

基于多任务学习的慢病成果转化价值二元评估与预测研究
——以糖尿病专利为例

[摘要] **目的/意义。**专利转化是衡量慢病防治科技成果转化（以下简称“慢病成果转化”）的重要指标。本文对慢病专利转化价值与转化时间的综合预测进行了研究，为慢病成果转化提供了决策支撑。**方法/过程。**本文基于2000—2024年期间13373条糖尿病相关专利数据，设计了专利转化特征体系，构建了兼顾专利转让与否与专利转化时间长度**结果/结论。**特征重要性分析表明，申请到公开时间（天）、家族被引证次数、专利类型是影响专利转让的关键因素

[关键词] 慢病；糖尿病；专利转化；转让时长；机器学习

英文题目 Multi-Task Learning-Based Binary Evaluation and Prediction for Commercialization Value of Chronic Disease Research Achievements — A Case Study of Diabetes Patents

SONG Xinrong, XU Qingyuan, GUO Wei, SHAN Zhiguang

State Information Center, Beijing, 100045, China

[Abstract] **Purpose/Significance:** Patent transformation is an important indicator for measuring the technological transformation of chronic disease prevention and treatment. This paper conducts a comprehensive prediction study on the value and time of patent transformation, providing decision support for the technological transformation in this field. **Method/Process.** Based on 13,373 diabetes-related patent records from 2000 to

[Keywords] Chronic diseases; Diabetes; Patent commercialization; Patent transfer duration; Machine learning

1 引言

慢性非传染性疾病（以下简称“慢病”）具有发病率高、病程长、有效控制率低、经济负担重等特点。根据国家卫生健康委统计，2024年我国因慢病导致的死亡占总死亡比例超过80%^[1]。慢病成果转化是慢病防治的重要抓手，而专利是慢病成果中相对标准化的载体之一。糖尿病是重要的代谢类慢病，中国拥有全球最多的糖尿病患者，患者总数约占全球糖尿病患者总数的22%^[2]

学界关于慢病领域的科技成果转化价值分析与评价主要围绕价值维度、评价方法及评价特征展开。在价值维度方面，已有研究重点关注慢病成果转化价值的临床、技术、法律、经济、社会、产业等维度^[4-6]。在评价方法方面，近年来科技成果转化价值评价尤其是专利价值的评价方法逐步向定量化、智能化方向发展，如使用Logistic回归、层次分析法、机器学习和大语言模型文本分析法等^[7-9]。从具体开展的评价实践来看，慢病成果以专利、新药、医疗器械为主要载体，其价值评价需兼顾技术属性、市场需求与临床适用性^[10]。在专利成果转化领域，机器学习被广泛应用于可转化专利识别、价值评估等方面，例如商慧珺等采用Stacking集成学习模型提升可转化专利识别的准确率^[11]；肖宇锋等将机器学习与熵权法结合，实现医药领域高价值专利筛选^[6]。

^[7,11-15]。上述研究在慢病成果转化价值评价方式、专利转让预测等方面已形成较为系统的积累，但对专利转让的时间维度与长期价值关注不足，存在一定局限性。



2 数据和方法

2.1 数据收集

运用 CNdataseed 基于国家知识产权局专利数据整理的“中国专利数据库”进行关键词检索得到 2000 年至 2024 年的糖尿病相关专利。运用 CNdataseed 的“中国专利转让数据库”进行关键词检索得到 2000 年至 2024 年的糖尿病相关专利转让情况。对糖尿病全口径专利和转让专利进行匹配、去重、治理得到了研究数据集。包括 13373 条糖尿病相关专利，其中已转让专利 3489 条，占比 26.08%，未转让专利 9884 条，占比 73.92%。按专利类型看，发明专利 12501 件，实用新型专利 786 件，外观设计专利 86 件。如图 1—3 所示。

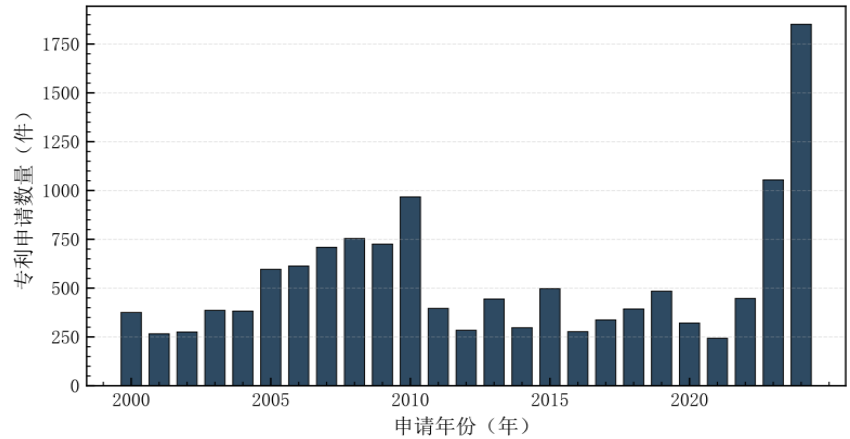


图 1 糖尿病相关专利历年申请数量

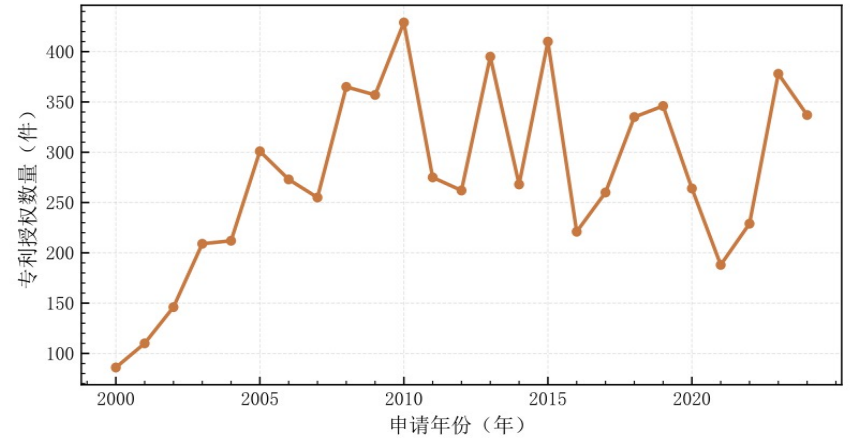


图 2 糖尿病相关专利授权数量趋势

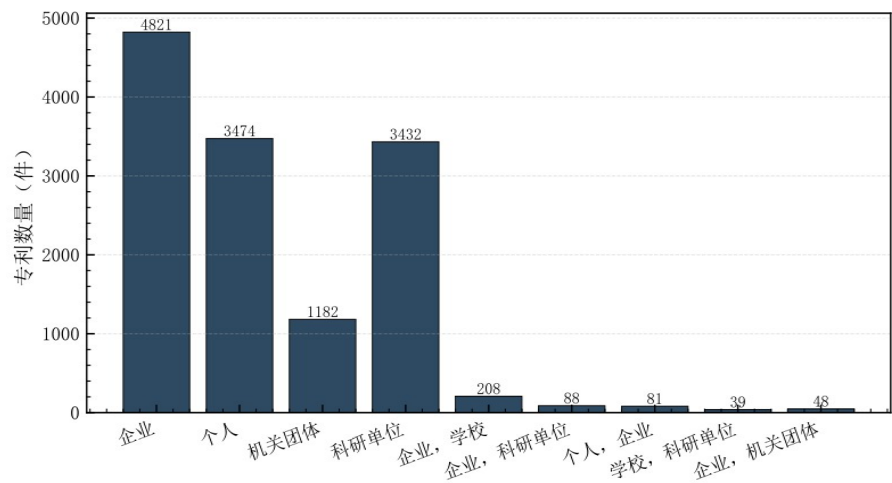


图 3 糖尿病相关专利申请人类型分布

基于专利转让特征，依托多层感知机（MLP）神经网络，构建集专利转让与否和专利转让时长综合预测于一体的转化价值评估模型，并进行了模型优化。

2.2.1 专利 特征

根据文献调研，专利转让特征通常涉及专利的技术价值、法律价值、经济价值等。结合数据可得性，梳理汇总得到了申请人个数、发明人个数、申请到公开时间、公开到授权时间、专利名称字数、专利摘要字数、申请人类型、专利类型、IPC 分类号个数、引证次数、家族引证次数、家族被引证次数这 12 个特征，如表 1 所示。

表 1 专利转让特征表

序号	特征名称	特征含义	特征维度
1	申请人个数 (个)	专利权利归属主体数量	反映专利申请的多主体合作情况
2	发明人个数 (个)	专利发明的研发人员数量	反映研发人员投入情况
3	申请到公开时间 (天)	从提交申请到专利公开的时间间隔	反映专利法律流程和时效
4	公开到授权时间 (天)	从专利公开后进入实审到最终授权的时长	反映专利法律流程和时效
5	专利名称字数	专利名称的总字数	反映技术命名详细程度、技术范围宽窄

6	专利摘要字数	专利摘要的总字数	反映技术信息披露完整度、文本丰富度
7	申请人类型	专利申请机构的类型，区分个人、企业、高校、科研单位、机关团体等	反映创新主体的属性和产学研合作情况
8	专利类型	分为发明专利、实用新型、外观设计三类	反映专利技术含量
9	IPC 分类号个数	专利涉及的 IPC 分类号个数	反映技术领域范围
10	引证次数	专利引用前人专利的数量	反映技术继承性和创新起点
11	家族引证次数	专利所涉及的同族专利整体的引证专利的数量	反映全球专利布局关联、同族技术参考广度特征
12	家族被引证次数	所在专利家族中的专利被其他专利引用的总次数	反映专利家族技术影响力、后续创新贡献度

用于机器学习训练的标签为是否转让、转让时长。专利转让分为申请权转让和专利权转让。专利权转让模式下转让的是已授权专利的专利权，转化后相对更易产业化落地、实际价值更高，因此本文选择研究专利权转让。转让时长定义为从专利授权日到转让生效日的间隔时间。由于转让时长为连续变量，是否转让为离散变量，混合变量类型训练对模型性能不利，因此将转让时长也转化为离散变量。

阈值选取的敏感性分析见

3.3.1 节。

2.2.2 转化价值评估模型

结合慢病专利转化的真实逻辑构建双任务学习模型。一方面，转让时长是转让行为的事后条件观测属性，仅当转让行为真实发生时，转让时长才有可观测的有效值。数学表达式如公式（1）所示：

$$T_{obs} = T \mathbf{g}(Y = 1) + Nan \times I(Y = 0) \tag{1}$$

其中， T_{obs} 为专利转让时长的观测值， Y 为是否转让的二元变量（1=已转让，0=未转让）， T 为专利真实转让时长， Nan 表示缺失值。 $I(Y)$ 为指示函数，当括号内条件成立时取值为1，否则为0。即当专利已转让时， $T_{obs}=T$ ，未转让时， T_{obs} 取缺失值。该标签公式避免了未转让样本对时长预测任务的标签污染。

另一方面，专利转让行为与转让时长是由专利底层特征联合决定的内生关联变量，不存在绝对的单向因果关系。专利的技术质量、市场化潜力等核心特征会同时影响转让的可能性与转让的效率，专利权人对转让时长的预期也会反向影响转让策略与转让结果。通过共享特征层联合学习两个任务，能够捕捉专利转让的底层规律，实现两个任务的正向协同。

基于上述理论框架，本研究构建了双任务 MLP 网络结构，主要包含三大模块：共享特征骨干网络、转让分类头、时长分类头，结构示意图如图 4 所示。

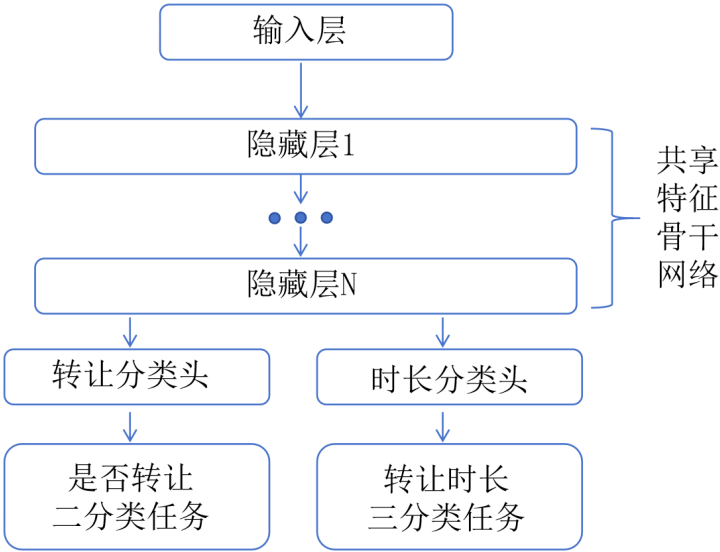


图4 双任务MLP网络结构示意图

其中，共享特征骨干网络由多层全连接单元堆叠构成，每一层依次包含线性变换、激活函数、批量归一化与Dropout正则化操作，实现两个预测任务之间的耦合。转让分类头采用两层全连接网络结构，末端连接Sigmoid激活函数。时长分类头为两层全连接网络，并引入残差连接优化梯度传递，末端连接Sigmoid激活函数；在训练过程中引入掩码机制，仅对已发生转让的有效样本进行损失计算，保证时长分类任务建模的合理性。

$$L_{total} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{reg} \tag{2}$$

其中， L_{total} 为总损失， $\mathcal{L}_{cls}$ 为二分类交叉熵损失，全量样本均参与计算。 $\mathcal{L}_{reg}$ 为三分类交叉熵损失，通过掩码矩阵 M 约束仅已转让样本参与损失计算。 λ_1, λ_2 为损失权重系数。 λ_1 与 λ_2 和为 1。

损失权重敏感性分析见 3.3.2 节。

2.2.3 模型优化

本研究对神经网络激活函数、参数与超参数进行了优化。寻优结果表明，SiLU 激活函数在本研究场景下表现最优，其综合得分达到 0.708，显著优于其他激活函数。本研究还通过 Optuna 超参数搜索工具完成了 50 次迭代的超参数全局寻优，优化目标为是否转让分类 AUC 与转让时长预测的宏 F1 加权值，最优超参数组合下隐藏层数量为 3，第一层隐藏层神经元数量为 32，第二层神经元数量为 256，第三层神经元数量为 192。

3 结果

3.1 模型训练与测试结果

选用主流的 XGBoost、随机森林、逻辑回归作为对比模型。所有模型均在相同数据划分及预处理流程下重新训练。选取 AUC、准确率、精确率、召回率、F1 分数等指标评估模型在转让与否任务分类上的性能。选取准确率、宏 F1 分数、加权 F1 分数等指标评估模型在转让时长分类任务上的性能。

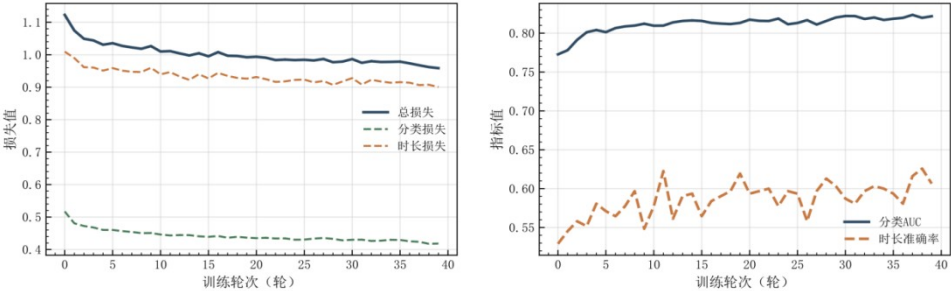


图5 模型训练过程曲线

最优模型的训练过程曲线如图 5 所示。最优模型在测试集上的全指标性能如表 2 和图 6—7 所示。在转让分类任务上，所提出的模型 AUC 达到 0.8207，F1 分数达到 0.533，表明模型具备一定的整体类别区分能力。考虑到已转让样本占比为 26.08%，样本类别存在不平衡，F1 分数绝对水平不高，反映了少数类识别任务的难度。但与对照基线模型相比，本模型 F1 分数显著提升。在转让时长分类任务上，模型准确率达到 0.624，宏 F1 分数达到 0.613，加权 F1 分数 0.611，实现了对转让效率的预测。

表 2 测试集转让与否预测性能统计表

任务类型	AUC	准 确 率	精 确 率	召 回 率	F1 分数
双 任 务 MLP (本 文)	0.8207	0.821	0.833	0.392	0.533
XGBoost	0.7789	0.801	0.873	0.275	0.419
随机森林	0.7726	0.801	0.907	0.262	0.407
逻辑回归	0.7430	0.782	0.690	0.298	0.417

所提出模型在转让与否分类任务上的 ROC 曲线对比如图 6 所示。

图 7

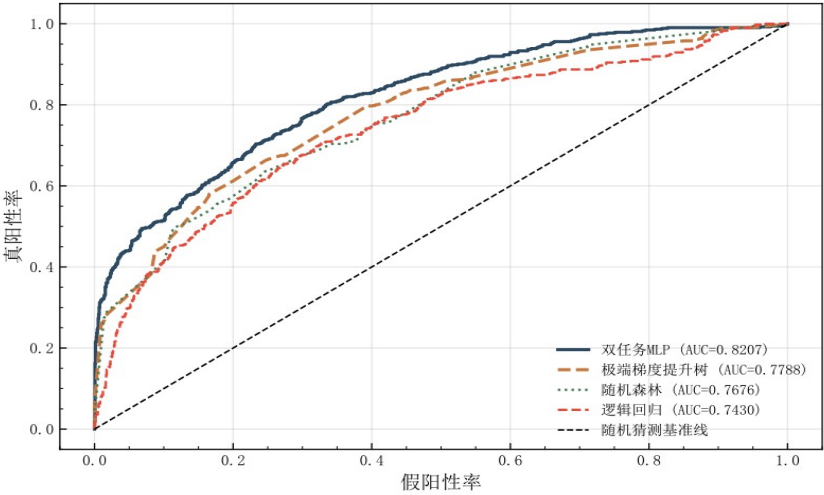


图 6 模型 ROC 曲线对比

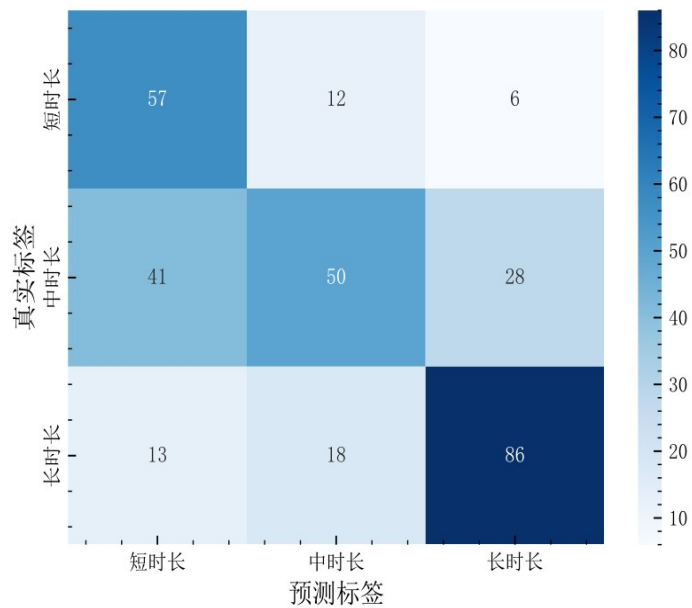


图7 转让时长分类混淆矩阵

3.2 特征重要性分析

[REDACTED]

基于梯度法计算的双任务 MLP 模型特征重要度结果如表3所示。结果表明，申请到公开时间（天）、家族被引证次数、专利类型是影响糖尿病专利转让的核心因素。

表3 双任务 MLP 模型特征重要度

序号	特征	重 要 性
1	申请到公开时间 (天)	0.154
2	家族被引证次数	0.136
3	专利类型	0.084
4	申请人类型	0.082
5	公开到授权时间 (天)	0.079

10.12201/bmr.202610.00040V1

6	引证次数	0.067
7	家族引证次数	0.061
8	IPC 分类号个数	0.045
9	发明人个数（个）	0.041
10	专利名称字数	0.040
11	专利摘要字数	0.039
12	申请人个数（个）	0.031

供
参考
表 4 所示

表 4

3.3.2 损失权重敏感性分析

λ_1 与 λ_2 和为 1，故对其中一个进行优化即可得到一对优化参数。本研究对 λ_1 参数取值在 [0.3, 0.7] 区间内以 0.1 为步长进行了扫描分析，固定其余最优超参数与随机种子，逐取值重新训练模型，并基于测试集 2000 次 bootstrap 重抽样估计各指标的 95% 置信区间，以 $\lambda_1=0.6$ 为参照进行配对 bootstrap 差异检验。结果如表 5 所示， λ_1 在区间内，转让分类任务 AUC 波动范围为 0.8198-0.8226，时长分类任务宏 F1 波动范围为 0.5925-0.6155。配对 bootstrap 检验显示，各取值与 $\lambda_1=0.6$ 相比，转让分类任务 AUC 与时长分类任务宏 F1 的差异 p 值均大于等于 0.05，差异无统计学意义，表明模型性能对损失权重系数在较大取值范围内保持稳健。故选取寻优确定的 $\lambda_1=0.6$ 兼顾了转让分类与转让时长分类任务性能。

表 5 损失权重阈值敏感性测试结果

λ_1	转让分类 AUC (95%置信区间)	时长分类宏 F1 (95%置信区间)	AUC 差异 p 值	宏 F1 差异 p 值
0.3	0.8226 (0.8009-0.8427)	0.6102 (0.5525~0.6642)	0.151	0.710
0.4	0.8198 (0.7983-0.8400)	0.5925 (0.5323~0.6455)	0.718	0.052
0.5	0.8200 (0.7984-0.8402)	0.6155 (0.5571~0.6694)	0.564	0.991
0.6 (本文采用)	0.8207 (0.7979-0.8395)	0.6134 (0.5545~0.6691)	—	—
0.7	0.8200 (0.7982-0.8403)	0.6111 (0.5531~0.6668)	0.562	0.671

作者贡献：宋心荣负责提出选题、研究设计、数据整理与分析、论文撰写；徐清源负责选题分析、文献调研；郭伟负责政策梳理、文献调研；单志广负责研究设计、论文审核与修订。

利益声明：所有作者均声明不存在利益冲突。

参考文献

[1] 国务院新闻办公室.我国慢病发病率呈上升趋势 慢病死亡人数占比超过 80%[EB/OL]. (2024-9-12) [2026-08-21].<https://news.cctv.com/2024/09/12/ARTI1uiaQICinOmYL8uioAt5240912.shtml>

- [2] Zhou Y C , Liu J M , Zhao Z P ,et al.The national and provincial prevalence and non-fatal burdens of diabetes in China from 2005 to 2023 with projections of prevalence to 2050[J].Military Medical Research, 2025, 12(1).DOI:10.1186/s40779-025-00615-1.
- [3] 陈喆,蒋薇薇,朱佳敏,等.全球重大慢病专利演进态势及我国发展对策研究[J].中国发明与专利,2025,22(3):65-73.
- [4] 杨宁燕,刘欢梅,阎早芳,等.2016~2023 年全球 10 个主要国家心血管领域技术创新专利的现状分析和评价[J].中国循环杂志,2024,39(10):956-967.
- [5] 李玺,孟天华,温杰.试论市场经济体制下医药卫生科技成果内涵价值的六要素[J].中华医学科研管理杂志,1998,(3):33-34+32.
- [6] 肖宇锋,杨雪梅,黄雅兰,等.基于机器学习和熵权法的医药领域高价值专利识别研究[J].中国发明与专利,2025,22(04):13-21.
- [7] 胡嘉蕴,赵楠.适用于医药生物领域审查阶段的专利价值评估体系的建立与验证[J].中国发明与专利,2019,16(9):90-97.
- [8] 李依诺.技术转化视角下高价值专利的识别研究[D].南京:东南大学,2024.DOI:10.27014/d.cnki.gdnau.2024.003103.
- [9] 高腾愉.高校高价值专利识别与研发策略研究[D].镇江:江苏科技大学,2024.
- [10] 张洋,王春颖,许吉,等.6 所中医药院校专利转化趋势及技术布局的可视化分析[J].中医药管理杂志,2023,31(15):1-7.DOI:10.16690/j.cnki.1007-9203.2023.15.017.
- [11] 商慧琚.基于 Stacking 集成学习的可转化专利识别研究[D].济南:山东大学,2023.DOI:10.27272/d.cnki.gshdu.2023.001342.
- [12] 程春铭.基于迁移学习的专利转让预测和分析模型研究[D].呼和浩特:内蒙古大学,2024.DOI:10.27224/d.cnki.gnmdu.2024.001186.
- [13] 来音,李婉丽,李峥嵘,等.深度注意力机制下的高校专利价值评估方法——基于历史转化特征锚定效应视角[J].情报杂志,2023,42(08):200-207.
- [14] 王若雅.基于主体异质性的专利转化影响机制研究[D].北京:北京交通大学,2023.DOI:10.26944/d.cnki.gbfju.2023.003898
- [15] 孙笑明,熊旺,宇文乐薇,等.高校与科研院所沉睡专利价值评估[J].科技进步与对策,2025,42(06):99-108.