合作所累积的信任与协同效能仍然能有效维持合作的稳定性。这证明高频合作更多依赖于在异质性基础上构建的互补机制，而非静态的、同质化的合作者特征。鉴于 H 指数差异和学术年龄差异、均被引次数差异、研究主题相似度等特征间的交互效应结果，与论文数量差异的分布形态具有高度一致性，故图 8j~图 8l 仅展现 H 指数差异与论文数量差异的交互效应图。上述特征对的中频合作效应图主要围绕 SHAP 值为 0 的中线对称分布，并且不同颜色点呈现明显的分散状态，无法得出显著影响模型变化的结论。综上，本研究探讨的合作频次影响因素主要体现在特征重要性、分布趋势的分化、高频合作的独特性以及中频合作的多样性上，共同揭示学术合作中复杂多变的互动关系。

5 总 结

本研究基于生物医学领域的大规模合著数据，通过 XGBoost-SHAP 可解释的算法归因框架，系统识别并量化不同科研合作频次背后的关键影响因素及其作用机制，为该领域合作行为的动态解析提供新的方法与视角。首先，基于学者间已有的合作关系构建多维度特征体系，实现对合作频次的精准分类预测；其次，借助 SHAP 值对特征因素在分类预测中的贡献进行量化评估，揭示各特征在低、中、高合作频次类别中的贡献方向、阈值变化与交互效应。

本研究发现，不同特征因素对合作频次的影响程度存在显著差异。研究主题相似度是影响合作频次的关键因素，其次为论文数量差异、总被引用次数差异和均被引次数差异，而学术年龄差异、H 指数差异、贡献度差异和署名顺序差异则影响较小。具体而言：①在低频合作中，研究主题相似度呈现促进与抑制并存的双重效应，反映出该类别下既存在领域相近的偶发性合作，也存在跨领域的探索性协作；在高频合作中，高主题相似度显著提升模型的预测概率，证明其是维持高频稳定合作的核心条件，并与低频和中频合作形成明显区分。②论文数量差异小对于模型识别低频和中频合作具有正向影响，反映作者间学术产出的均衡性；在高频合作中则为负向影响，揭示该合作类别下可能存在以互补为导向的非对称合作结构。③引用特征方面，初次合作可能更关注总被引次数所体现的学术积累，而高频合作则注重均篇被引次数所代表的成果质量一致性。④由特征因素的主效应分析可知，作者间研究主题相似度、论文数量差异与 H 指数差异的作用机制分别为双峰对称型、阈值稳定型及边际递减型。⑤基于交互效应的合作模式识别表明，低频合作常伴随主题差异与论文产出规模不均；高频合作以高度主题契合与产出均衡为典型特征，呈现较强

的学术共同体属性；中频合作则表现出由低频向高频过渡的动态演变特征。综上，高频合作对知识结构、影响力构成等异质性因素展现出一定的包容性，其背后往往形成了以互补为核心、以信任为基础的有效分工机制。本研究从特征驱动视角揭示了科研合作频次背后的内在逻辑，为理解合作动态与优化科研组织模式提供了实证依据。

本研究的结论主要反映生物医学领域科研合作频次的影响因素及其作用机制，为挖掘不同合作频次类别背后的内在逻辑提供了新思路。然而，该领域具有合作规模相对较大、受重大公共事件驱动显著、数据共享文化较成熟等特点，如新冠疫情的暴发可能导致研究主题相似度的重要性在数据中被强化。不同学科因其知识生产模式、合作文化、数据可获得性的差异，合作频次的关键驱动因素及其作用机制可能有所不同。例如，数学领域可能更注重个体思维深度，小规模长期合作模式占主导，主题相似度和研究产出的作用模式或与本研究结论存在差异。另外，本研究仅侧重于文献计量指标的定量归因，可能忽略了人际信任、学术网络嵌入等难以量化的隐性社会资本因素。后续研究可拓展到多学科数据，系统揭示合作频次影响因素与长效作用机制，并融合定性研究方法，以深化对科研合作关系与模式的理解。

参　考　文　献

[1] Fatt C K, Abu Ujum E, Ratnavelu K. The structure of collaboration in the journal of finance[J]. Scientometrics, 2010, 85(3): 849-860.

[2] 曾粤亮, 司莉. 跨学科科研合作: 背景、理论研究与实践进展[J]. 图书情报工作, 2021, 65(10): 127-140.

[3] Song X, Zhang Y, Pan R, et al. Link prediction for statistical collaboration networks incorporating institutes and research interests [J]. IEEE Access, 2022, 10: 104954-104965.

[4] Pisani N, Boekhout H D, Heemskerk E M, et al. China's rise as global scientific powerhouse: a trajectory of international collaboration and specialization in high-impact research[J]. Research Policy, 2025, 54(8): 105288.

[5] Hou J, Pan H X, Guo T, et al. Prediction methods and applications in the science of science: a survey[J]. Computer Science Review, 2019, 34: 100197.

[6] Newman M E J. Coauthorship networks and patterns of scientific collaboration[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(suppl_1): 5200-5205.

[7] Bozeman B, Corley E. Scientists' collaboration strategies: implications for scientific and technical human capital[J]. Research Policy, 2004, 33(4): 599-616.

[8] Leahey E, Reikowsky R C. Research specialization and collaboration patterns in sociology[J]. Social Studies of Science, 2008, 38 (3): 425-440.

[9] Leahey E. From sole investigator to team scientist: trends in the practice and study of research collaboration[J]. Annual Review of Sociology, 2016, 42: 81-100.

[10] Liu J W, Guo X F, Xu S, et al. A new interpretation of scientific collaboration patterns from the perspective of symbiosis: an investigation for long-term collaboration in publications[J]. Journal of Informetrics, 2023, 17(1): 101372.

[11] 刘飞, 毛进, 李纲. 科学计量学领域科研合作特征及空间集聚模式研究[J]. 情报科学, 2022, 40(1): 166-175.

[12] 刘晓婷, 黄颖, 李瑞婧, 等. 内聚-耦合视角下科研团队合作模式识别与对比研究[J]. 情报科学, 2022, 40(12): 170-180.

[13] Gazni A, Sugimoto C R, Didegah F. Mapping world scientific collaboration: authors, institutions, and countries[J]. Journal of the American Society for Information Science and Technology, 2012, 63(2): 323-335.

[14] Finardi U, Buratti A. Scientific collaboration framework of BRICS countries: an analysis of international coauthorship[J]. Scientometrics, 2016, 109(1): 433-446.

[15] 欧桂燕, 岳名亮, 吴江, 等. 杰出青年科研人员学术职业生涯科研合作特征演变分析——以化学领域为例[J]. 情报学报, 2021, 40(7): 756-767.

[16] 刘晓娟, 余梦霞, 赵卓婧, 等. "长江学者奖励计划"资助前后科研合作变化与分析——以2005年度长江学者为例[J]. 图书情报工作, 2020, 64(3): 100-110.

[17] 曹仁猛, 谢维熙, 耿屿, 等. 新冠疫情暴发前后我国国际科研合作模式的变化——基于bioRxiv的分析[J]. 图书情报知识, 2022, 39(3): 41-49.

[18] 任妍, 杨金庆. 科研合作角色演化模式及其外生驱动因素研究[J]. 情报资料工作, 2025, 46(2): 37-45.

[19] Petersen A M. Quantifying the impact of weak, strong, and super ties in scientific careers[J]. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112(34): E4671-E4680.

[20] Bu Y, Ding Y, Liang X K, et al. Understanding persistent scientific collaboration[J]. Journal of the Association for Information Science and Technology, 2018, 69(3): 438-448.

[21] Bu Y, Murray D S, Ding Y, et al. Measuring the stability of scientific collaboration[J]. Scientometrics, 2018, 114(2): 463-479.

[22] Wu L Y, Yi F, Bu Y, et al. Toward scientific collaboration: a cost-benefit perspective[J]. Research Policy, 2024, 53(2): 104943.

[23] Wang G F, Gan Y T, Yang H D. The inverted U-shaped relationship between knowledge diversity of researchers and societal impact[J]. Scientific Reports, 2022, 12: Article No.18585.

[24] 马荣康, 李真真. 高被引还是零被引: 基于论文被引的最佳科

研合作规模研究——来自 Financial Times TOP 45 商学院期刊的证据[J]. 情报学报, 2020, 39(11): 1182-1190.

[25] Bozeman B, Gaughan M. How do men and women differ in research collaborations? An analysis of the collaborative motives and strategies of academic researchers[J]. Research Policy, 2011, 40(10): 1393-1402.

[26] 张明妍. 科研人员学术合作的性别差异[J]. 科学学研究, 2025, 43(6): 1197-1205.

[27] Wang W, Yu S, Bekele T M, et al. Scientific collaboration patterns vary with scholars' academic ages[J]. Scientometrics, 2017, 112(1): 329-343.

[28] Benckendorff P, Zehrer A. Career and collaboration patterns in tourism research[J]. Current Issues in Tourism, 2016, 19(14): 1386-1404.

[29] 郭宇森, 杨艳萍. 国际科研合作网络视角下关键联络人识别模型构建[J]. 情报学报, 2024, 43(10): 1142-1153.

[30] 刘苗苗, 姜华, 刘盛博, 等. 不同学科科研合作差异的比较研究——以 2017 年教育部创新团队 114 位带头人为例[J]. 科技管理研究, 2019, 39(16): 100-107.

[31] Liu Z F, Wang C L, Yang J Q. The effects of scientific collaboration network structures on impact and innovation: a perspective from project teams[J]. Journal of Informetrics, 2025, 19(1): 101611.

[32] Gui Q C, Xu W, Jiang S D, et al. Unpacking the dynamics of international research collaboration network: structural effects and dyadic effects[J]. Technology in Society, 2025, 82: 102954.

[33] Tao J, Zhou L N, Hickey K. Making sense of the black-boxes: toward interpretable text classification using deep learning models [J]. Journal of the Association for Information Science and Technology, 2023, 74(6): 685-700.

[34] Ayoub J, Yang X J, Zhou F. Combat COVID-19 infodemic using explainable natural language processing models[J]. Information Processing & Management, 2021, 58(4): 102569.

[35] 安璐, 陈苗苗. 突发事件情境下政务微博信息发布有效性评估[J]. 情报学报, 2022, 41(7): 692-706.

[36] 崔蕴学, 王贤文, 王勇臻. 基于归因分析的引用模式挖掘及其实证研究[J]. 情报学报, 2023, 42(4): 381-392.

[37] Chen T Q, Guestrin C. XGBoost: a scalable tree boosting system [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 785-794.

[38] Zhang Y L, Feng T, Wang S D, et al. A novel XGBoost method to identify cancer tissue-of-origin based on copy number variations [J]. Frontiers in Genetics, 2020, 11: 585029.

[39] Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms[J]. Artificial Intelligence Review, 2021, 54(3): 1937-1967.

[40] Borisov V, Leemann T, Seßler K, et al. Deep neural networks and tabular data: a survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(6): 7499-7519.

[41] Lundberg S M, Lee S I. A unified approach to interpreting model predictions[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2017: 4768-4777.

[42] Shapley L S. A value for n-person games[M]// Contributions to the Theory of Games. Princeton: Princeton University Press, 2016: 307-318.

[43] Maschler M, Peleg B, Shapley L S. The kernel and bargaining set for convex games[J]. International Journal of Game Theory, 1971, 1(1): 73-93.

[44] Lundberg S M, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees[J]. Nature Machine Intelligence, 2020, 2(1): 56-67.

[45] Xu J, Yu C, Xu J W, et al. PubMed knowledge graph 2.0: connecting papers, patents, and clinical trials in biomedical science[J]. Scientific Data, 2025, 12: Article No.1018.

[46] Karimi F, Wagner C, Lemmerich F, et al. Inferring gender from names on the web: a comparative evaluation of gender detection methods[C]// Proceedings of the 25th International Conference Companion on World Wide Web. Republic and Canton of Geneva: International World Wide Web Conferences Steering Committee, 2016: 53-54.

[47] Mueller J, Stumme G. Gender inference using statistical name characteristics in Twitter[C]// Proceedings of the 3rd Multidisciplinary International Social Networks Conference on SocialInformatics. New York: ACM Press, 2016: Article No.47.

[48] Shen H Q, Xie J, Li J, et al. The correlation between scientific collaboration and citation count at the paper level: a meta-analysis [J]. Scientometrics, 2021, 126(4): 3443-3470.

[49] Larivière V, Ni C Q, Gingras Y, et al. Bibliometrics: global gender disparities in science[J]. Nature, 2013, 504(7479): 211-213.

[50] Benenson J F, Markovits H, Wrangham R. Rank influences human sex differences in dyadic cooperation[J]. Current Biology, 2014, 24(5): R190-R191.

[51] 杨杰, 孔嘉, 张艺炜, 等. 融合论文颠覆性与巩固性的学者二元影响力测度[J]. 情报学报, 2023, 42(12): 1412-1423.

[52] 杨杰, 王左戎, 邓三鸿, 等. 基于参考文献的论文跨学科性、跨时域性及其影响力研究[J]. 情报学报, 2024, 43(9): 1003-1014.

[53] Wang J, Thijs B, Glänzel W. Interdisciplinarity and impact: distinct effects of variety, balance, and disparity[J]. PLoS One, 2015, 10(5): e0127298.

[54] Conley J P, Önder A S. The research productivity of new PhDs in economics: the surprisingly high non-success of the successful[J]. The Journal of Economic Perspectives, 2014, 28(3): 205-215.

[55] Hagen N T. Harmonic allocation of authorship credit: source-level correction of bibliometric bias assures accurate publication and

citation analysis[J]. PLoS One, 2008, 3(12): e4021.

[56] Wang X F, Li R R, Ren S M, et al. Collaboration network and pattern analysis: case study of dye-sensitized solar cells[J]. Scientometrics, 2014, 98(3): 1745-1762.

[57] Larivière V, Desrochers N, Macaluso B, et al. Contributorship and division of labor in knowledge production[J]. Social Studies of Science, 2016, 46(3): 417-435.

[58] 张琳, 孙蓓蓓, 黄颖 . 跨学科合作模式下的交叉科学测度研究——以 ESI 社会科学领域高被引学者为例[J]. 情报学报, 2018, 37(3): 231-242.

[59] Tuninetti M, Aleta A, Paolotti D, et al. Prediction of new scientific collaborations through multiplex networks[J]. EPJ Data Science, 2021, 10: Article No.25.

[60] Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values[J]. Artificial Intelligence, 2021, 298: 103502.

[61] Zeng A, Fan Y, Di Z R, et al. Impactful scientists have higher tendency to involve collaborators in new topics[J]. Proceedings of the National Academy of Sciences of the United States of America, 2022, 119(33): e2207436119.

[62] Boudreau K J, Guinan E C, Lakhani K R, et al. Looking across and looking beyond the knowledge frontier: intellectual distance, novelty, and resource allocation in science[J]. Management Science, 2016, 62(10): 2765-2783.

（责任编辑　魏瑞斌）