融入技术匹配的高校可转化专利识别研究*

俞琰[1,2]　董优优[2]　刘攀[2]　杨舸[1]
（1.南京工业大学图书馆，南京 211816；2.南京工业大学经济与管理学院，南京 211816）

摘要：高校可转化专利识别是提升高校科技成果产业化水平的重要课题。然而，面对大量高异质、技术话语密集的专利数据，如何从中有效识别出具备转化潜质的专利，仍缺乏成熟、可操作的评价工具和筛选机制。目前相关研究多基于机器学习挖掘高校专利本身特征，侧重于评估高校专利的内在价值，忽略了高校专利与受让企业之间的技术供需适配性。为此，本研究提出一种融入技术匹配的高校可转化专利识别方法：首先构建融入技术匹配的可转化专利识别特征，其次构建应对稀疏专利转化场景的高校专利转化困难负采样，再次构建多算法优势的高校可转化专利加权集成识别模型，最后基于博弈归因分析高校可转化专利驱动因素分析。本研究选取膜技术领域开展高校可转化专利识别实验，结果表明该识别方法相较于现有仅关注内在价值的识别方法，识别准确性显著提升，展现出更好的工业实际应用价值。

关键词：可转化专利；高校；企业；技术匹配；技术互补；技术相似
中图分类号：G25　DOI：10.3772/j.issn.1673–2286.2026.04.008
引文格式：俞琰，董优优，刘攀，等. 融入技术匹配的高校可转化专利识别研究[J]. 数字图书馆论坛，2026，22（4）：78–88.

高校作为国家创新体系中的核心主体之一，其专利转化对驱动经济发展至关重要。近年来，我国高校专利申请量呈快速增长趋势，但中国的科技成果转化率却不足30%[1]，大量专利处于沉睡状态，高产出、低转化的结构性矛盾凸显。如何从有大量存量的高校专利中精准识别出具备转化潜质的专利是提升科技成果转化的重要课题，但从大量高异质、技术话语密集的专利数据中筛选出可转化专利仍处于方法探索阶段，缺乏成熟、可操作的评价工具和筛选机制。因此，本研究聚焦于高校可转化专利的识别问题，旨在构建一套有效的识别方法，为高校提供更具操作性的筛选路径，从而支撑高校存量专利的盘活与转化，助力企业技术创新。

1　相关研究

可转化专利识别是指从高校持有的大量存量专利中精准识别出具备向产业端转移并实现商业化应用潜质的专利，从而为科技成果转化效率的提升提供决策支撑。围绕这一目标，既有文献在表述上使用可转移专利、可转让专利、可转化专利等不同术语，但其核心指向并无实质差异，均聚焦于专利从高校向产业端流动与应用这一根本问题，在分析对象和研究目标上具有内在一致性。

识别特征构建是识别专利转化潜力的基础，现有研究主要从技术、经济和法律3个维度展开。在技术维度，研究者通过发明人数、引用数及IPC小类数等指标刻画专利的创新质量与协作深度[2–4]，同时运用SBERT-LDA混合模型、LDA与K-means聚类或主题建模等语义解析技术有效捕获专利的技术新颖性与领域特征[5–7]。在经济维度，同族专利数是核心考量指标，其被证实与专利转化率或转化价值具有显著的正相关关系[8–9]，能直观反映专利在全球市场布

收稿日期：2026–03–03
*本研究得到国家社会科学基金一般项目“数据驱动的高校技术转移供需信息挖掘模式构建研究”（编号：23BTQ098）资助。

局中的商业化潜力。在法律维度，权利要求数和专利维持期是常用表征指标[10−11]，其中独立权利要求长度和从属权利要求数量被视为评估法律强度的有效指标，专利维持期则反映权利人对专利价值的持续认可与维护意愿，两者共同表征专利的保护范围与权利稳定性。

综合来看，上述技术、经济、法律三维特征体系已较为系统地覆盖了专利内在价值的主要方面，也是目前专利价值评估领域的主流分析框架。然而，该体系本质上是对专利静态属性的刻画，较少涉及专利与企业需求之间的动态匹配关系，尤其是专利技术与潜在受让企业技术能力、业务方向之间的适配程度。换言之，一项专利即便在技术、经济、法律维度均表现出较高价值，若与企业的技术需求不匹配，仍难以实现有效转化。

可转化专利的识别方法主要为机器学习方法。该方法因具备处理高维特征与非线性关系的优势，已成为该领域的主流建模工具。在模型构建的演进路径上，研究者首先着力于特征空间的优化，通过指标约减算法剔除冗余信息[12]，以提升模型输入质量；在此基础上，分别采用神经网络[1,13]和随机森林[2]等典型算法构建识别模型。面对专利文本语义复杂、信息密度高的挑战，研究者进一步引入深度神经网络[10]以增强语义表征能力，或构建融合文本特征与外部多源信息的综合模型[6]以实现更全面的知识表达。针对高校专利场景的多项研究[4−5,13]表明，深度学习及主流分类算法在提升转化识别准确性方面表现出显著优势。

综上所述，现有高校可转化专利研究主要基于专利本身的价值特征，识别其是否具备转化潜力。然而，高校专利技术能否实现向企业的有效转化，不仅关注技术本身的价值，更需要考量该技术是否契合企业需求，以填补研发空白或构筑技术壁垒。鉴于此，区别于现有研究主要依据专利价值特征进行识别，本文引入校企专利的技术匹配特征，将其纳入高校可转化专利的识别框架，以期更准确地识别具备转化潜力的高校专利，为高校挖掘高价值可转化专利、实现精准科技成果转化提供科学支撑。

2　基于技术匹配的高校可转化专利识别方法

本研究构建的高校可转化专利识别模型包括特征构建、样本优化、模型集成、归因分析4个核心模块。①特征构建：旨在构建一套能够有效表征专利转化潜力的特征体系，作为识别模型的基础与前提。②样本优化：为应对稀疏专利转化场景构建高校专利转化困难负样本，旨在为识别模型提供更具判别边界的高质量训练数据集。③模型集成：构建融合多算法优势的高校可转化专利加权集成识别模型，作为核心的机器学习模型。④归因分析：对高校可转化专利驱动因素进行博弈归因分析，旨在实现对识别模型结果的因果解释与核心驱动因素解析。上述4个模块共同构成了完整技术链条，其中特征构建与样本优化共同服务于模型训练，模型输出结果则由归因分析进行解释。

2.1　特征构建

2.1.1　技术匹配特征

技术匹配是指高校与企业在技术层面的关联适配程度，直接反映了双方技术协同与资源整合的潜在可能性。为量化这一匹配程度，本研究从技术相似度和技术互补度两个维度对技术匹配进行量化分析。

2.1.1.1　技术相似度特征计算

技术相似度反映了高校专利技术与企业技术吸收能力的匹配程度。这种共性不仅体现了双方在技术领域、功能路径及应用场景上的重叠，更是降低技术转移门槛、加速技术内部化过程的关键因素。本研究构建了分类与语义双维测度体系，分别从国际专利分类号（International Patent Code，IPC）结构化特征与文本非结构化特征两个视角评估技术相似度。

（1）IPC技术相似度。IPC分类体系提供了技术知识的层次化表征。利用IPC小组这一细粒度分类特征，可以判断高校专利与企业已有技术布局的重合程度。IPC技术相似度可用于衡量高校专利特征在企业技术库中的覆盖率，计算公式如式（1）所示。该指标值越大，表明该专利技术分类与企业已有技术布局的契合度越高。

$$\mathrm{Sim}_{\mathrm{IPC}}\left(p,E\right)=\frac{\left|S_p\cap S_E\right|}{\left|S_p\right|} \tag{1}$$

其中：S_p为高校专利p的IPC小组集合，S_E为企业E的全量专利IPC小组集合。

（2）文本技术相似度。为捕捉更深层的语义关联，本研究通过计算专利文本向量的余弦距离来量化内容相似度。将高校专利p与企业E存量专利库中的专利进行语义对比并取均值。在文本层面，文本技术相似度通过比较高校专利文本与企业已有专利文本的语义向量得到，计算公式如式（2）所示。

$$\mathrm{Sim}_{\mathrm{Txt}}(p,E)=\frac{1}{n}\frac{\sum_{i\in E}\cos(\boldsymbol{v}_p,\boldsymbol{v}_i)}{n} \quad (2)$$

其中：$\boldsymbol{v}_p$表示高校专利p文本向量，$\boldsymbol{v}_i$代表企业专利向量集合中第i条专利对应的文本向量，n为企业参与计算的专利样本总数量。该指标衡量高校专利在语义空间中与企业整体技术背景的契合度。

2.1.1.2 技术互补度特征计算

技术互补度旨在表征校企双方技术的异质互补程度，即衡量高校专利是否具备弥补企业技术短板、填补空白或提供差异化资源的能力。该维度体现了企业通过引入外部异质性资源实现技术跨界融合，进而突破技术瓶颈、产出创新成果或拓展技术边界的潜力。基于功能特征的差异与协同机制，本研究提出基于IPC层级结构与IPC共现的两种互补度测度方法。

（1）IPC层次技术互补度。根据Makri等[14]和Frenken等[15]对技术互补度的界定，互补技术指在同一个宏观知识领域内，功能或知识层面相互依赖，但在具体技术实现路径或应用方式上存在差异的技术单元，即总体相同、微观不同。在专利分类中，IPC分类号采用分层结构：大组代表较为宽泛的技术主题或功能领域；小组则是对大组内更细化的技术分支或具体实现方式的划分，表示特定的加工步骤或材料处理方式。因此，当两件专利具有相同的大组时，意味着二者共享同一宏观技术领域，具备了互补的知识基础；若小组不同，则说明二者在具体技术路径、实现方法或功能环节上存在差异。这种大组相同、小组不同的IPC分层差异，能标示出企业技术链条中不同环节之间的功能依赖与协同关系，从而有效表征技术互补度。本研究利用专利IPC分类体系的层次特性（见表1），将IPC大组定义为广义共同领域，界定双方处于同一功能性技术模块；将IPC小组定义为狭义异质领域，以识别同一功能导向下差异化的具体工艺或者材料路径。

高校与企业间的IPC层次技术互补度$\mathrm{Com}_{\mathrm{Hie}}(p,E)$计算公式如式（3）所示。

表1 IPC层级示例

IPC层级	示例	含义	技术表征
部	B	作业；运输	宏观产业领域
大类	B01	物理或化学的方法或装置	产业技术门类
小类	B01D	分离	特定技术系统
大组	B01D 71/00	以材料为特征的用于分离工艺或设备的半透膜；专用于半透膜制备的加工工艺	功能性技术主题（广义共有领域）
小组	B01D 71/56	聚酰胺，例如聚酯酰胺	具体技术实现路径（狭义异质领域）

$$\mathrm{Com}_{\mathrm{Hie}}(p,E)=\frac{\sum_{i\in(M_p\cap M_E)}\left|S_{pi}\setminus S_{Ei}\right|}{\left|S_p\right|} \quad (3)$$

其中：M_p表示专利p所包含的IPC大组集合；M_E表示企业E拥有的专利包含的IPC大组集合；IPC大组集合是高校专利带来的新技术集合；$M_p\cap M_E$表示专利p与企业E专利集所共有的技术主题；S_{pi}表示专利p在共有大组i下所属的小组集合；S_{Ei}表示企业E专利在共同大组i下拥有的所有小组集合；$\left|S_{pi}\setminus S_{Ei}\right|$表示在共同大组$i$下，专利$p$拥有但企业$E$尚未拥有的差异化小组数量；$\left|S_p\right|$表示专利$p$拥有的全部小组总数，用于归一化处理。

（2）IPC共现技术互补度。IPC共现技术互补度基于组合创新理论，旨在衡量新引入的高校技术与企业现有知识体系的兼容与协同程度。若两个技术元素在领域中频繁共同出现，表明二者在功能上具有显著的协同效应，且技术融合的验证成本较低[16]。鉴于IPC大组能较好反映独立的技术功能模块，本研究选取它作为基本技术单元。本研究将IPC共现技术互补度定义为企业现有技术体系与高校专利所引入新技术之间的平均技术关联强度。

首先，采用Jaccard系数计算任意两个IPC大组所代表的技术元素i与j间的关联强度ω_{ij}，计算公式如式（4）所示。

$$\omega_{ij}=\frac{C_{ij}}{C_i+C_j-C_{ij}} \quad (4)$$

其中：C_{ij}为技术元素i与j同时出现的频次，C_i、C_j分别表示包含技术元素i或者j的单独出现频次。ω_{ij}越高表明这两个技术元素在领域内共同出现的确定性越强，即二者在功能或应用上具有更高的互补协同性。

其次，计算高校专利p所引入的新技术与企业E现有技术体系之间的平均关联强度$\mathrm{Com}_{\mathrm{co}}(p,E)$，用以度量两者之间的技术互补度，计算公式如式（5）所示。

$$\mathrm{Com}_{\mathrm{co}}(p,E)=\frac{\sum_{i\in M_p\setminus M_E}\sum_{j\in M_E}\omega_{ij}}{\left|M_p\setminus M_E\right|\times\left|M_E\right|} \quad (5)$$

其中：$M_p\setminus M_E$表示高校专利p中企业E尚未拥有的技术元素，$\left|M_p\setminus M_E\right|$与$\left|M_E\right|$分别表示两个集合的元素个数，用于归一化。

2.1.2　特征体系

为构建高校可转化专利识别模型，需建立一套能够有效表征专利转化潜力的特征体系，作为模型输入的基础与前提。在特征选取上，本研究借鉴现有研究成果，从技术、法律及经济3个维度提取专利价值特征。同时，考虑到校企合作的实际场景，本研究进一步引入地理特征与技术匹配特征，构建高校可转化专利特征体系（见表2），使可转换专利识别更契合真实的转化决策逻辑。

具体而言，在专利价值特征中，技术价值通过发明人数、引用数及IPC小类数等指标刻画专利的创新质量与协作深度[2-6]；法律价值利用权利要求数和专利维持期表征专利的保护范围与权利稳定性[7-9]；经济价值则以同族专利数反映其市场布局能力与商业化潜力[10-11]。地理特征通过同省和同市两个二元变量进行量化，这两个特征用以表征校企间的空间关联，及其在降低交易成本与沟通障碍方面的优势。技术匹配特征包括技术相似度与技术互补度，衡量校企双方在技术层面的关联适配程度，反映技术协同的潜在可能性。

2.2　样本优化

在高校专利转化识别中，实际发生转化的样本相对于海量的潜在配对空间占比极低，极端的样本不平衡会导致模型产生严重的偏差。传统随机负采样引入的简单负样本特征差异过大，使模型难以捕捉驱动转化的细粒度特征。因此，本研究设计了一种基于聚类的困难负样本挖掘策略，通过生成特征相近、转化状态相异的高质量对抗样本，迫使模型在更严苛的条件下学习判别逻辑。需说明的是，本研究负样本指未观测到成功转化而非必然不可转化的配对，构造此类样本有助于缓解因信息不对称等非匹配因素带来的噪声。

该方法的具体实施路径如下。首先，语义向量化与聚类。基于专利摘要构建语义向量，利用K-means算法

表2　可转化专利特征体系

一级特征	二级特征	三级特征	描述与赋值
专利价值特征	技术价值	发明人数	在一定程度上反映技术复杂度与协作深度，通常复杂技术需更多领域专家协作，间接体现其技术价值
		申请人数	在一定程度上反映专利技术的市场认可度与合作价值，通常申请人越多，意味该技术被多方认可或具有更广泛的应用潜力，其技术价值往往相对更高
		被引用数	表明专利技术被后续创新借鉴或依赖的程度
		引用数	表明该专利需要借鉴现有技术的数量，反映出其技术的复杂性
		IPC小类数	专利中包含的国际专利分类号小类数量，反映专利涉及的技术领域范围与技术应用场景广度，数值越高，其技术应用潜力越大
	法律价值	权利要求数	体现了专利的法律保护范围，权利要求数量越多，表明其法律保护越严密，排他性越强，法律价值越高
		专利维持期	专利维持有效法律状态的年份。专利维持期越长，通常表明权利人认为法律保护带来的收益值得持续投入，体现该专利法律价值较高
	经济价值	同族专利数	专利共享完全相同的优先权申请号的简单同族专利数量。同族专利数越多，表明专利在更多市场获得保护，市场覆盖范围越广，商业价值越大
地理特征	地理邻近性	同省	二元变量，若专利所属高校与受让企业位于同一省份，则赋值为1，否则赋值为0，用以反映校企在省级层面的地理邻近性
		同市	二元变量，若专利所属高校与受让企业位于同一地级市，则赋值为1，否则赋值为0，用以反映校企在市级层面的地理邻近性
技术匹配特征	技术相似度	IPC相似度	基于IPC的技术相似度衡量高校专利技术与受让企业已有专利技术的重合程度
		文本相似度	基于专利摘要文本相似度衡量高校专利文本与受让企业已有专利文本的语义相似度
	技术互补度	IPC层次技术互补度	基于IPC层次衡量双方在相同领域下的技术差异
		IPC共现技术互补度	基于IPC的共现频次，衡量双方技术组合的功能互补度

将全量专利划分为若干内部特征趋同的技术主题簇。其次，校企主题映射。根据最大隶属原则，将企业映射至其持有专利分布最集中的技术簇中，建立企业与技术领域的关联逻辑。最后，同簇对抗采样。在正样本所在的聚类簇内实施随机采样，抽取同簇内未发生实际转化合作的专利或企业组成负样本。这种细粒度差异样本不仅能有效剔除简单噪声，更能最大程度模拟真实的供需匹配逻辑，提升模型的鲁棒性与判别精度。

2.3 模型集成

为了突破单一模型在处理高维、非线性特征及样本不平衡时的性能瓶颈，提升识别结果的泛化能力与鲁棒性，本研究构建了一种基于性能加权的集成学习框架。

硬投票或简单平均软投票等传统集成策略通常假设所有基分类器的决策权重完全一致，忽略了不同模型在特定数据分布下的性能差异。鉴于此，本研究提出一种加权软投票机制：根据各基分类器在验证集上的历史表现（F1值）分配差异化权重，赋予高性能模型更大的决策影响力，从而实现多模型的优势互补。

2.3.1 基模型选择

为确保决策视角多元化并提升集成效能，本研究选取6种具有代表性的基模型：逻辑回归（Logistic Regression，LR）用于捕捉线性特征关联；支持向量机（Support Vector Machine，SVM）用于处理高维空间的复杂边界；轻量级梯度提升机（Light Gradient Boosting Machine，LightGBM）与极端梯度提升（eXtreme Gradient Boosting，XGBoost）为高效的梯度提升框架，分别优化训练速度与防过拟合能力；RF利用引导聚集算法思想，通过降低模型方差提升模型的抗噪性；多层感知机（Multilayer Perceptron，MLP）通过非线性激活函数捕捉高阶特征交互。

2.3.2 加权投票

本研究引入验证集F1值作为衡量基模型可信度的标准，作为权重(w)。对于待识别样本x，最终的集成转化概率$P(x)$根据对各基模型识别的正类概率$P_i(x)$归一化得到，计算公式如式（6）所示。

$$P(x)=\frac{\sum_i w_i \cdot P_i(x)}{\sum_i w_i} \tag{6}$$

其中：$P(x)$为对于待识别专利样本x的最终的集成转化概率；i指第i个基模型；$P_i(x)$第i个基模型输出的正类识别概率；w_i是第i个基模型的权重（由验证集F1值确定）。

最终集成模型识别结果如式（7）所示。

$$\hat{y}=\begin{cases}+1\text{（可转化）}, & \text{if } P(x)>\tau \\ -1\text{（不可转化）}, & \text{otherwise}\end{cases} \tag{7}$$

其中：τ表示待识别专利样本x可转化概率阈值。

2.4 归因分析

SHAP（SHapley Additive exPlanations）模型是一种基于博弈论的机器学习模型解释方法，通过计算每个特征对机器学习模型输出的边际贡献值，实现机器学习模型特征重要性量化与解释[17]。本文将SHAP模型应用于高校可转化专利识别模型的结果解析，量化各识别特征（如技术相似度、技术互补度、地理邻近性等）对专利可转化概率的边际贡献，进而识别影响高校专利转化的关键驱动因素。

3 实证研究

膜技术作为利用特殊孔隙结构进行分离净化的先进技术，在水处理及能源等领域应用广泛。近年来该领域专利数量激增，且高校在膜材料等核心方向积累深厚，但相关科技成果长期面临高沉淀、低转化的困境。因此，本研究选取膜技术领域作为实验对象，旨在验证所提方法的有效性，并为破解高校科技成果转化难题提供实证参考。

3.1 样本选取

本研究数据采集自壹专利数据库（https://www.patyee.com/home），该数据库覆盖170余个国家、地区和组织，收录2亿余条深加工专利数据，能提供智能检索、多维分析、预警、导出与协同等服务。检索时间跨度为2014—2024年，专利类型限定为发明专利，法律状态限定为有效及审查中。为了保障数据完备性，本研究构

建了覆盖材料类型（如有机膜、无机膜等）、工艺技术（如微滤、超滤等）及应用形态（如膜组件等）的组合中英文检索式。经数据去重、简单同族合并处理，共获取膜技术领域数据58 052条。在此基础上，本研究聚焦高校科技成果转化场景，甄别出权利人由高校变更为企业（转让）或由高校许可给企业使用（许可）的转化专利。最终，获得474条有效转化数据作为正样本，正样本涉及128所高校与208家受让企业。需要说明的是，本研究以已发生转让或许可作为正样本的界定标准，主要是基于数据可得性与可操作性的考虑。诚然，已交易专利并不完全等同于高可转化专利，未交易专利也不完全等同于不可转化专利。但鉴于大规模真实转化数据的获取难度，本研究采用已交易专利作为正样本替代指标。

3.2　正负样本构建与评估指标设定

3.2.1　正负样本构建

本研究将实际发生转让或许可的高校专利定义为正样本，即已成功实现向产业端转移的专利；将未观测到转化行为的高校专利定义为负样本，即尚未发生转让或许可的专利。根据获得的474条有效专利转化数据作为正样本，在此基础上分3个步骤构建正负样本。

（1）训练集与测试集正样本构建：将正样本按8∶2划分为训练集正样本（379条）与测试集正样本。该比率遵循机器学习模型开发中通用的数据划分惯例。

（2）训练集负样本构建：在训练集正样本的基础上，本研究参照已有文献关于负采样比例与统计检验力及模型训练效率的经典设计与研究实践[18]，选择正负样本比例为1∶4，并引入聚类引导的困难负样本采样策略。具体实施路径：首先进行语义向量化与聚类，基于专利摘要构建语义向量，利用K-means算法将全量专利划分为若干内部技术特征趋同的主题簇；其次进行校企主题映射，根据最大隶属度原则，将企业映射至其持有专利分布最集中的技术簇中，以构建企业—技术领域的关联逻辑；最后实施同簇对抗采样，在正样本所在的同一聚类空间（关系相近的技术主题领域）内实施对抗随机采样，抽取该簇内未发生实际成交或合作的其他专利或企业组装成高难度的对抗负样本。

（3）测试集负样本构建：为模拟真实专利转化识别的应用场景，采用全量候选集配对策略构建测试集负样本。为保持与训练集正负比例的一致性，对全量候选负样本进行随机下采样，使测试集中正样本与负样本的总量比例同样控制在1∶4。其中，负样本由供需错配组与干扰组按1∶1构成。供需错配组由190条已转化专利与非真实受让的所有企业组合而成，用于评估模型对正确受让方的识别精度；干扰组由190条未转化专利与所有企业组合而成，用于评估模型对专利转化价值的判断力。测试集共包含95条正样本及380条负样本。该比例设置一方面与训练集保持一致，便于横向比较模型性能；另一方面避免了真实场景中极端稀疏分布对评估指标的干扰，能够更清晰地检验模型对转化专利的判别能力。

3.2.2　评估指标设定

在高校专利转化过程中存在类别不平衡分布的问题，即转化专利（正样本）数量远少于非转化专利（负样本）。若采用传统准确率作为评价指标，模型会将所有样本识别为负样本以获得虚高评分，导致指标失效。为克服这一问题，本研究选用Hits@N与精确率—召回率曲线下面积（Area Under the Precision-Recall Curve，AUPRC）作为评估指标。Hits@N能评估模型对少数正样本的召回能力，AUPRC对类别不平衡问题具备良好鲁棒性，二者可更客观地反映模型在稀疏转化场景下的真实判别性能。Hits@N指标仅在正样本和供需错配组负样本集合上计算，定义为真实企业排名前N的比例。AUPRC通过计算精确率和召回率在不同阈值下的变化轨迹下面积获得，相较于接收者操作特征曲线（Receiver Operating Characteristic Curve，ROC）曲线，AUPRC更适于评估极度不平衡数据中的正样本识别质量。AUPRC的取值范围为[0，1]。该指标越高，说明模型不仅能准确识别专利与企业，还能有效识别并过滤那些不转移的专利，具有更强的抗干扰能力。

3.3　实验结果

3.3.1　基于技术匹配的识别特征有效性评估

本研究在基模型与随机负采样策略（从非转化专利中随机抽取样本作为负样本）保持不变的条件下，通过对比是否引入本研究所构建的识别特征，检验其对模型性能的提升作用，评估结果如表3所示。仅依赖

专利价值特征时，各模型的AUPRC值均低于0.022，Hits@20值均低于0.260，整体识别表现较差。这一结果说明，仅依靠供给侧指标（如专利价值特征）难以有效区分转化与非转化专利，模型基本不具备正类识别能力，原因在于专利转化本质上是供需匹配的结果，而单一供给侧特征无法反映技术需求方的采纳意愿与匹配程度。引入地理特征与技术匹配特征后，各模型的AUPRC值均有提升，最高可提升至0.292；Hits@20最高为0.665，模型性能发生质变，进一步证明供需适配是专利转化的决定因素。

表3　识别特征有效性评估结果

模型	负采样策略	识别特征	AUPRC	Hits@20
LR	随机	专利价值特征	0.015	0.215
	随机	专利价值特征+地理特征+技术匹配特征	0.045	0.425
SVM	随机	专利价值特征	0.018	0.232
	随机	专利价值特征+地理特征+技术匹配特征	0.068	0.462
LightGBM	随机	专利价值特征	0.021	0.258
	随机	专利价值特征+地理特征+技术匹配特征	**0.292**	**0.665**
XGBoost	随机	专利价值特征	0.014	0.180
	随机	专利价值特征+地理特征+技术匹配特征	0.264	0.620
RF	随机	专利价值特征	0.019	0.245
	随机	专利价值特征+地理特征+技术匹配特征	0.272	0.645
MLP	随机	专利价值特征	0.020	0.255
	随机	专利价值特征+地理特征+技术匹配特征	0.125	0.425

注：加粗数值表示最佳实验结果。

3.3.2　聚类负采样有效性评估

本实验在检验基于聚类的困难负采样策略在解决正负样本不平衡及语义稀疏性问题上的效能。实验设置了两组对比方案：基准组采用传统的全局随机负采样，即从所有非受让企业中随机抽取负样本；实验组则采用本研究提出的聚类负采样策略。

在聚类参数选择与质量评估方面，本文利用K-means算法对样本进行聚类。为科学确定最佳聚类数，实验采用了基于误差平方和（Sum of Squared Errors，SSE）的肘部法则与轮廓系数进行双重验证。实验显示，随着K值的增加，SSE曲线在附近出现明显的转折，下降速度趋于平缓。同时，在该K值下，聚类样本的平均轮廓系数达到局部峰值0.42，表明簇内紧密度与簇间分离度达到较优平衡。基于构建技术主题簇，并在与正样本同簇的非匹配企业中抽取样本，以此构造具备高混淆性的特征相近、转化状态相异的负样本。两组实验均基于全量特征集进行测试。

实验结果如表4所示，引入聚类负采样策略后，所有模型的识别性能均得到进一步提升。性能差异的根本原因在于样本的对抗性强度：随机采样生成的负样本大多与正样本差异巨大（简单负样本），模型无需深入学习特征交互即可轻易区分，导致决策边界松散，泛化能力弱。相反，聚类采样筛选出了技术领域相似但未发生转化的企业，迫使模型去学习更细粒度的特征边界（如IPC互补度的微小差异）。结果表明，该策略有效增强了模型在面对大量高相似干扰项时的鉴别能力与鲁棒性。

表4　聚类负采样有效性评估结果

模型	负采样策略	AUCPRC	Hits@20
LR	聚类	0.158	0.452
SVM	聚类	0.215	0.554
LightGBM	聚类	**0.538**	**0.892**
XGBoost	聚类	0.498	0.882
RF	聚类	0.506	0.888
MLP	聚类	0.285	0.668

注：加粗数值表示最佳实验结果。

3.3.3　加权集成识别模型有效性评估

本节旨在评估本研究所提出的加权集成识别模型的有效性。该模型通过为不同基模型分配差异化权重，融合各算法的识别优势并抑制单一模型的识别偏差，从而提升整体识别性能。为验证该模型的有效性，本研究将其与基模型及其他集成方案（硬投票、平均软投票）在相同实验设置下进行对比。由表5可知，本研究提出的集成模型（加权软集成策略）优于基模型及集成方案（硬投票、平均软投票）。该方法利用各基学习器间的互补度修正偏差，在模拟真实转化场景（全量候选集配对）中展现出较强的鲁棒性，在模拟真实转化场景（全量候选集配对）的测试中，本研究所提出的

加权集成模型表现出稳健的判别性能。加强集成模型的AUPRC达0.592，Hits@20达0.926，均显著优于所有参与对比的单模型（以最优单模型LightGBM为基准，AUPRC相对提升10.0%，Hits@20相对提升3.8%），验证了该模型在稀疏专利转化场景下的鲁棒性。

表 5　集成模型与基模型性能对比

类型	模型	负采样策略	AUPRC	Hits@20
基模型	LR	聚类	0.158	0.452
	SVM	聚类	0.215	0.554
	LightGBM	聚类	0.538	0.892
	XGBoost	聚类	0.498	0.882
	RF	聚类	0.506	0.888
	MLP	聚类	0.285	0.668
集成方案	硬投票	聚类	0.515	0.890
	平均软投票	聚类	0.542	0.915
集成模型	加权集成	聚类	**0.592**	**0.926**

注：加粗数值表示最佳实验结果。

从以上数据分析结果可知，模型性能随策略叠加增强。技术匹配特征与地理特征、聚类负采样以及加权集成学习策略均能提高识别的准确性。

3.4　高校可转化专利驱动因素分析

3.4.1　全局重要性分析

本研究利用SHAP方法量化各特征对识别结果的影响程度（见图1）。结果显示，技术匹配特征是模型决策的首要因素。其中，IPC层次技术互补度、文本相似度和IPC共现互补度位列前三，其影响程度显著高于地理位置（同市区）及专利本身质量（被引数）等特征。IPC层次互补度影响最大，表明高校与企业间的技术转化更多是基于产业链互补而非技术同质。这验证了企业引入高校成果的主要动机在于获取外部异质技术以弥补自身技术短板，而非单纯累积相似专利。

3.4.2　核心维度雷达图分析

图2以表2中二级特征为维度构建，通过对各二级特征下属三级特征的平均SHAP值取算术平均得到各维度分值，分值反映对应二级特征整体对高校可转化专利识别的平均贡献度。从图2中可以观察到，技术互补度的特征重要性得分优于技术相似度，且这两种反映供需双方匹配度的技术匹配特征（互补度、相似度）以及地理特征的整体贡献度高出了专利价值特征（包括法律价值、技术价值及经济价值）。这揭示了高校专利转化的内在逻辑：一方面，虽然技术相似度是企业理解、消化和吸收外部技术成果的认知基础，但在真实的产学研交易中，企业付费购买高校专利的内在动力在于寻求技术互补，即通过引进高校异质性的基础研究成果来填补自身产业链中的技术空白并实现协同增值，这也表明技术相似是技术受让的基本前提，而技术互补是驱动交易达成的实质原因；另一方面，这一数据层级的差异表现反映出高校专利转化的深层内核并非取决于专利固有价值，而在于双方双向契合关系。即使专利的法律、技术和经济等固有价值属性再高，若脱离了与企业工艺所对应的技术匹配需求，或由于地理邻近缺失导致面对面沟通及知识传递的协调阻力过高，其绝对价值也难以转化为现实的成交决策。

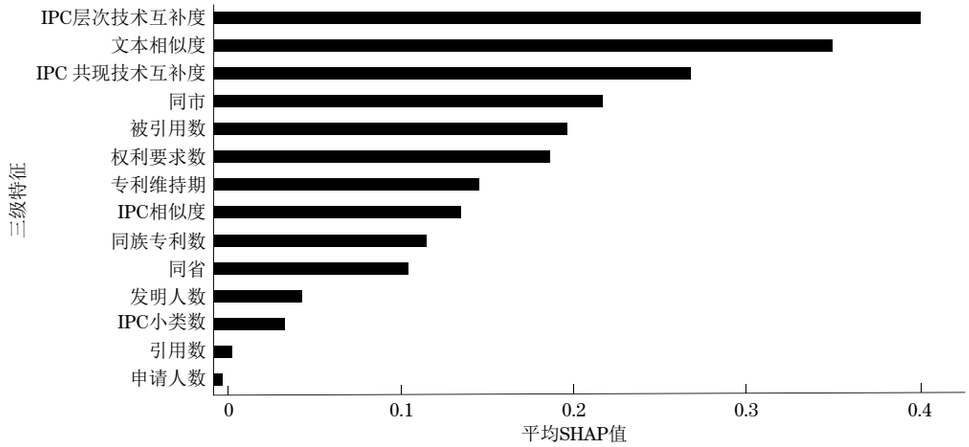

图1　基于平均SHAP值的三级特征全局重要性排序

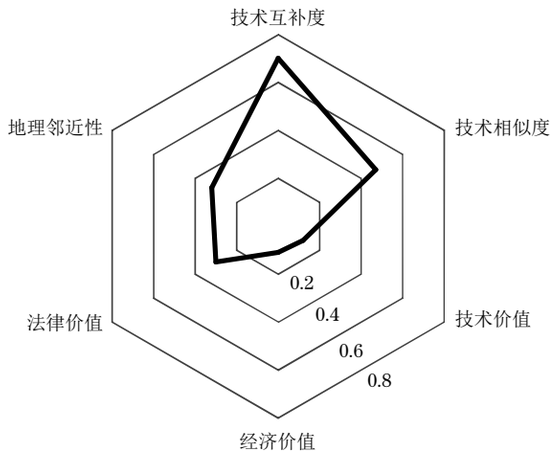

图2　二级特征雷达图

3.4.3　特征影响分布

图3揭示了各特征数值高低对专利转化概率识别结果的正负影响关系。图中纵轴自上而下依次为各个特征，横轴为SHAP值；图内每个散点对应一条研究样本。散点灰度表征该样本对应的特征原始取值大小，深色、黑色代表特征低值，浅色、浅白色代表特征高值；散点在横轴上的位置反映SHAP值大小，位于横轴零值右侧代表该特征对专利转化起到正向促进转化作用，位于零值左侧则代表对专利转化存在负向抑制作用。

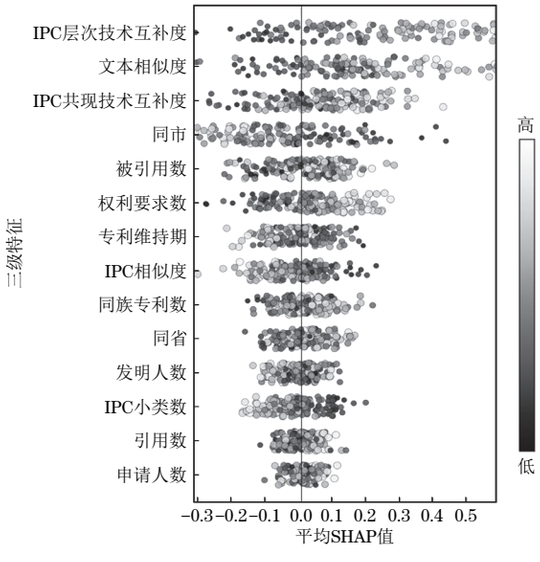

图3　SHAP散点图

分析结果表明技术匹配特征的主导作用：排名前三的特征（IPC层次技术互补度、文本相似度、IPC共现技术互补度）均属于本研究构建的技术匹配特征。特别是IPC层次技术互补度，其高值点（浅色、浅白色）显著集中在右侧，表明高校专利与企业业务在技术层级上的互补度是驱动转化的最强因素。这意味着企业倾向于购买那些能填补其技术空白、完善其技术链条的高校专利，而非仅仅是技术雷同的专利。

4　结语

4.1　结论与建议

针对海量高校专利转化率低、供需匹配难的问题，本研究构建了基于机器学习的专利转化识别框架，并选择膜技术领域进行实证研究。

本研究发现，高校专利能否成功转化，关键在于技术宏观领域同向与微观技术错位的协同作用。一方面，大方向一致（吸收能力）是前提。高校与企业需处于同一个大组或具备高度相似的语义背景。这保证了企业具备足够的吸收能力，能够理解并消化高校的基础技术。另一方面，小路径异质（互补价值）是关键。真正促成交易的并非单纯的同质化叠加，而是高校在具体技术路径上的差异化创新。如企业擅长瓣膜的结构与植入，而高校则提供瓣膜材料的抗血栓处理。这种宏观领域耦合，微观错位的模式，构成专利转化契合点。

基于研究结论，本研究提出以下推动高校专利转化的建议。①优化评价视角，强调需方驱动。高校科研管理应突破单一的质量自评，引入相似度与互补度双维评价，精准刻画专利对目标企业的价值。②聚焦异质互补，引导精准对接。基于产业链互补的微观技术互补是驱动转化的核心动机。高校应重点推介能补足企业技术短板或支持跨界融合的专利，而非简单的同质化成果复制。③推动数智化转型，辅助精准识别。推广应用基于机器学习的匹配识别模型，在转化概率低的背景下挖掘潜在关联，缓解信息不对称，提高存量资产盘活效率。

4.2　贡献与不足

本研究作出如下贡献：首先，基于校企技术匹配视角，重构了多维度识别特征体系；其次，提出了基于聚类的困难负采样策略，有效解决了正负样本严重失衡及语义稀疏性难题；再次，设计了基于性能加权的软集成策略，显著提升了模型在复杂场景下的识别精度与鲁棒性；最后，利用SHAP框架打开模型“黑箱”，量化解析

了驱动转化的关键要素。

本研究存在两个方面不足。一是本研究结论主要源于膜技术领域的实证分析，具有明显的领域特定性。由于材料科学与其他技术门类（如电子信息、生物工程）在技术结构上存在差异，未来研究应进一步将识别框架推广至生物医药、人工智能等技术领域，通过跨领域实证对比，检验本研究所揭示规律在不同学科知识表征体系下的适用性与泛化能力。二是本研究通过随机划分数据集验证了模型在统计层面的泛化能力，但专利转化本质上是一个具有时间滞后性和动态演化特征的决策过程。现有静态采样策略未能充分挖掘专利技术生命周期的演进趋势对转化结果的长期影响。在未来的研究中，应考虑遵从时间序列进行数据集划分，引入时间窗回测方法，通过以过去识别未来的逻辑验证模型在动态时间轴上的识别精度与稳定性。这不仅能有效避免数据泄露风险，更能模拟真实的专利运营管理场景，从而提升识别结果的实用价值。

参考文献

[1] 方曦，李清瑞. 高校专利申请前价值评估体系构建及实证研究[J]. 科技管理研究，2024，44（8）：48–56.

[2] 张更平，王薇，陈红艺. 高校技术转移预测模型构建及归因分析：以区块链技术为例[J]. 图书馆杂志，2025，405（44）：61–73.

[3] 冉从敬，宋凯. 高校可转化专利识别模型构建：以人工智能领域为例[J]. 情报理论与实践，2020，43（11）：79–85.

[4] DENG D，CHEN T. Using machine learning with technological innovation factors to predict the transferability of university patents[J]. IEEE Access，2023，11：131562–131572.

[5] LEE S，KIM C H，YOON J. Technology-transferability analysis of universities and public research institutes using Deep Neural Networks[J]. IEEE Access，2023，11：15196–15205.

[6] 张彪，吴红，高道斌. 基于特征融合的高校可转移专利识别研究[J]. 情报杂志，2022，41（9）：160–165.

[7] 李睿，范九江. 高校专利可转化性与其文献计量特征相关性研究[J]. 科技进步与对策，2021，38（17）：43–50.

[8] 魏太琛，刘敏榕，陈振标. 高校专利技术转移转化价值影响因素实证分析：基于11所一流高校专利转移转化数据[J]. 图书情报工作，2022，66（9）：103–116.

[9] 吴红，李剑飞，崔哲. 基于专利文献的高校专利可转移性特征研究[J]. 情报杂志，2021，40（9）：187–194.

[10] 武玉英，才久然，何喜军. 结合高阶神经元的深度神经网络用于专利可转让性评价：以电子信息技术领域为例[J]. 情报杂志，2021，40（2）：63–68.

[11] 窦志强，马佰莲，许红. 基于发明人特征的高校专利可转化性研究[J]. 情报杂志，2023，42（5）：169–174，53.

[12] 李欣，冯野，马晓迪. 一种基于机器学习的专利可转让性评估方法研究[J]. 情报杂志，2023，42（3）：85–93.

[13] 唐健廷，卢志平. 地方高校专利可转化性评估研究：基于文献计量与BPNN模型对11所高校的分析[J]. 情报杂志，2023，42（2）：175–183.

[14] MAKRI M，HITT M A，LANE P J. Complementary technologies, knowledge relatedness, and invention outcomes in high technology mergers and acquisitions[J]. Strategic Management Journal，2010，31（6）：602–628.

[15] FRENKEN K，VAN OORT F，VERBURG T. Related variety, unrelated variety and regional economic growth[J]. Regional Studies，2007，41（5）：685–697.

[16] 曾德明，周涛. 企业知识基础结构与技术创新绩效关系研究：知识元素间关系维度新视角[J]. 科学学与科学技术管理，2015，10：80–88.

[17] 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.

[18] 刘家伟，白晨，卢超，等. 基于图表征学习与多源特征融合的产学研专利合作预测研究 [J]. 情报学报，2023，42（4）：412–425.

作者简介

俞琰，女，博士，教授，硕士生导师，研究方向：专利数据挖掘、专利分析，E-mail：yuyanyuyan2004@126.com。
董优优，女，硕士研究生，研究方向：专利数据挖掘。
刘攀，男，硕士研究生，研究方向：专利数据挖掘。
杨舸，男，硕士研究生，研究方向：大数据分析。

Research on Identifying University Transferable Patents Integrated with Technology Matching

YU Yan[1,2] DONG YouYou[2] LIU Pan[2] YANG Ge[1]
(1. Nanjing Tech University Library, Nanjing 211816, P. R. China;
2. School of Economics and Management, Nanjing Tech University, Nanjing 211816, P. R. China)

Abstract: Research on the identifying of transferable university patents is an important topic for enhancing the industrialisation level of scientific and technological achievements in universities. However, faced with large volumes of highly heterogeneous and technically dense patent data, there remains a lack of mature and operable evaluation tools and screening mechanisms to effectively identify patents with transfer potential, and related work is still in the exploratory methodological stage. Current research predominantly employs machine learning to mine the intrinsic characteristics of university patents, focusing on assessing their inherent value while neglecting the technological supply-demand compatibility between university patents and potential corporate assignees. To address this gap, this study proposes a technology-matching-based approach for identifying transferable university patents. First, this study constructs identifying features integrated with technology matching. Then, designs a negative sampling method for difficult university patent transformation in sparse patent transformation scenarios. Next, constructs a weighted ensemble identifying model for university transferable patents that leverages the advantages of multiple algorithms. Finally, analyzes the driving factors of university transferable patents based on game-theoretic attribution. Experiments on transferable patent identifying in the membrane field demonstrate that the proposed method is both feasible and effective.

Keywords: Transferable Patent; Universities; Enterprises; Technology Matching; Technology Complementarity; Technology Similarity