研究论文
DOI：10.13998/j.cnki.issn1002-1248.23-0214

基于自适应特征选择的科研论文跨学科性测度方法研究

王晋飞[1]，孙　巍[1,2*]，张学福[1,2]，杨　璐[1]
（1. 中国农业科学院农业信息研究所，北京　100081；2. 农业农村部 农业大数据重点实验室，北京　100081）

摘　要：［目的 / 意义］跨学科研究能够通过知识整合和渗透，创造性地解决自然环境和人类社会中的复杂问题。随着跨学科研究成果的大量增长，跨学科性测度评估变得越来越有必要，如何构建有效的跨学科性测度方法，实现对论文跨学科性综合全面的测度是亟待解决的问题。［方法 / 过程］本研究首先基于跨学科研究的内涵和特点，从学科属性、知识网络拓扑结构和知识整合文本内容 3 个维度提取科研论文跨学科性特征指标，并给出特征指标的计算方法；其次，对跨学科性特征指标进行自适应计算，构建一种基于机器学习的跨学科性测度方法；最后，以植物纳米生物技术领域为例进行实证研究，对领域中高跨学科性的论文进行识别和筛选。［结果 / 结论］本文提出的自适应特征选择能够对跨学科性相关特征指标进行有效筛选，提升结果的可靠性，实现全面、深入的科研论文跨学科性测度。该测度方法避免了定性评估可能会出现的主观性缺陷以及不同测度指标可能出现相互矛盾结果的问题，为跨学科性测度提供新的思路与方向。

关键词：跨学科性；自适应特征选择；论文测度
中图分类号：TP391.1；G353.1　　**文献标识码：**A　　**文章编号：**1002-1248（2023）03-0052-19

引用本文：王晋飞，孙巍，张学福，等. 基于自适应特征选择的科研论文跨学科性测度方法研究[J]. 农业图书情报学报，2023，35(3)：52-70.

1　引　言

跨学科研究是取得重大科学发现和产生原创成果重大突破的重要方式，也是提升创新能力的重要途径。回顾国内外近现代科学的发展历程，许多重大科学进展包括纳米技术、富勒烯、生物医学工程等都产生于不同学科之间的融合交叉以及相互渗透[1]。同时，跨学科团队合作是解决自然环境和人类社会复杂问题的重要途径，也是科技创新的重要手段，比如在全球性重

收稿日期：2023-02-04
基金项目：NSTL 文献专项任务基金项目“下一代开放知识服务平台关键技术优化集成与系统研发”（2022XM28）子课题“基于专利的领域技术分析与预测系统集成”；中国农业科学院基本科研业务费专项课题“农业研究热点前沿探测及领域技术布局分析研究”（Y2022ZK07）
作者简介：王晋飞（1999-），女，硕士，研究方向为信息管理与信息系统。张学福（1966-），男，博士，研究员，博士生导师，研究方向为信息资源管理。杨璐（1988-），男，博士，研究方向为数据分析与模型计算。
*通信作者：孙巍（1973-），女，博士，研究员，博士生导师，研究方向为信息检索可视化、数据挖掘、知识组织。Email：sunwei@caas.cn

大突发公共卫生事件影响下，科学界强调多学科知识与方法的融合与应用，呼吁通过跨学科合作实现联合攻关[2]。

跨学科研究已经得到了研究人员和科学政策的重视，许多国家制定了各种政策和资助计划，以鼓励跨学科研究。欧盟发布了“跨学科研究探索”[3]；美国 INSPIRE 计划、NSF 会聚项目及 OLA、OMA 等多个跨学科管理部门，也高度重视跨学科研究（IDR）的融资[4]；推动跨学科性研究也是中国“科技强国”战略的关键一环，2020 年中国正式在自然科学基金委下设“交叉科学部”，以促进学科的进一步交叉融合；2022 年 1 月 1 日起实施《中华人民共和国科学技术进步法》，其中也提到“国家鼓励科学技术研究开发与高等教育、产业发展相结合，鼓励学科交叉融合和相互促进”“国家完善学科布局和知识体系建设，推进学科交叉融合，促进基础研究与应用研究协调发展”[5]。

跨学科性的测度是跨学科研究中的重要问题，是从目的性、学科性和整合性等角度分析不同学科知识深度融合与扩散的程度，从而实现对跨学科数据的定量分析。随着跨学科研究成果大量增长，跨学科性评估变得越来越有必要。在此背景下，学界不断加深对跨学科性这一概念的理解，对跨学科性的研究提出了多方面的度量指标，通过对相异学科知识在深度融合过程中发生交叉的程度进行度量，以期将其更好地应用于科研优化、科研管理、国家政策制定和科研资助规划等。随着研究的不断深入，选择科研论文作为跨学科性研究对象具有操作性强、解释性高、粒度合适、适用范围广等特点，逐渐成为关注的重点，学者们从不同理论视角出发，产出了许多跨学科性指标测度，用于测度论文跨学科性程度、发现跨学科知识融合态势等，对创新决策等具有重要参考价值，但是，多数跨学科性测度体系仅停留在理论层面，或由于指标难以获取而缺乏数据支持，不同指标之间缺乏有效的组织，难以综合地测度跨学科性。立足于此，本研究将不同视角的科研论文跨学科性指标进行综合，通过自适应的方法进行特征选择，开展科研论文跨学科性的测度研究，以辅助科技决策、跨学科创新评估等工作。

2　相关研究

2.1　跨学科性特征指标

跨学科性是衡量学科交叉强弱的一个综合指标[6]，也称为学科交叉性。目前，对科研论文的跨学科性测度主流研究方法主要有基于文献计量、社会网络和文本内容的方法。

在文献计量分析方面，学者不断从学科丰富度、均衡度、差异度等角度全面度量跨学科性程度。信息熵指标由于能较好地综合表达论文参考文献所属学科数量的丰富性与分布的均衡性，一度成为国内外学者常用的学科跨学科性指标之一，但是，STIRLING[7]对比生物学中多样性理论，认为客观表征跨学科性还需包含学科差异度。因此，为更好表征论文的跨学科性，Rao-Stirling 指标[8]全面考虑了论文所涉及学科的数量、学科分布的均匀程度以及学科性质的差异，在国际和国内跨学科性研究中产生了较大的影响，成为跨学科性测度领域的热门指标。ZHANG 等[9]提出 Ture Diversity（TD）以解决 Rao-Stirling 指标在测度跨学科性程度中存在的较低区分度问题。LEYDESDORFF[10]将 Rao-Stirling（RS）指标修改为新的 DIV 和 DIV* 指标，解决了 RS 指标由于事先定义的丰富性和平衡性而导致的结果与实际情况不符的反常现象。

在社会网络分析方面，学者利用引文、科研合作、共被引等数据构建相应网络，从网络凝聚性、节点属性等角度衡量跨学科性程度，了解研究中已整合知识的程度。其中，凝聚性用来表示科研论文在网络结构中的关系强度[11]，可细分为网络的密度，强度和差异度 3 个方面[12]，也可以结合时间导数评估凝聚性的变化，了解跨学科研究中知识整合的动态过程。LEYDESDORFF[13]使用网络密度指标，节点总数、最大连通网络节点数、中介中心性等辅助指标进行凝聚性分析，从而测度跨学科情况。

也有学者基于文本内容进行跨学科性测度，基于文献计量和社会网络可以有效反映跨学科研究的学科结构与知识来源，而基于文本内容的方法与这两种方

DOI：10.13998/j.cnki.issn1002-1248.23-0214

法相比，可以更直观地表达跨学科内容，体现知识的质量和价值属性。其中，XU 等[14]提出了一种新的跨学科主题挖掘度量指标 TI，通过计算 TI 值、Bet 值、词频值等来识别知识融合领域，黄菡[15]根据复杂网络的原理提出了 TIc 值及其计算方式，从主题的多样性和凝聚性两个方面对主题进行测度。

针对论文跨学科性测度指标研究的深度和广度正在加强，但目前研究仍存在不足之处。一是论文跨学科性测度指标仅从引用、网络关系和文本内容等单一视角出发，无法对论文的跨学科性进行综合性评判。二是已有指标多从参考文献视角测度跨学科性，根据文献内容提取的跨学科性测度指标较少，相关研究还处于简易的层面，无法直观表达跨学科内容，忽视了论文的知识质量和价值属性。因此，本研究从论文外部学科属性、网络拓扑结构和论文内部文本语义角度对科研论文的跨学科性进行解构，对跨学科性进行综合全面的测度。

2.2 特征选择方法

特征选择方法是从特征集中找到最大程度表征原始特征且具有较高准确度的最小特征子集[16]，从而简化学习模型，加快算法学习过程。传统的文本特征选择方法主要有：卡方检验（CHI）、信息增益（IG）、互信息（MI）、文档频率（DF）等。为提高模型精度和最大程度表示原始特征，学者们在特征筛选方面开展了深入的研究。一方面，为提升特征选择模型的精度，CHEN 等将 LDA（Latent Dirichlet Allocation）、决策树、粗糙集以及 F-score 方法和 SVM 结合构建模型[17]，提升了单个 SVM 模型的性能；另一方面，为最大程度表示原始特征，VLASSELAER 等[18]提出了关注数据内在特征和网络特征的特征提取方法。而对于不同的数据类型，熊志斌[19]构建了 GCFS-SVM（Gebelein CFS SVM）模型，对非线性数据进行有效的特征提取。DAHIYA 等[20]将特征选择和混合 Bagging（Bootstrap Aggregating）模型结合对数值型数据和非数值型数据进行特征筛选。近年来，特征选择方法在论文原创性、创新性、专利价值属性测度等方面得到了广泛应用。赵蕴华等[21]采用决策树、支持向量机和神经网络对专利价值相关特征进行评估。何向等[22]进一步通过支持向量机和随机森林对比，对高校专利价值相关特征进行特征选择。李欣等利用 SVM、ANN、RF 和 AdaBoost 组合的方法进行特征选择，分别用于筛选高质量专利重要指标[23]、论文中研究前沿重要特征[24]。钱玲飞等[25]在相关性分析和特征选择的基础上，筛选与学术创新力相关的重要特征。同时，特征识别在文本分类、文本聚类等方面也具有广泛的应用。

但是传统的基于完全搜索策略、启发式策略和随机搜索策略的特征选择不适用于具有高维度、强关联性特点跨学科性特征指标。而随着自适应算法的快速发展，越来越多学者将其应用在特征筛选中，自适应是一个不断以优化决策参数去实现目标的过程，这类算法对不同环境的适应能力非常强[26]。SHAFIQ 等[27]提出了一种名为 WMI_AUC 的基于机器学习的自适应特征选择算法，利用加权互信息方法和 ROC 曲线下面积从流量的所有特征中选择出最有效的特征，刘凯等[28]通过采用合适的自适应算法进行特征选择以挑选出每个决策树的特定特征并自动融合多维度的特征。

由于跨学科性具备多维的概念，因此，从不同角度和层次进行指标选择，并对不同指标进行有效组合可以更综合、客观地度量跨学科性的不同属性。以往的人工筛选存在一定主观性，会直接影响评价结果，而特征选择作为机器学习方法，能够对多维性数据进行自动筛选，利用较少的变量最大程度反映跨学科性，简化模型结构，提高模型预测精度和准确度。自适应特征选择作为特征选择的一种有效手段，可以用于处理跨学科性特征指标的高维性和高相关性问题，提升跨学科性特征指标的可解释性。因此，本文基于自适应的特征选择方法，对论文的跨学科性特征指标进行自动化筛选，同时针对不同的学科特点，构建符合该领域的测度指标，提出一种更加合理有效的科研论文跨学科性测度方法。

3 论文跨学科性测度方法

由于跨学科研究本身的复杂性和多样性，跨学科

性测度研究同样呈现出较为复杂的情况。不同的测度视角、方法、指标等，在测度同一个研究对象时可能得到不同的结果，同时数据的差异和变化也会影响模型的可迁移性。本文提出一种基于自适应特征选择的论文跨学科性测度方法，基本流程如图 1 所示。从跨学科性内涵出发，从学科属性、知识网络拓扑结构和知识整合内容特征 3 个维度构建跨学科性原始特征集，在原始特征集中通过自适应特征选择结合机器学习的方法筛选出最优特征子集，通过评价函数对该最优特征子集进行评价，最终实现对科研论文跨学科性的测度计算。

3.1　跨学科性原始特征集构建

为全面度量科研论文的跨学科性，本文根据跨学科研究的广泛定义，即“以团队合作或独立个体为基础，结合两个或两个以上的学科或专业知识群体的信息、数据、技能、工具、观点、概念和理论的科学研究方式，来加深对解决一个学科或研究领域无法解决的问题的基本理解”[29]，同时梳理了科研论文跨学科性测度所涉及的跨学科性驱动力、跨学科性模式、跨学科性影响因素、跨学科性评价、跨学科性机理研究等，构建了一个包含了知识整合学科属性、网络结构、文本内容等要素的跨学科性分析框架，明确跨学科性测度中需要考虑 3 个维度，如表 1 所示。

（1）学科属性特征。跨学科研究主要属性是来自不同学科知识的融合渗透，主要依托于相关参考文献，不同学科分布的高质量参考文献能为研究提供有价值的知识理论和方法体系[30]。因此，本文将参考文献的相关学科属性特征作为测度指标之一，评估论文在跨学科研究过程中对不同学科知识的吸收和整合程度。为衡量科研论文的知识整合程度，客观表征论文的跨学科性，本文除了选择学科均衡度、差异度和丰富度等重要指标外，综合考量了 Rao-Stirling 指标[7]、Ture Diversity（TD）[9]以及 Div* 指标[10]，以测度科研论文学科多样性。

（2）知识网络结构的拓扑特征。跨学科研究中的知识可以移动或传播到以前未曾使用过的知识体系，主要体现在引文扩散到不同领域的情况[31]。因此，本文通过选择文献被引用的网络特征进行量化分析，剖析跨学科知识扩散程度，利用引文网络，从网络密度、凝聚性以及中介中心性的角度[11-13]，衡量科研论文的知识扩散程度。

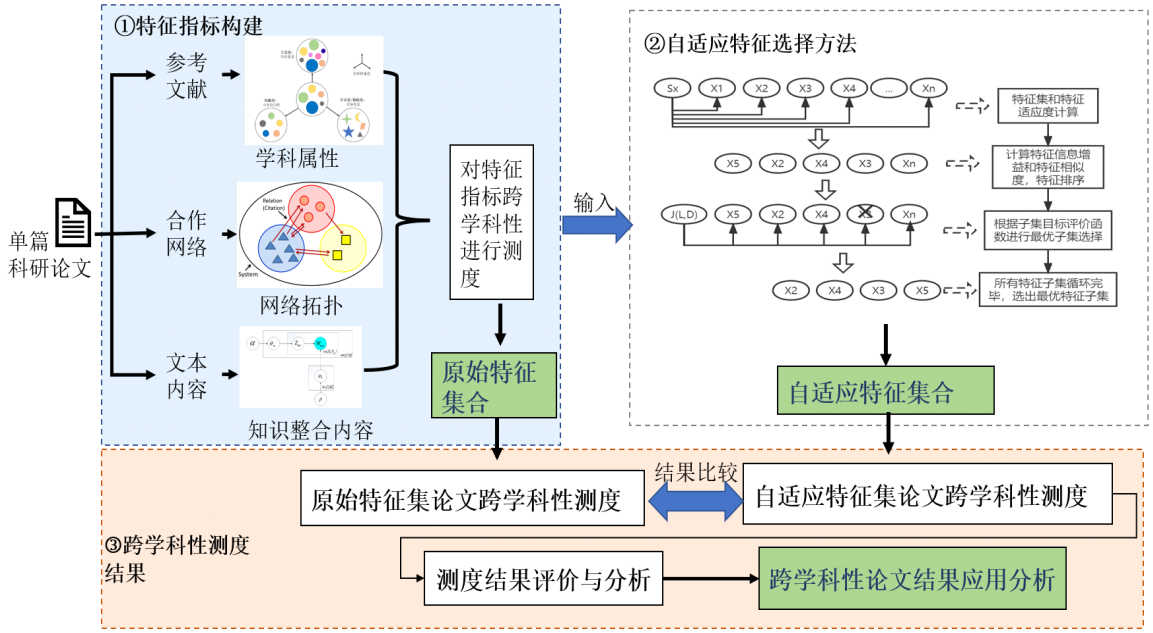


图 1　科研论文跨学科性测度图

Fig.1　Measurement chart of interdisciplinarity of research papers

DOI：10.13998/j.cnki.issn1002-1248.23-0214

表 1　跨学科性测度原始特征集

Table 1　Original feature set of interdisciplinarity measure

要素维度	特征指标	测量公式		公式描述	测度作用
学科属性特征	丰富度	$\text{variety} = \text{学科类别数} = v$	(1)	v 为数据集文献参考文献所归属学科类别数量[7]	用于测度研究领域覆盖面的宽窄程度
	均衡度	$1 - Gini = 1 - \frac{\sum_i (2i - v - 1)x_i}{v \sum_i x_i}$	(2)	i 是类别序号，x_i 是属于第 i 个类别的学科类别数量，类别依据 x_i 由小到大进行升序排列[7]	用于测度研究领域学科分布的均匀程度
	差异度	$disparity = \frac{1}{v(v-1)} d_{ij}$	(3)	d_{ij} 是学科距离，需要一个先验的学科分类系统对学科进行划分，在学科分类的基础上通过构建学科引用网络或学科共被引网络，对学科引用矩阵进行相似度计算[7]	用于测度参考文献所在学科分布的差异程度
	Rao-Stirling	$\sum_{\substack{i,j=1 \\ i \neq j}}^{n} (p_i p_j) d_{ij}$	(4)	用于测度学科多样性，n 为该系统所包含的元素个数，p_i，p_j 表示元素 i，j 在该系统中所占的比例，d_{ij} 表示元素 i 和元素 j 之间的距离[8]	用于测度学科多样性
	True Diversity	$\frac{1}{\sum (1 - d_{ij})(p_i p_j)}$	(5)	相对于 Rao-Stirling 指标的表现更优秀且具有较高的区分度[9]	用于测度学科多样性
	Div*	$n_c * (1 - Gini_c) * \sum_{i=1, j=1, i \neq j}^{i=n_c, j=n_c} d_{i,j}$	(6)	针对 DIV 指标中存在的异议，进行修正后得到的最新 DIV*指标[10]	用于测度学科多样性
知识网络拓扑特征	凝聚性	$Coherence = \sum_{ij(i \neq j)} p_{ij} d_{ij}$	(7)	将共被引作为学科之间链接的基础，凝聚性代表网络结构紧凑的程度[11]	用于测度网络结构紧凑的程度
	中介中心性	$BC = \sum_{j,k} \frac{b_{ijk}}{b_{jk}}$	(8)	任意两个节点需要通过某节点的最短路径数量与两个节点所有最短路径的比值，衡量网络中经过特定节点的捷径数[12]	用于测度节点在网络中的重要程度
	聚类系数	$C_i = \frac{\sum a_{ij} a_{ik} a_{jk}}{\sum a_{ij} a_{jk}}, i \neq j \neq k$	(9)	网络结构中各个节点之间关系的紧密程度[12]	用于测度研究对象与其他学科之间知识融合的紧密程度
	特征中心性	$EC_i = \mathrm{c} \sum_{j=1}^{n} a_{ij} x_j$	(10)	网络节点中，节点与自身得分较高的节点相连接的紧密程度[13]	用于测度研究对象与其他学科之间知识融合的紧密程度
知识整合内容特征	主题术语跨学科性（TI）	$TI = d * \log tf$	(11)	TI 值越大，此主题术语将链接到的学科越多[14]	用于测度题术语跨学科程度
	主题分布广度	$div_{topic} = = \frac{1}{\lvert Z \rvert} \sum_{z_i \in Z} -p(z_i \lvert d) \log p(z_i \lvert d)$	(12)	其中 Z 是论文涉及到的主题集合，$p(z_i \lvert d)$表示文档 d 所分配的主题的概率分布	用于测度文献中抽取到的主题分布涉及的学术领域越广度
	主题跨学科性（TIc）	$TIc = n_t * (-\sum_{i=1}^{n} p_i \log p_i) * \sum_{i=1, j=1, i \neq j}^{i=n_c, j=n_c} d_{i,j} * \frac{2L}{N(N-1)}$	(13)	TIc 值越大则主题的学科交叉性越强[15]	用于测度研究主题的学科多样性和学科凝聚性

（3）知识整合内容特征。跨学科研究中的知识整合内容，可以是观点、概念和理论的整合，也可以是信息、数据和工具的整合。知识整合的多种形式主要体现在文献内容中[32]，从内部微观层面如句法学、语义学等方面探究内容语言层面的文本特征，揭示跨学科不同内容特征，研究文献的内在联系和科学结构，以及其对应的研究方向间的关系。本文利用 LDA 主题模型，对科研论文主题进行抽取，以主题分布广度、主题术语跨学科性（TI）[14]以及主题跨学科性（TIc）[15]作为指标，衡量科研论文的知识整合内容。

3.2　跨学科性自适应特征选择

在特征集构建的基础上，为选择对跨学科性影响较大的某些特征或特征组合，并且对不同测度指标之间的关联关系和重要程度进行分析，需要对特征进行特征适应度计算，并对结果进行选择评价。为解决跨

学科性测度数据中高相关性、非线性的数据问题，本文在初始化过程中，加入了特征适应度作为权重评估算法；在更新机制上，使用自适应参数选择策略替代原始的更新机制。

3.2.1　特征适应度计算

自适应特征选择的首先步骤是计算特征适应度并进行特征排序。由于跨学科性特征之间关联性较强且每个特征分别能测度跨学科性的不同特点，因此本文结合自适应的方法，对特征的分布、自动调整、特征的处理顺序以及特征相关权重的参数进行合理处理，按照规则对算法进行挑选并自动化调节系数，最终实现系统的稳定性和质量的可靠性。

（1）特征适应度。信息增益是衡量加入一个特征对区分数据样本分类的度量，通过计算加入某个特征后带来的信息熵的差值判断特征信息量，特征信息量越大，说明此特征越重要。同时，引入余弦相似度计算每一个特征与其他所有特征之间的相关度，因为本文希望筛选出的最优特征子集对分类高度相关，但彼此互不相关，故采用特征的信息增益与特征余弦相似度的比值[33]，设置了特征适应度，通过适应度可以去除冗余度高的特征，达到降低模型复杂度、提高分类性能的目的，如公式（14）~（16）所示。

$$IG(Y,x) = H(Y) - H(Y|x) \tag{14}$$

$$\cos(x,x') = \frac{\sum_{i=1}^{n}(x_i * x_i')}{\sqrt{\sum_{i=1}^{n}(x_i)^2} * \sqrt{\sum_{i=1}^{n}(x_i')^2}} \tag{15}$$

$$S_x = \frac{IG(Y,x)}{\cos x} \tag{16}$$

其中，$IG(X,Y)$是指加入特征 x 带来的信息增益，cos(*x*, *x*')指特征之间的余弦相似度，x_i 和 x_i' 分别代表特征 x 和特征 x' 对应每一个样本的值，S_x 是结合信息增益与余弦相似度的特征适应度，即特征 x 的信息增益与 1/ 余弦相似度坐标线与坐标轴所围矩形的面积越大[33]，则其特征适应度越高。

（2）特征排序。根据适应度计算的值对特征进行降序排序，构造以特征数为横坐标、以特征适应度为纵坐标的二维空间，结合空间坐标节点依次选择适应度结果高于其余特征适应度的特征构成特征子集用于特征选择，特征适应度如图 2 所示。

3.2.2　特征选择与评价

在跨学科性自适应特征选择阶段，将特征选择定义为：给定一个学习算法 L，一个数据集 D，数据集 D 来自一个具有 n 个特征 $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$的具有类别标记 y 的符合分布的例子空间，则一个最优特征子集是使得

1.INPUT：论文数据集 D，论文跨学科性相关特征集 X={X(1),X(2),...,X(n)}//n 代表跨学科性相关特征维数
2.OUTPUT：特征适应度值 Sx(1,2,…,n)和 X 排序
3.BEGIN
4.初始化数据集 D 中特征适应度值 w(1,2,…,n)=0;
5.for i=1 to n do:
6.获取特征 X(i);
7.根据公式（14）计算 X(i)信息增益;
8.for j=1 to n do:
9.根据公式（15）搜索计算 X(i)和 X'(i)余弦相似度;
10.根据公式（16）计算特征适应 Sx;
11.根据 Sx 对特征进行降序排序;
12.end for
13.end for

图 2　特征适应度算法

Fig.2　Feature fitness algorithm

DOI：10.13998/j.cnki.issn1002-1248.23-0214

某个评价准则 J=J(L,D)最优的特征子集。本文将论文的特征集和通过特征计算方法得到的特征数据集作为输入，其中，每篇论文的特征向量由原始特征集的特征指标构成，类别标识 Y 根据文献集本身的学科分布特征进行构建。通过机器学习分类器对分类准确度进行计算，同时结合最小化特征数量来进行特征评价目标函数构建，对搜索得到的特征子集进行评估，判断特征子集和特征组合接近全体特征的分辨能力，来进行自适应特征选择，如图 3 所示。

在构造特征评价目标函数时，首先构建基于特征子集的机器学习分类器，以分类器的性能评价所选特征子集的分类性能，进而评价提出自适应特征选择算法的性能[34]。为评价不同机器学习算法的性能，本文选用准确率（Accuracy）、精确率（Precision）、召回率（Recall）和 *F*1 值 4 个指标对结果进行评价。其中，准确率是指正确分类的文献数占文献总数的比值，计算公式（17）所示；精确率是指被分类为正的文献数量中实际为正的文献数量的概率，计算公式（18）所示；召回率是指实际为实际为证的文献数量中被预测为正的文献数量的概率，计算公式如（19）所示；*F*1 值综合 Precision 和 Recall 的结果，取值在[0,1]之间，计算公式如（20）所示。

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{17}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{18}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{19}$$

$$F1 = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \tag{20}$$

其中，公式（17）~（20）中，TP（True Positive）指预测是正确的正样本；FP（False Positive）指预测是错误的正样本；TN（True Negative）指预测是正确的负样本；FP（False Positive）指预测是错误的负样本。

```
1.INPUT:论文数据集 D,论文跨学科性相关特征集 X={X(1),X(2),....,X(n)}//n 代表跨学科性相关特征数量
2.OUTPUT:G=(X*,J(X))//X*代表筛选后的跨学科性特征子集,J(X)代表各个特征集合特征评价值
3.BEGIN
4.t=0;
5.E=∞;
6.d=|X|;
7.While t<T do
8.从特征集 X 中随机产生一个特征子集 X';
9.根据算法 1 得出特征适应度;
9.E'=CrossValidation(£(DX'));//使用机器学习器计算在特征子集 X'上的误差
10.使用公式（17）计算特征子集评价函数 J(S)
11.if(J'<J)∪((J'=J)? (d'<d))then
12.J=J′;
13.d=d′;
14.X*=X?
15.else
16.t=t+1;
17.end if
18.end while
```

图 3　特征选择

Fig.3　eature selection process

在利用机器学习计算分类准确度的基础上，为了最大化分类准确度和最小化特征数量[21]，采用本文采用特征子集评价函数 J(L,D)衡量特征子集效果，设置初始特征评价值 J 为一个极大值（便于后续特征子集的特征评价值取代它，从而找到特征评价值最小的特征子集），然后计算每一个特征子集的特征评价值 J'，如果小于 J，则把 J' 的值赋予 J。由于特征子集的特征数太少会导致数据集有效信息大量丢失，以至于分类准确率极低，可根据情况设置特征子集的最少特征数。特征评价函数如公式（21）所示。

$$J(L,D)=\alpha E'+(1-\alpha)\frac{i}{n} \tag{21}$$

其中，$J(S)$为含有 k 个特征变量子集 X* 的评价值，E'表示用机器学习在特征子集数据集上进行五折交叉验证的平均分类错误率，i 表示特征子集中的特征数，n 表示数据集预处理后的特征总数，α 和$(1-\alpha)$表示分类错误率和选择特征数量的权重。

3.3　论文跨学科性计算

通过自动化调节系数，识别可以纳入科研论文跨学科性测度的特征指标，最终得到了优于全体特征和其他特征子集跨学科性分辨能力的特征组合及其适应度的值。

将训练好的特征子集根据特征适应度进行加权求和计算该领域论文跨学科性，可以与学科属性、拓扑结构和文本内容特征指标遴选方法构建的原始特征集进行对比，选取该领域新的数据集，根据阈值筛选高跨学科性论文作为潜在跨学科性论文，以实现论文跨学科性的计算和测度，如图 4 所示。

4　实证研究

植物纳米生物技术作为新兴前沿交叉研究领域，它源自纳米材料技术与农业科学的深度融合，涵盖了作物纳米抗逆生物学（包括纳米材料种子引发技术）、纳米智能作物构建、作物纳米仿生学和作物纳米毒理学等，在农业领域具有重大的发展前景。因此，本文选择以植物纳米生物技术的论文作为跨学科性研究对象，验证上述所列指标体系和自适应特征选择的合理性。依据文献调研和专家咨询得到的研究主题构建检索策略，以 WOS 学科分类体系作为数据基础，对参考文献学科属性、文献引用网络拓扑结构和知识整合文本内容中的跨学科性相关指标进行学科层面和内容层面的分析，对跨学科性进行综合测度分析。

4.1　数据集

本文在选取数据集时主要考虑数据的代表性、适用性和易获得性，选取植物纳米生物技术领域，构建

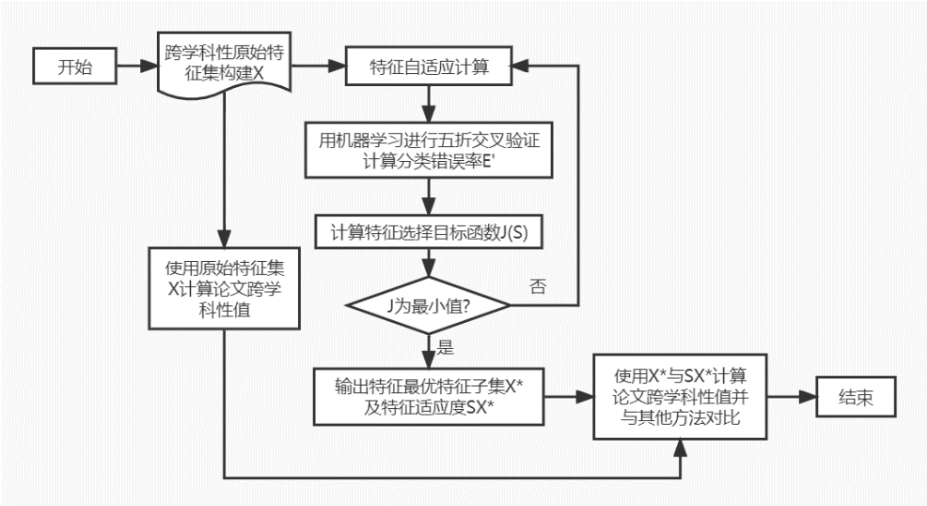


图 4　科研论文跨学科性计算流程图

Fig.4　Flow chart of interdisciplinarity calculation of scientific research papers

DOI：10.13998/j.cnki.issn1002-1248.23-0214

跨学科数据集。在限定检索领域后，本文选择科睿唯安的 Web of ScienceTM 核心合集作为科研论文检索数据源，通过文献调研和专家咨询构建检索式为：(TS=((nano* NOT (NANOG OR NANO2 OR NaNO3 OR NanoDrop OR Nanos2 OR Nanorana OR nanograms)) and ((pesticide* or insecticide* or fungicide* or herbicide*) or ("gene" or "genes" or "genetics" or "genomics" or "Synthetic biology" or "molecular biology" or "Gene sequencing" or "DNA sequencing" or "RNA interference" or "RNAi" or "multi-gene" or "intelligent plant sensor" or "DNA self-assembly technology") or（"growth regulator*" or " fertilizer*" or " plant nutrition" or "plant nutrient") or (film and environment*)) and （"crop" or "plant" or "wheat" or "rice" or "corn" or "cotton" or "soybean" or "rape" or "broomcorn" or "cabbage" or "watermelon" or "tomato" or "papaya" or "arabidopsis thaliana" or "tabacum" or "murphy" or "potato" or "crops" or "plants" or "wheats" or "rices" or "corns" or "cottons" or "soybeans" or "rapes" or "broomcorns" or "cabbages" or "watermelons" or "tomatoes" or "papayas" or "arabidopsis thalianas" or "tabacums" or "murphys" or "potatoes"))) AND (TI=(nano*) OR AB=(nano*) OR AK=(nano*))，语言为英语，文件类型包括 ARTICLE、PROCEEDINGSPAPER 和 REVIEW，限定时间在 2007 到 2021 年的数据，共检索得到 5 225 篇文献数据，经过数据集过滤清洗最终确定 5 206 篇文献数据。同时，为了测度该领域最新的论文跨学科性，研究选取了 2022 年的论文数据进行应用验证分析，采集时间为 2022 年 12 月 10 日，共检索得到论文 1 385 篇，经过数据集过滤清洗最终确定 1 370 篇文献数据。

4.2 实验过程

4.2.1 原始特征集构建

本文在 3.1 中从学科属性、拓扑结构和文本内容 3 方面选择了跨学科性 13 项相关特征并提出计算方法，此处对植物纳米生物技术的论文样本构建跨学科性原始特征集，对单个论文样本进行特征计算，并对数据进行标准化处理。本文共计算了 5 206 个样本的原始特征值，以论文样本中的一篇论文为例（Facile Method for the Selective Growth of Rice Like Rutile TiO2 from Peroxotitanate Gel and Its Photo-Activity）展示了各原始特征值的计算结果，如表 2 所示。

4.2.2 特征适应度计算与特征排序

根据期刊所属学科分类对单篇论文样本进行标注

表 2 跨学科性原始特征集计算示例

Table 2 Examples of interdisciplinarity calculation of original feature sets

评估指标	符号标记	特征值	预先处理方式
丰富度	X1	0.001 5	根据文献-期刊-所属学科对参考文献进行学科划分后根据公式（1）~（6）计算
均衡度	X2	0.877 8	
差异度	X3	0.999 0	
Rao-Stirling	X4	0.810 7	
True Diversity	X5	0.428 5	
Div*	X6	0.027 4	
凝聚性	X7	0.605 8	根据文献-所属学科构建学科网络后根据公公式（7）~（10）进行计算
中介中心性	X8	0.135 8	
聚类系数	X9	0.142 6	
特征中心性	X10	0.601 2	
主题术语跨学科性（TI）	X11	0.474 4	对标题和摘要使用 LDA 后根据公公式（11）~（13）计算
主题分布广度	X12	0.250 0	
主题跨学科性 TIc	X13	0.338 0	

得到实验数据集，选取前 50%的论文作为跨学科性高的论文，后 50%的论文作为跨学科性较低的论文[28]。本文将跨学科性高低用 y 表示，1 代表高跨学科性，0 代表低跨学科性。最终论文样本中包括高跨学科性文献 2 718 篇，低跨学科性论文 2 488 篇。

将 5 206 篇论文样本集跨学科性数据进行标准化处理，根据特征自适应度计算不同指标之间的余弦相似度，同时计算不同指标与 y 值之间的信息增益，运用信息增益和余弦相似度综合方法计算得出特征自适应度，得到非线性特征适应度矩阵，如图 5 所示。

对图 5 中结果进行分析，其中，x1~x13 列表示对应特征之间的余弦相似度，分析这些特征之间的相关性，可知 x2、x3、x4、x5、x7、x10、x12 之间相似度较高，x6 与 x1 之间相似度较高，在相关性分析的基础上，结合第 y 列中样本分类与某一特征之间的信息增益值分析特征与分类相互依赖程度，信息增益值越大说明此特征越重要，x2、x4、x5、x7、x8、x9、x10、x11 信息增益值较高。根据 3.2 中的适应度计算算法对 13 个原始指标的适应度进行计算并对特征进行排序，如图 6 所示。

4.3　方法评估

在特征适应度计算的基础上随机进行特征子集构建，将特征子集及其适应度的值作为新的输入，构建机器学习评估模型。本文利用分类器模型包括 RF、SVM、XGBoost 和 RNN 4 种分类器模型作为跨学科性评估模型。这里本文以原始特征集 -RF 模型构建为例来说明机器学习评估模型的构建（其他模型构建方法完全类似）。在构建 RF 模型过程中，需要对选择的参数进行调试，包括学习器的个数、树的最大深度、叶子节点最少样本数等，本研究对训练集论文样本采用 5 重交叉验证方法，并利用网格搜索（Grid-Search）优化算法寻找模型最优参数。本文 RF 算法程序使用的工具是 Python 中的 Random Forest Regressor，通过学习训练，确定好模型参数，评估模型也就构建完毕。使用精确度（Accuracy）、准确度（Precision）、$F1$ 值（F1-Score）和召回率（Recall）对分类准确率进行综合判断。

使用本文提出的基于自适应特征选择的方法在植物纳米生物技术论文特征数据集上对特征适应度求最

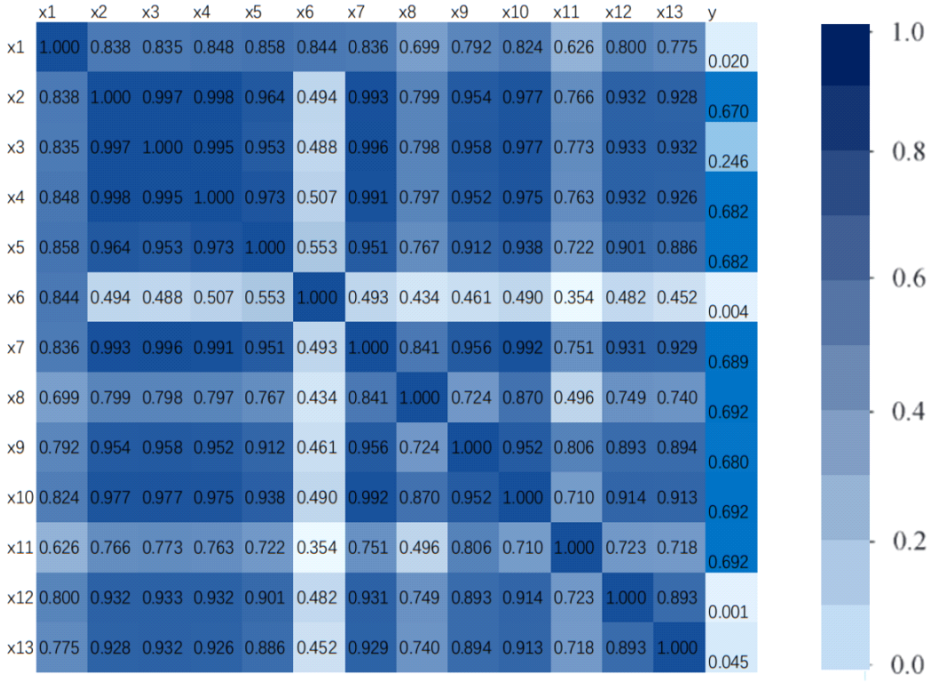


图 5　特征自适应度计算

Fig.5　Feature adaptive calculation

DOI：10.13998/j.cnki.issn1002-1248.23-0214

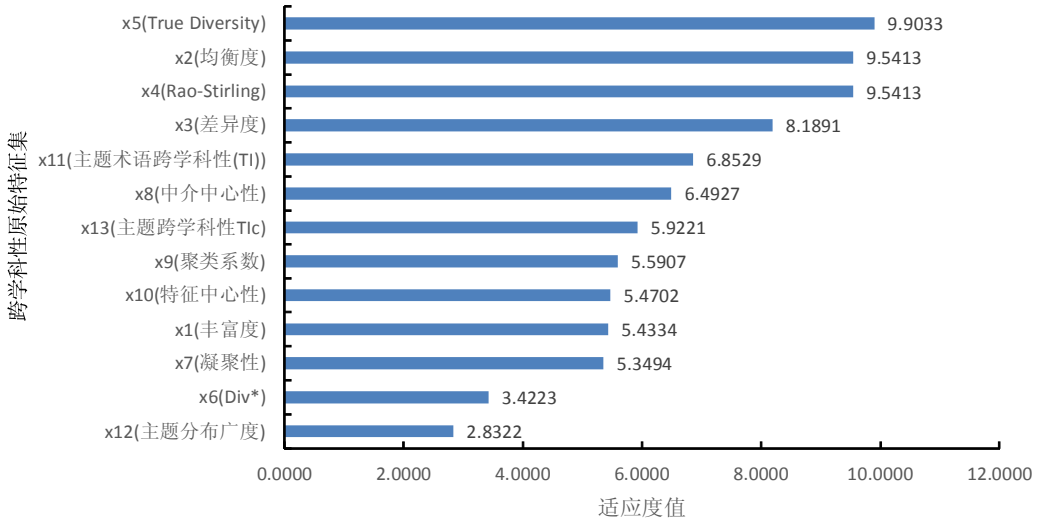


图 6　特征适应度计算

Fig.6　Feature fitness calculation

小值，其中，经过文献调研[33]并进行实验测试，选择 α 取值 0.5，最后求得最优特征子集为特征适应度总体较高的 8 个特征，具体如表 3 所示。

其中，学科属性特征中的丰富度指标被剔除，说明学科数量对跨学科性测度影响较小，均衡度和差异度对跨学科性测度作用较大[35,36]；拓扑特征整体效果较好，可能原因在于引文网络跨学科引用的结构因素对跨学科性测度影响较大，研究人员倾向于引用具有更高凝聚性、中介中心性和聚类系数较高的论文，语义特征中主题分布广度对跨学科性测度的作用较大，可能由于主题分布较广的论文一般学科分布较广，TI 和 TIc 对跨学科性测度影响较小被剔除，可能由于这两个指标与学科属性特征表示的信息相关度较高导致冗余。通过特征自适应计算结合机器学习、指标自适应特征选择方法，本文最终确定 8 个对植物纳米生物技术的跨学科性测量影响最大的指标。

将原始特征和自适应特征分别组配 4 种机器学习算法，对自适应跨学科性结果进行分析，结果如表 4

表 3　基于样本数据的自适应特征选择方法结果

Table 3　Results using the adaptive feature selection method based on sample data

特征选择方法	选择的变量	选择变量数量
自适应特征选择	x2、x4、x5、x7、x8、x9、x10、x11	8

表 4　机器学习模型性能指标结果对比

Table 4　Comparison of performance indicators of machine learning models

类别	指标	RF/%	SVM/%	XGBoost/%	RNN/%
原始	准确率	97.695	77.628	96.424	58.579
	精确率	97.668	85.178	96.410	60.420
	召回率	97.723	76.152	96.434	59.573
	*F*1	97.691	75.641	96.421	58.040
最优特征子集	准确率	97.997	95.343	96.552	62.228
	精确率	97.982	95.925	96.546	63.475
	召回率	98.005	95.048	96.554	62.956
	*F*1	97.993	95.230	96.550	62.040

所示。综合来看，对于 4 种机器学习方法，采用本研究自适应特征选择训练模型所构建的机器学习评估模型均效果表现突出，其中采用自适应特征和使用 RF 模型得到的跨学科性评估模型在所有模型中表现最优，最优模型的 F1 值为 97.993%，采用 XGBoost 构建的模型评价指标表现略差，但是差距不大，SVM 模型中自适应特征相比与原始特征集对模型效果有一定程度的提升。

通过 ROC 曲线进一步验证 RF、SVM、XGBoost 和 RNN 模型在特征筛选和构建跨学科性评估模型的性能，其中，使用所有特征构建的跨学科性评估模型 ROC 曲线如图 7（a）所示，使用自适应特征构建的跨学科性评估模型 ROC 曲线如图 7（b）所示。

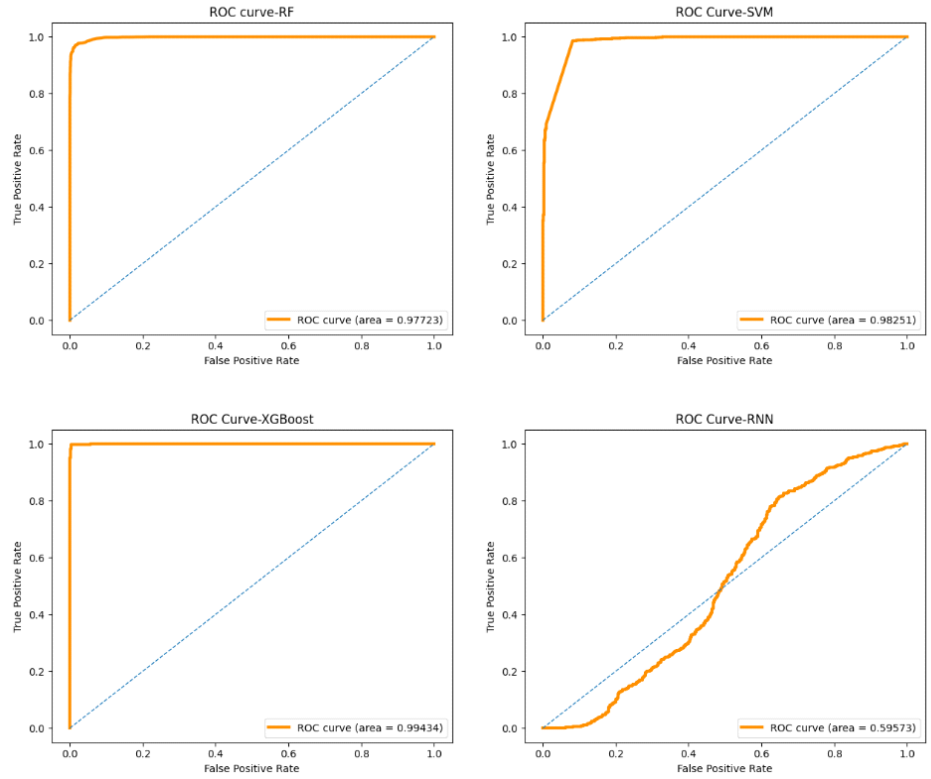
实证结果表明，本文经过自适应特征选择的结果通过了 4 种机器学习模型的检验（准确率较原始特征集均有提高），都说明了自适应特征对于提升模型效果的正向影响，证实了本文使用的自适应特征选择跨学科性测度方法可以作为判断论文跨学科性的有效参考，可在后续跨学科性评估、创新评估中被广泛应用。

4.4　结果分析

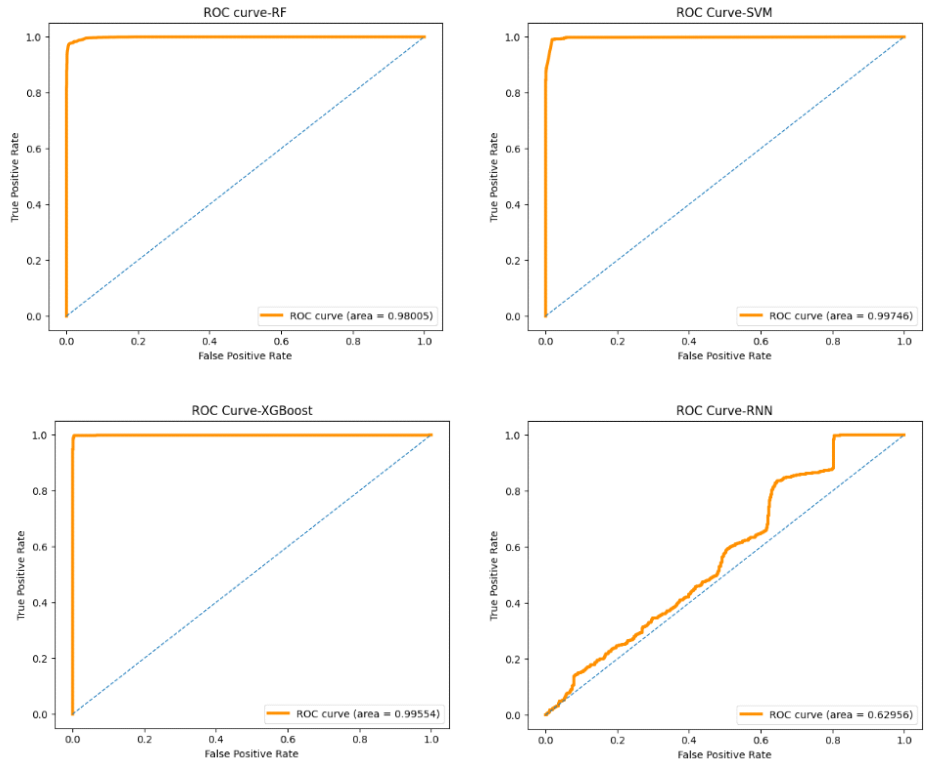
鉴于自适应特征 -RF 跨学科性机器学习模型评估效果较好，使用该方法遴选高跨学科性论文。对 2022 年植物纳米生物技术领域论文集（数据说明见 4.1 节）进行跨学科性计算，遴选高跨学科性论文。对验证数据集论文进行自适应跨学科性计算，按照测度值从大到小进行排序，自适应跨学科性数值越大，则该文献跨学科程度越高，表 5 主要对该领域部分论文跨学科性测度结果进行展示，同时对比原始特征计算的跨学科性结果，总体来看自适应跨学科性呈现区分度更高的结果。对验证数据集论文跨学科性数值的频率分布进行直方图展示如图 8，可以发现，自适应跨学科性大多集中于 0.4~0.7 之间。本文结合文献调研[37]和直方图结果将文献跨学科性分为 3 个水平：大部分文献的自适应跨学科性数值在 0.4~0.7 之间，属于中跨学科性文献；大于 0.7 的文献属于高跨学科性文献；小于 0.4 的文献属于低跨学科性文献。可以根据研究需要设定不同的阈值，筛选领域跨学科性文献用于进一步的学科交叉主题识别和内容分析等[37]。

对自适应特征筛选出的高跨学科性科研论文总体进行分析，共包括 229 篇高跨学科性论文，涉及 247 个 WOS 学科类别，学科数量累计叠加为 1 776 个，平均每篇论文涉及 7.75 个学科，表现出非常高的学科多样性。对 229 篇高跨学科性论文集合进行分析解读，主要内容包括了研究纳米颗粒在农业中的作用和潜在应用，从纳米颗粒在害虫控制和施肥中的使用，到它们对植物生长、植物病原体和环境的影响。部分研究涉及各种类型的纳米颗粒，如银、铜、氧化镁和氧化锌纳米颗粒对作物、害虫和微生物的作用，包括银纳米颗粒可以抑制烟草花叶病毒在植物中的生长、银纳米颗粒可以诱导水稻植株产生抗性并导致水稻主要害虫褐飞虱死亡、从薄荷和嗜线虫杆菌中生物合成银纳米颗粒对晚疫病和斜纹夜蛾幼虫具有控制效果、包覆的铜掺杂 ZnO 和铜纳米颗粒在控制植物病原真菌和线虫方面显示出良好的效果；此外，还有部分研究讨论了纳米颗粒对植物生长和生理的影响，包括泡桐纳米肥在不同灌溉水平下对罗勒形态生理性状和干物质产量的不同影响、壳聚糖 / 碳纳米颗粒和壳聚糖 / 木质纤维素纳米纤维复合材料是泥炭田水稻种子的有效生长介质、掺锌和掺镁羟基磷灰石纳米颗粒经尿素改性后也被开发为智能氮肥等；还有研究探讨了纳米颗粒对环境的影响，包括其对有益昆虫和微生物的毒性、银纳米颗粒对真菌毒素产生真菌镰刀菌（Fusarium Graminearum）和卵菌病原体疫霉（Phytophthora Infestans）的毒性和作用机制、银和氧化锌纳米颗粒对主要农业害虫棉铃虫（Helicoverpa Armigera）的生物学和生命表参数的毒性等。这些论文全面概述了纳米颗粒在可持续农业中使用的研究现状，强调了它们的潜在好处和潜在风险。虽然纳米颗粒在控制害虫和促进植物生长方面显示出前景，但需要进一步研究它们对环境和非目标生物的影响，以确保它们的安全和可持续使用。总的来说，这些论文涉及范围较广，且结合了很多领域的知识和研究方法，表现出学科交叉的态势，可为相关研究人员提供一定的参考。本研究提出的基

DOI：10.13998/j.cnki.issn1002-1248.23-0214



（a）使用所有特征



（b）使用自适应特征选择特征

图 7　跨学科性机器学习评估 ROC 曲线对比

Fig.7　Comparison of ROC curves of interdisciplinary in machine learning evaluation

表 5　科研论文跨学科性测度结果（部分）

Table 5　Measurement results of interdisciplinarity of research papers (partial)

ID	科研论文标题	所属 WOS 中学科分类	自适应特征跨学科性计算	原始特征跨学科性计算
n396	*Nanocidal Effect of Rice Husk-Based Silver Nanoparticles on Antioxidant Enzymes of Aphid*	Biochemistry & Molecular Biology; Endocrinology & Metabolism	0.999 9	0.920 1
n15	*Potential of Chitosan/Carbon Nanoparticles and Chitosan/Lignocellulose Nanofiber Composite as Growth Media for Peatland Paddy Seeds*	Environmental Sciences; Public, Environmental & Occupational Health	0.967 1	0.916 5
n1088	*Simple and Efficient Enzymatic Procedure for P-Coumaric Acid Synthesis: Complete Bioconversion and Biocatalyst Recycling under Alkaline Condition*	Biotechnology & Applied Microbiology; Engineering, Chemical	0.920 7	0.915 9
n68	*Toxicity Effects and Mechanisms of MgO Nanoparticles on the Oomycete Pathogen Phytophthora Infestans and Its Host Solanum Tuberosum*	Environmental Sciences; Toxicology	0.876 9	0.912 8
n419	*An Overview of the Role of Nanoparticles in Sustainable Agriculture*	Biotechnology & Applied Microbiology	0.876 2	0.902 5
n813	*Biosynthesis and Characterization of Iron Oxide Nanoparticles from Mentha Spicata and Screening Its Combating Potential Against Phytophthora Infestans*	Plant Sciences	0.867 6	0.903 5
n106	*Silver and Copper-Oxide Nanoparticles Prepared with GA3 Induced Defense in Rice Plants and Caused Mortalities to the Brown Planthopper, Nilaparvata Lugens (Stal)*	Environmental Sciences; Nanoscience & Nanotechnology	0.865 4	0.901 1
n760	*Nanopesticides: Current status and Scope for Their Application in Agriculture*	Agronomy; Plant Sciences	0.864 5	0.895 4
n691	*The Effect of Nano-Fertilizer of Paulownia on Morpho-Physiological Traits and Dry Matter Yield of Basil Under Different Irrigation Levels*	Plant Sciences	0.864 0	0.899 1
n890	*Toxicity and Action Mechanisms of Silver Nanoparticles Against the Mycotoxin-Producing Fungus Fusarium Graminearum*	Multidisciplinary Sciences	0.859 2	0.914 3
n809	*Coated Cu-Doped ZnO and Cu Nanoparticles as Control Agents Against Plant Pathogenic Fungi and Nematodes*	Environmental Sciences; Nanoscience & Nanotechnology	0.857 1	0.898 7
n616	*Biosynthesized Silver Nanoparticles Inhibit Pseudomonas Syringae pv. Tabaci by Directly Destroying Bacteria and Inducing Plant Resistance in Nicotiana Benthamiana*	Plant Sciences	0.850 4	0.908 0
n613	*Controlled and Prolonged Release Systems of Urea from Micro- and Nanomaterials as an Alternative for Developing a Sustainable Agriculture: A Review*	Nanoscience & Nanotechnology; Materials Science, Multidisciplinary	0.838 4	0.903 3
…	…	…	…	…
n1153	*Intelligent Modeling of Unconfined Compressive Strength (UCS) of Hybrid Cement-Modified Unsaturated Soil with Nanostructured Quarry Fines Inclusion*	Engineering, Civil	0.000 1	0.246 5

DOI：10.13998/j.cnki.issn1002-1248.23-0214

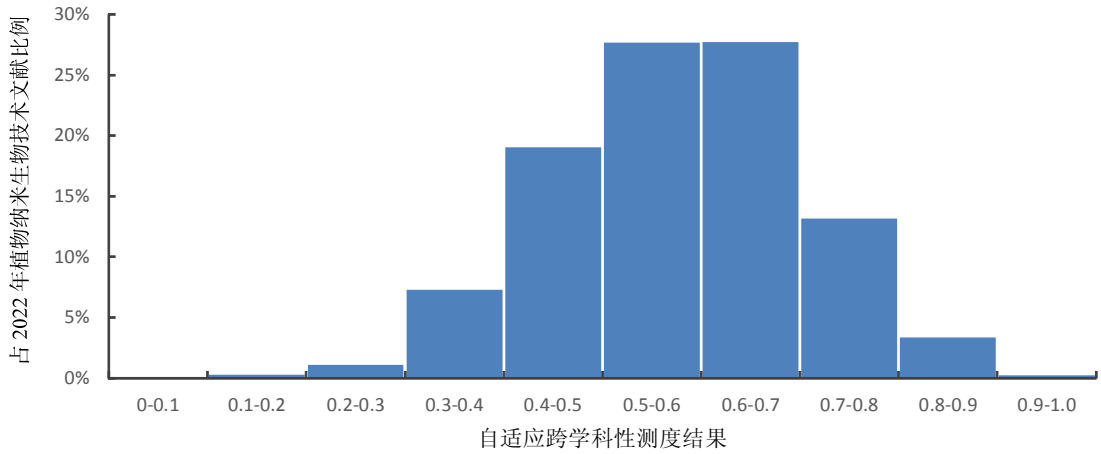


图 8　2022 年植物纳米生物技术文献基于自适应特征选择跨学科性测度频率直方图

Fig.8 Frequency histogram of interdisciplinary measures selected by plant nanobiology literature in 2022 based on adaptive features

于自适应特征选择的论文跨学科性测度方法与原始特征方法相比，自适应地综合了不同维度的跨学科性测度结果和测度角度，避免了常用的跨学科性测度中学科属性特征计算结果区分度较低，拓扑结构计算结果可能存在引用滞后等问题，同时结合语义特征增大了论文测度结果的区分度，总体而言，自适应机器学习方法能结合各个维度的不同指标自适应计算跨学科性数值，在实际操作中将自适应特征选择方法迁移到不同的学科领域，根据不同学科的数据特点、学科属性、拓扑结构和文本内容特征自适应进行特征选择，同时可以按照阈值自定义单个领域 Top 高跨学科性论文数量，选择合适的论文数量，降低人工判读的工作量，提高跨学科论文筛选的效率。

5 结　语

随着跨学科性测度任务越来越繁重，掌握跨学科研究的特点、发展规律和测度方式，对理解知识融合扩散情况与学科布局管理决策至关重要。本文从跨学科性丰富的内涵出发，从学科属性、拓扑结构和文本内容特征 3 个维度提取跨学科性相关关键技术指标，从不同的维度构建跨学科性原始特征集，对特征集中输入指标和数据进行信息增益和特征相似性分析，提出一种基于特征自适应选择的论文跨学科性测度方法，利用机器学习分类器的方法最大化特征分类准确率，同时结合最小化特征数量自适应选择能最优表达跨学科性的特征子集，将选择出的自适应特征集用于论文跨学科性计算中，并对比原始特征集跨学科性计算结果进行综合分析，结果表明：①本研究提出的结合学科属性、知识网络拓扑结构和知识整合文本内容特征构建的特征集经过自适应计算展示出良好的效果，可以结合不同维度实现对跨学科性的度量。其中，在学科属性特征中，均衡度和差异度对跨学科性评估作用较大；知识网络拓扑结构特征整体效果较好，知识整合文本内容特征中主题分布广度对跨学科性评估的作用较大，各特征根据适应度加权求和进一步提升了计算效果。②利用自适应特征选择能够提升方法的准确率，自适应特征在 4 种机器学习分类器中均表现出较高的准确率、精确率和 F1 值，其中，自适应特征结合 RF 的方法，在 4 种分类器种显示出最高的分类准确率。③本研究提出的基于自适应特征选择的跨学科性测度方法可以拓展应用于各个学科领域，从而实现跨学科性智能化测度。在原始特征集中综合了传统测度方法中对跨学科性不同维度的研究，并在自适应特征选择的方法中实现对各个指标的有效验证，基于这种自适应的跨学科性测度方法，可以用于比较文献或文献集合的跨学科性程度，进一步对比机构之间、期刊之间或不同领域之间文献的跨学科性，有利于了解机构、期刊和学科领域中知识的发展和变化，更科学地对学科进行规划。

本研究基于提出的基于自适应特征选择的论文跨学科性测度方法实现了对论文自适应跨学科性的测度与分析，但仍有不足之处。本研究从跨学科性的内涵出发，仅从学科属性、网络拓扑和文本内容特征 3 个维度对跨学科性进行了分析，对跨学科研究的产生机制和发展动力等并未涉及；同时，本文在对跨学科性自适应特征选择验证时，选择了数据库 Web of ScienceTM 进行数据采集和学科类别划分，这种学科分类也存在一些问题，难以迁移到其他数据库或文档内容的分类中，也难以反映知识的动态变化情况，未来的研究需要更准确、科学地对文献进行分类，后续将在这些方面进行更加深入的研究。

参考文献：

[1] GLÄNZEL W, DEBACKERE K. Various aspects of interdisciplinarity in research and how to quantify and measure those[J]. Scientometrics, 2022, 127(9): 5551–5569.

[2] 曾粤亮, 李玉海. 基于生态系统理论的跨学科科研合作运行框架与关键问题[J]. 情报资料工作, 2022, 43(3): 34–42.
ZENG Y L, LI Y H. Operational framework and key issues of interdisciplinary scientific research cooperation based on ecological systems theory[J]. Information and documentation services, 2022, 43(3): 34–42.

[3] European commission. Directorate general for research and innovation., research, innovation, and science policy experts (rise)[M/OL]. Quests for interdisciplinarity: A challenge for the ERA and HORIZON 2020, LU: Publications Office, 2015. https://data.europa.eu/doi/10.2777/499518.

[4] 樊春良, 樊天. 国外学科交叉研究的发展趋势及启示[J]. 中国科学基金, 2019, 33(5): 446–452.
FAN C L, FAN T. The trends of development interdisciplinary research abroad and its inspiration[J]. Bulletin of national natural science foundation of China, 2019, 33(5): 446–452.

[5] 中华人民共和国科学技术进步法 _ 中国人大网[EB/OL]. (2021–12–24)[2022–08–01].http://www.npc.gov.cn/npc/c30834/202112/1f4abe22e8ba49198acdf239889f822c.shtml.

[6] 步一, 陈洪侃, 许家伟, 等. 跨学科研究的范式解析: 理解情报学术中的"范式"[J]. 情报理论与实践, 2022, 45(3): 12–18, 34.
BU Y, CHEN H K, XU J W, et al. Connotations of interdisciplinarity from the perspective of paradigms: Towards "paradigms" in information science research and practices[J]. Information studies: Theory & application, 2022, 45(3): 12–18, 34.

[7] STIRLING A. A general framework for analysing diversity in science, technology and society[J]. Journal of the royal society interface, 2007, 4(15): 707–719.

[8] PORTER A L, RAFOLS I. Is science becoming more interdisciplinary? Measuring and mapping six research fields over time[J]. Scientometrics, 2009, 81(3): 719–745.

[9] ZHANG L, ROUSSEAU R, GL?NZEL W. Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account[J]. Journal of the association for information science and technology, 2016, 67(5): 1257–1265.

[10] LEYDESDORFF L, WAGNER C S, BORNMANN L. Interdisciplinarity as diversity in citation patterns among journals: Rao–Stirling diversity, relative variety, and the Gini coefficient[J]. Journal of informetrics, 2019, 13(1): 255–269.

[11] RAFOLS I, MEYER M. Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience[J]. Scientometrics, 2010, 82(2): 263–287.

[12] RAFOLS I. Knowledge integration and diffusion: Measures and mapping of diversity and coherence[M]//DING Y, ROUSSEAU R, WOLFRAM D. Measuring scholarly impact. Cham: Springer, 2014: 169–190.

[13] LEYDESDORFF L, WOUTERS P, BORNMANN L. Professional and citizen bibliometrics: Complementarities and ambivalences in the development and use of indicators – A state–of–the–art report[J]. Scientometrics, 2016, 109(3): 2129–2150.

[14] XU H Y, GUO T, YUE Z H, et al. Interdisciplinary topics of information science: A study based on the terms interdisciplinarity index series[J]. Scientometrics, 2016, 106(2): 583–601.

[15] 黄菡, 王晓光, 王依蒙. 复杂网络视角下的研究主题学科交叉测度研究[J]. 图书情报工作, 2022, 66(19): 99–109.
HUANG H, WANG X G, WANG Y M. Research on the interdisciplinary measurement of research topics from the perspective of

DOI：10.13998/j.cnki.issn1002-1248.23-0214

complex networks[J]. Library and information service, 2022, 66(19): 99–109.

[16] 姚旭, 王晓丹, 张玉玺, 等. 特征选择方法综述[J]. 控制与决策, 2012, 27(2): 161–166, 192.
YAO X, WANG X D, ZHANG Y X, et al. Summary of feature selection algorithms[J]. Control and decision, 2012, 27(2): 161–166, 192.

[17] CHEN M X, CHU X Q, SUBBALAKSHMI K P. MMCoVaR: Multimodal COVID–19 vaccine focused data repository for fake news detection and a baseline architecture for classification[C]//Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. New York: ACM, 2021: 31–38.

[18] VAN VLASSELAER V, BRAVO C, CAELEN O, et al. APATE: A novel approach for automated credit card transaction fraud detection using network–based extensions[J]. Decision support systems, 2015, 75: 38–48.

[19] 熊志斌. 信用评估中的特征选择方法研究[J]. 数量经济技术经济研究, 2016, 33(1): 142–155.
XIONG Z B. Research on feature selection method in credit evaluation[J]. The journal of quantitative & technical economics, 2016, 33(1): 142–155.

[20] DAHIYA S, HANDA S S, SINGH N P. A feature selection enabled hybrid–bagging algorithm for credit risk evaluation[J]. Expert Systems, 2017, 34(6): e12217.

[21] 赵蕴华, 张静, 李岩, 等. 基于机器学习的专利价值评估方法研究[J]. 情报科学, 2013, 31(12): 15–18.
ZHAO Y H, ZHANG J, LI Y, et al. Study on evaluation for patent value based on machine learning[J]. Information science, 2013, 31(12): 15–18.

[22] 何向, 李莉, 王小绪. 基于机器学习的高校专利价值评估体系构建[J]. 情报工程, 2020, 6(1): 50–58.
HE X, LI L, WANG X X. The construction of assessing college patent value system based on machine learning[J]. Technology intelligence engineering, 2020, 6(1): 50–58.

[23] 李欣, 范明姐, 黄鲁成. 基于机器学习的专利质量评价研究[J]. 科技进步与对策, 2020, 37(24): 116–124.
LI X, FAN M J, HUANG L C. Research on patent quality evaluation using machine learning[J]. Science & technology progress and policy, 2020, 37(24): 116–124.

[24] 李欣, 温阳, 黄鲁成, 等. 一种基于机器学习的研究前沿识别方法研究[J]. 科研管理, 2021, 42(1): 20–32.
LI X, WEN Y, HUANG L C, et al. A study of the research front identification method based on machine learning[J]. Science research management, 2021, 42(1): 20–32.

[25] 钱玲飞, 贺婉莹, 杨建林. 论文学术创新力特征指标体系研究[J]. 情报科学, 2021, 39(1): 56–64.
QIAN L F, HE W Y, YANG J L. The characteristic index system of academic innovation ability[J]. Information science, 2021, 39(1): 56–64.

[26] 李道全, 李腾, 李玉秀. 基于自适应特征选择与 KNN 的网络流量分类研究[J/OL]. 计算机工程与应用: 1–9[2023–05–08]. http://kns.cnki.net/kcms/detail/11.2127.TP.20220510.1353.002.html.
LI D Q, LI T, LI Y X. Research on network traffic classification based on adaptive feature selection and KNN [J/OL]. Computer Engineering and Applications: 1–9 [2023–05–08]. http://kns.cnki.net/kcms/detail/11.2127.TP.20220510.1353.002.html.

[27] SHAFIQ M, YU X Z, BASHIR A K, et al. A machine learning approach for feature selection traffic classification using security analysis[J]. The journal of supercomputing, 2018, 74(10): 4867–4892.

[28] 刘凯. 随机森林自适应特征选择和参数优化算法研究[D]. 长春: 长春工业大学, 2018.
LIU K. Research on adaptive feature selection and parameter optimization algorithm for random forest[D]. Changchun: Changchun University of Technology, 2018.

[29] National academy of sciences, national academy of engineering, institute of medicine[M]//Facilitating interdisciplinary research. Washington, D.C.: The National Academies Press, 2005.

[30] 黄颖, 张琳, 孙蓓蓓, 等. 跨学科的三维测度——外部知识融合、内在知识会聚与科学合作模式[J]. 科学学研究, 2019, 37(1): 25–35.
HUANG Y, ZHANG L, SUN B B, et al. Interdisciplinarity measurement: External knowledge integration, internal information convergence and research activity pattern[J]. Studies in science of sci-

ence, 2019, 37(1): 25–35.

[31] ZENG B, LYU H H, ZHAO Z Y, et al. Exploring the direction and diversity of interdisciplinary knowledge diffusion: A case study of professor Zeyuan Liu´s scientific publications[J]. Scientometrics, 2021, 126(7): 6253–6272.

[32] 张琳, 刘冬东, 吕琦, 等. 论文学科交叉测度研究: 从全部引文到章节引文[J]. 情报学报, 2020, 39(5): 492–499.
ZHANG L, LIU D D, LYU Q, et al. Interdisciplinarity measurement in publications: From full reference analysis to sectional reference analysis[J]. Journal of the China society for scientific and technical information, 2020, 39(5): 492–499.

[33] 谢娟英, 吴肇中, 郑清泉. 基于信息增益与皮尔森相关系数的 2D 自适应特征选择算法[J]. 陕西师范大学学报(自然科学版), 2020, 48(6): 69–81.
XIE J Y, WU Z Z, ZHENG Q Q. An adaptive 2D feature selection algorithm based on information gain and Pearson correlation coefficient[J]. Journal of Shaanxi normal university (natural science edition), 2020, 48(6): 69–81.

[34] CHEN R C, CARAKA R E, PILIANG A, et al. An end to end of scalable tree boosting system[J]. Sylwan, 2020, 164(5): 140–151.

[35] 谭慧颖, 耿骞, 靳健. 一种基于重叠社区标签传播的学科划分方法[J]. 农业图书情报学报, 2021, 33(1): 41–52.
TI H Y, GENG Q, JIN J. A COPRA based algorithm for subject division[J]. Journal of library and information science in agriculture, 2021, 33(1): 41–52.

[36] 张宝隆, 王昊, 张卫. 学科交叉视角下的学科区分能力测度方法及分析研究[J]. 情报学报, 2022, 41(4): 375–387.
ZHANG B L, WANG H, ZHANG W. Measurement and analysis of disciplinary discriminative capacity from an interdisciplinary perspective[J]. Journal of the China society for scientific and technical information, 2022, 41(4): 375–387.

[37] 韩正琪, 刘小平, 寇晶晶. 基于 Rao–stirling 指数和 LDA 模型的领域学科交叉主题识别——以纳米科技为例[J]. 情报科学, 2020, 38(2): 116–124.
HAN Z Q, LIU X P, KOU J J. Interdisciplinary literature discovery based on Rao–stirling diversity indices: Case studies in nanoscience and nanotechnology[J]. Information science, 2020, 38(2): 116–124.

Interdisciplinarity Measurement Method of Scientific Research Papers based on Adaptive Feature Selection

WANG Jinfei[1], SUN Wei[1,2*], ZHANG Xuefu[1,2], YANG Lu[1]

(1. Institute of Agricultural Information, Chinese Academy of Agricultural Sciences, Beijing 100081;
2. Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081)

Abstract: [Purpose/Significance] Interdisciplinary research can creatively solve complex problems in natural environment and human society through knowledge integration and penetration. With the increase of interdisciplinary research results, the evaluation of interdisciplinarity becomes increasingly necessary. How to establish an effective method for interdisciplinarity measurement and achieve a comprehensive measurement of scientific research papers is an urgent problem to be solved. [Method/Process] Based on the above background, this study takes the data of scientific research papers as the analysis source, deconstructs the interdisciplinarity of scientific

DOI：10.13998/j.cnki.issn1002-1248.23-0214

research papers from multiple dimensions, constructs the feature set of interdisciplinarity of scientific research papers, and on this basis proposes the method for measuring interdisciplinarity based on the adaptive method of machine learning, and conducts a comprehensive measurement of interdisciplinarity. This study has certain positive significance for researchers to understand the interdisciplinary papers in the field. The work process is as follows: First of all, the basic concepts of interdisciplinarity are sorted out and related concepts are discriminated, and the index of interdisciplinarity of different dimensions is analyzed. Based on the connotation and characteristics of interdisciplinary research, the characteristic index of interdisciplinarity of scientific research papers is extracted from three dimensions: subject attribute, knowledge network topology and knowledge integration text content. Secondly, an interdisciplinarity measurement method based on machine learning is constructed. By analyzing information gain and feature similarity of input indexes and data in feature sets, a feature selection calculation method based on adaptive feature selection is proposed, and the accuracy of feature classification is maximized by machine learning classifier. At the same time, the feature subset that can best express the interdisciplinary is selected based on the adaptive selection of the minimum number of features, and the selected adaptive feature set is used in the calculation of the interdisciplinary of the paper, and the results of the calculation of the original feature set are analyzed comprehensively. Finally, an empirical study was carried out in the field of plant nanobiotechnology to verify the effectiveness of the index system and adaptive feature selection listed above, identify and screen papers with high interdisciplinarity in the field, measure the interdisciplinarity of papers and identify key influencing factors based on the calculation of subject attributes, knowledge network topology and knowledge integration text content features. [Results/Conclusions] The main empirical results show that, among the subject attributes, the balance degree and the difference degree have a greater effect on the interdisciplinary evaluation. The overall effect of knowledge network topology structure features is satisfactory, the distribution breadth of knowledge integration text content features has a greater effect on interdisciplinary evaluation, and the calculation effect is further improved by fitness weighted summation of each feature. The results demonstrate that the adaptive feature selection proposed in this paper can effectively screen the interdisciplinary related feature indexes, improve the reliability of the results, and achieve a comprehensive and in-depth measurement of the interdisciplinary of scientific research papers. This measure method avoids the subjective defects that may occur in qualitative evaluation and the problems that different measure indicators may produce contradictory results. It provides a new idea and direction for interdisciplinary measurement.

Keywords: interdisciplinarity; adaptive feature selection; paper measurement