计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(9): 2231−2247, 2026

面向可解释人机协同的知识引导主动学习

邵杰晶　杨　铭　杨骁文　李宇峰
（南京大学人工智能学院　南京　210023）
（计算机软件新技术全国重点实验室（南京大学）　南京　210023）
（shaojj@lamda.nju.edu.cn）

Knowledge-Guided Active Learning for Explainable Human-AI Collaboration

Shao Jiejing, Yang Ming, Yang Xiaowen, and Li Yufeng
（*School of Artificial Intelligence*, *Nanjing University*, *Nanjing* 210023）
（*National Key Laboratory for Novel Software Technology* (*Nanjing University*), *Nanjing* 210023）

Abstract　Label scarcity remains a major obstacle to deploying machine learning in real applications, because many tasks require costly annotation by domain experts. Active learning reduces this cost by selecting valuable unlabeled instances for expert annotation. Existing methods usually design query criteria from a learner-centric view, such as uncertainty, diversity, committee disagreement, and core-set coverage. However, their gains are not stable across datasets, models, and annotation budgets, and an unsuitable strategy may even underperform random annotation. Meanwhile, conventional active learning often restricts experts to the role of label providers, lacking a collaborative medium that explains why samples are selected and what decision logic the model follows. As a result, expert feedback is difficult to elevate from individual labels into reusable knowledge rules, and the model cannot be effectively aligned with expert domain knowledge. To address these issues, we propose KALE, a knowledge-guided active learning method for explainable human-AI collaboration. KALE uses first-order logic rules as the collaborative medium between the model and experts. It first induces compact logic rules from the current model behavior, and then asks experts to judge their correctness according to domain knowledge. Correct rules are used for deductive weak-label expansion, while incorrect rules are treated as symbolic observations of knowledge-model disagreement and used to drive abductive disagreement querying. KALE further combines class-level rules, local rules, and sparse rule selection to keep expert interaction manageable. Experiments show that KALE achieves more robust improvement than existing active learning baselines, especially under small annotation budgets.

Key words　machine learning; active learning; human-AI collaboration; logic rule; weakly-supervised learning

摘　要　机器学习在真实场景中的有效应用通常依赖充足可靠的标注数据，而许多任务的标注需要领域专家参与，成本较高。主动学习通过选择高价值未标注样本交由专家标注，以降低数据构建成本。现有方法多从学习器视角设计查询准则，如不确定性、多样性、委员会分歧和核心集覆盖等，但其跨数据集、模型和预算条件的收益并不稳定，甚至可能劣于随机标注。同时，传统主动学习通常将专家限定为标签提供者，缺少解释样本选择依据和模型预测逻辑的协同媒介。专家反馈因此难以从样本标签上升为可复用的知识规则，模型也难以与专家领域知识有效对齐。针对上述问题，提出面向可解释人机协同的知识引

收稿日期：2026-06-06；**修回日期**：2026-07-16
基金项目：国家重点研发计划（2022ZD0114803）；国家自然科学基金项目（62576162）；江苏省基础研究计划重点项目（BK20243012）
This work was supported by the National Key Research and Development Program (2022ZD0114803), the National Natural Science Foundation of China (62576162), and the Key Program of Jiangsu Science Foundation (BK20243012).
通信作者：李宇峰（liyf@nju.edu.cn）

导主动学习方法KALE。KALE以一阶逻辑规则作为模型与专家的协同媒介,先将当前模型预测行为归纳为紧凑逻辑规则,再由专家依据领域知识判别规则正确性。对正确逻辑规则,KALE执行演绎式弱标注扩展;对错误逻辑规则,KALE将其视为知识-模型分歧的符号化观测,并据此开展反绎式分歧查询。方法同时结合类别级规则、局部规则和稀疏规则选择,以控制交互规模。实验表明,KALE较已有基线取得稳定性能提升,在小预算标注场景下优势更加明显。

关键词 机器学习;主动学习;人机协同;逻辑规则;弱监督学习

中图法分类号 TP181; TP391

DOI: 10.7544/issn1000-1239.202660454 **CSTR:** 32373.14.issn1000-1239.202660454

机器学习已经在语音识别、推荐系统、医疗诊断、金融风控和科学发现等场景中取得显著进展[1-3]。这些成功通常隐含一个重要前提:训练阶段能够获得数量充足、质量可靠且与目标场景分布匹配的标注数据。然而,在许多真实应用中,标注并不是简单的众包任务,而需要医生、工程师、化学实验人员、风控专家等领域专家参与判断。此类标注过程往往成本高、周期长,还可能受到实验条件、业务流程和专家时间的限制。特别是在系统上线初期,开发者通常只能获得少量标注样本,却需要尽快训练出可用模型。因此,如何在有限标注预算下选择最有价值的数据,是机器学习落地过程中长期存在的关键问题。

主动学习为降低标注成本提供了有效思路。其基本思想是将人类专家纳入学习闭环,由模型在未标注数据池中主动选择最值得标注的样本,再交由专家给出标签,从而用较少标注获得较高模型性能。根据样本选择准则的不同,典型主动学习方法大致可分为不确定性、多样性和分歧等类型。不确定性方法选择模型最难判断的样本,认为这些样本最可能改善决策边界[4-5];多样性方法选择能够代表整体数据分布的样本,以减少标注集偏置并提升泛化能力[6-8];分歧方法则通过委员会模型或假设空间差异选择能够缩小版本空间的样本[9]。近年来,核心集覆盖、梯度匹配和表示空间采样等方法进一步丰富了主动学习的查询准则,使其能够适配更复杂的模型和数据类型。

尽管主动学习已经在多个领域取得应用,现有方法的性能收益仍然存在明显不稳定性。近年研究者的系统性实验表明,同一主动学习策略在不同数据集、模型结构、特征表示和标注预算下可能呈现截然不同的效果[10-12]。某些策略在一个任务上显著优于随机标注,却可能在另一个任务上提升有限,甚至低于随机选择。对于真实应用而言,这一问题尤为突出,因为专家标注成本具有不可逆性:一旦错误策略消耗了有限预算,被选择样本的标注成本就无法回收。造成这一现象的重要原因在于,现有方法大多从学习器视角定义“样本价值”,即依据模型不确定性、样本分布覆盖、特征空间距离或模型间分歧来排序样本。这些准则虽然在机器学习意义上合理,但其有效性往往依赖具体模型和任务条件,难以保证跨任务稳定收益。

更重要的是,传统主动学习虽然在形式上把人类纳入训练过程,但这种“人在回路中”并不等同于深入的人机协同。大多数方法将领域专家限定为标签提供者:系统向专家提交被选样本,却很少解释样本为何被选中,也很少暴露模型作出预测所依据的决策逻辑。对于医生、工程师或化学专家而言,模型不确定性、梯度距离、嵌入空间覆盖等指标通常并不直观,专家只能被动完成标注,而难以判断模型当前学习到了哪些合理模式、哪些错误模式。由于缺少可解释的人机协同媒介,专家反馈往往停留在单个样本标签层面,难以上升为可复用的知识规则,从而导致模型无法与专家所掌握的丰富领域知识进行有效对齐。

从知识利用角度看,神经符号学习和反绎学习为缓解上述问题提供了重要启发。神经符号学习尝试将逻辑规则、语义约束、概率逻辑或答案集程序等符号知识嵌入模型训练过程,使模型预测更加符合任务知识并提升数据效率[13-19]。反绎学习则进一步强调利用背景知识发现和修正模型产生的错误符号事实,将知识约束作为诊断模型错误的依据[13-14]。这些研究表明,领域知识不仅可以作为训练正则项,也可以用于解释模型行为、发现模型错误并反馈学习过程。然而,这类方法通常假设任务知识已经被整理为较完备的规则系统。对于主动学习面临的初始数据构建阶段而言,这一假设往往过强,因为此时不仅标注数据稀缺,结构化知识库也未必已经存在。

已有知识驱动主动学习方法开始尝试将符号规则引入样本选择过程，例如利用给定规则的满足程度评价未标注样本价值[20]。这类方法说明，从知识视角设计主动学习策略有助于弥补单纯学习器视角的不足。但在真实场景中，领域知识通常并不以完整规则库形式预先给出，而是隐含在专家的经验判断和任务理解中。如果直接要求专家在系统使用前构建完整规则系统，既增加了知识获取成本，也不符合主动学习逐步探索数据空间的特点。因此，更合理的方式不是把规则库作为外部输入前提，而是在主动学习过程中逐步显式化专家知识：模型先暴露自身已经学到的预测逻辑，专家再围绕这些候选逻辑进行审核，系统进一步把专家审核结果转化为样本选择和弱标注信号。

基于上述认识，本文提出面向可解释人机协同的知识引导主动学习方法 KALE（knowledge-guided active learning）。KALE 以一阶逻辑规则作为模型与领域专家之间的协同媒介。首先，系统对当前模型在未标注数据池上的预测行为进行归纳式规则化，将样本级预测结果压缩为专家可理解的逻辑规则，用以描述模型在不同数据区域中的预测逻辑。其次，领域专家根据任务知识判断这些逻辑规则是否合理，从而把原本隐含在专家经验中的知识显式反馈给系统。最后，KALE 根据专家判断沿 2 条路径更新主动学习过程：对确认正确的逻辑规则，执行演绎式弱标注扩展，利用规则覆盖样本形成可信标注候选；对确认错误的逻辑规则，将其视为知识-模型分歧的符号化观测，并通过反绎式分歧查询选择最可能解释和修正该冲突的样本。

与传统主动学习相比，KALE 的关键区别在于，专家不再只是对孤立样本给出标签，而是能够审查模型归纳出的预测逻辑，并将领域知识反馈到样本选择机制中。与依赖预设规则库的知识驱动方法相比，KALE 不要求专家预先提供完整知识系统，而是围绕模型当前行为生成候选逻辑规则，降低了知识获取门槛。与一般可解释机器学习方法相比，KALE 也不止于解释模型预测，而是将解释结果进一步转化为主动学习中的弱标注扩展和分歧查询信号。由此，KALE 形成了“归纳—演绎—反绎”的知识引导闭环：归纳环节暴露模型预测逻辑，演绎环节扩展可信标注候选，反绎环节定位并修正知识-模型分歧。

本文的主要贡献包括：

1）提出一种面向可解释人机协同的知识引导主动学习方法，将模型逻辑规则解释、专家逻辑规则审核和主动样本查询统一到同一学习闭环中；

2）设计模型预测逻辑的归纳式规则化模块，通过类别级逻辑规则、局部逻辑规则和稀疏逻辑规则选择得到规模可控的可解释规则集合；

3）提出逻辑规则交互驱动的标注与查询机制，将正确逻辑规则用于演绎式弱标注扩展，将错误逻辑规则用于反绎式分歧查询；

4）在 9 个分类任务上与不确定性、多样性、分歧、核心集和知识驱动主动学习方法进行比较，验证 KALE 在不同任务和标注预算下的稳定性，尤其是在小预算标注场景中的有效性。

1 相关工作

本文研究面向可解释人机协同的知识引导主动学习，涉及主动学习、反绎学习与神经符号推理、可解释机器学习等相关方向。主动学习为降低标注成本提供了基本问题框架，但现有方法多从学习器视角设计查询准则，缺少对领域专家知识的显式利用；反绎学习和神经符号推理说明符号知识可以用于约束、诊断和修正模型行为，但通常依赖预先构建的规则系统；可解释机器学习则为暴露模型预测逻辑提供了技术基础，但多数方法主要面向预测解释，而非主动学习中的样本选择。下面分别回顾上述 3 类工作，并说明本文与已有研究的区别。

1.1 主动学习

主动学习研究关注如何用尽可能少的标注获得尽可能高的模型性能。不确定性采样基于“模型最不确定处最值得询问”的假设，常用边际置信度、熵或最小最大后验概率等指标[4-5]。多样性采样强调代表性，典型方法包括层次聚类采样和核心集选择[6-8]。传统分歧采样通过多个假设或贝叶斯近似估计模型分歧，以减少版本空间[9,21]。这些策略实现简单、适配性强，但它们多以学习器自身为中心，缺少与领域专家知识的直接对接。本文讨论的分歧不同于多模型间分歧，而是模型预测逻辑与领域知识之间的知识-模型分歧。

跨数据集实验表明，主动学习策略的相对排序会随数据集、模型、特征类型和标注预算而变化[10-12]。特别是在小预算阶段，错误选择策略的代价不可逆，因为已经消耗的专家标注无法回收。刘明阳等人[22]在视频问答任务中引入多模态知识主动学习，说明主动学习与外部知识融合能够提升复杂多模态推理任务的效果。这与本文的出发点一致，但本文进一

步关注通用分类任务中如何通过可解释逻辑规则实现专家知识注入，并把专家判断用于样本选择。

近年的主动学习研究进一步关注跨任务泛化、查询策略元学习和大模型增强。L2R-AL 通过多任务双层优化学习跨任务的样本排序函数，试图降低手工查询准则对特定任务的依赖[23]；CDALBench 从视觉、自然语言和表格数据等多个模态系统评估主动学习方法，进一步表明查询策略的相对优势会随数据域、模型和随机划分发生变化[24]；ActiveLLM 则利用大语言模型的语义能力辅助文本少样本实例选择[25]。这些方法借助跨任务先验或预训练语义增强查询，而 KALE 以逻辑规则将当前学习器行为显式化，并通过专家判断形成知识-模型分歧信号。由于现有大模型增强方法主要面向具有自然语言语义的任务，本文未将其直接并入结构化表格数据上的统一 GBDT 实验流水线，而将大语言模型辅助规则生成与预审核作为后续扩展方向。

1.2 反绎学习与神经符号推理

反绎学习旨在连接机器学习和逻辑推理：学习模型从数据中预测符号事实，逻辑推理模块利用背景知识发现不一致并修正错误事实，修正结果再反馈给模型训练[13-14]。这一思想说明，知识不只可以作为训练约束，也可以作为诊断模型错误的推理依据。本文借鉴这一思想，但关注点从给定知识库下的事实修正转向主动学习中的知识-模型分歧诊断和样本查询。

神经符号学习试图把可解释的符号知识与可微学习模型结合。DL2 通过逻辑公式构建可训练约束[15]，Semantic Loss 把符号知识转化为语义损失[16]，CL-STE 利用直通估计器把逻辑约束注入神经网络[17]。LTN、NeurASP 和 DeepProbLog 进一步结合模糊逻辑、答案集程序或概率逻辑，形成知识推理模型[18-19]。这些方法通常在训练目标层面使用规则，适合已有知识库较明确的任务。

主动学习场景的困难在于：逻辑规则往往尚未被完整整理，专家也未必能一次性表达全部知识。已有知识驱动主动学习 KAL 利用先验符号规则度量样本的规则满足程度[20]，但仍依赖预先构建的规则系统。本文与其不同，不把规则系统作为输入前提，而是从模型预测中归纳候选逻辑规则，让专家围绕候选逻辑规则做判断，再把判断结果组织为演绎式弱标注和反绎式查询信号。

1.3 可解释机器学习

可解释机器学习旨在把复杂模型的预测行为转化为人类可理解的形式。LIME 和 SHAP 用局部线性近似解释单个预测[26-27]，Anchor 用高精度 if-then 规则描述局部决策区域[28]，RuleFit 等方法用规则集或规则组合近似模型行为[29]。这些解释方法通常以“解释已有预测”为目标，而不是直接服务于数据选择。

在高可信智能软件中，可解释性不仅是结果展示问题，也是需求分析和人机协作问题。裴忠一等人[30]指出，机器学习应用需要把领域知识、业务逻辑、数据语义与模型行为显式关联起来，才能支撑工业和医疗等场景中的可解释性需求。本文沿着这一思路，把模型解释结果转化为专家可审核的逻辑规则，并进一步把审核结果变成主动学习中的样本选择依据。

2 预备知识

本文考虑池式批量主动学习设定下的分类任务。设输入特征空间为 X，标签空间为 $Y=\{1,2,\cdots,C\}$，其中 C 表示类别数。给定样本 $x\in X$，其真实标签 $y\in Y$ 通常需要由领域专家或可靠标注过程获得。学习目标是在有限标注预算约束下，训练分类模型 $h:X\to Y$，或等价地学习能够输出类别预测概率的模型 $h(y|x)$，使模型在未见样本上的预测性能尽可能高。与标准监督学习不同，主动学习并不假设训练初始阶段已经拥有充分标注数据，而是从一个规模较小的初始标注集 $L^0=\{(x_i,y_i)\}\{i=1\}^{n_{\text{lab}}}$ 和一个规模较大的未标注样本池 $U^0=\{x_i\}\{i=1\}^{n_{\text{unl}}}$ 出发，其中 $n_{\text{lab}}\ll n_{\text{unl}}$。由于初始标注集在样本数量、类别覆盖和特征分布上通常都不充分，仅依赖 L^0 训练得到的初始模型 h^0 往往难以形成可靠的分类边界，也难以支撑稳定的泛化预测。

在第 t 轮主动学习中，系统基于当前标注集 L^t、未标注池 U^t 以及当前模型 h^t，从 U^t 中选择一个大小为 b 的样本批 $Q_X^t\subseteq U^t$ 交由专家标注，其中 b 表示单轮标注预算。专家返回这些样本的真实标签后，得到新的标注样本集合 $Q^t=\{(x,y)|x\in Q_X^t,y\in Y\}$。随后，主动学习过程将 Q^t 加入当前标注集，并从未标注池中移除相应样本，即 $L^{\{t+1\}}=L^t\cup Q^t$，$U^{\{t+1\}}=U^t\setminus Q_X^t$；模型也随之由 h^t 更新为 h^{t+1}。经过多轮迭代后，主动学习希望在总标注预算 $B=T\times b$ 受限的条件下，优先获取对模型性能提升最有价值的样本标签，从而以低于全量标注的成本获得尽可能接近充分监督学习的分类效果。

在这一设定中，主动学习策略的核心问题可以表述为：在每一轮有限预算 b 下，如何从大量未标注

样本 U^t 中选择查询批 Q_X^t，使新增标注样本能够最大程度改进后续模型 h^{t+1} 的性能。传统方法通常依据模型不确定性、样本多样性或学习器分歧程度来度量样本价值，但这些度量主要从学习器自身状态出发，未必能够充分利用专家在标注过程中隐含的领域知识。因此，本文后续将在上述池式批量主动学习问题定义基础上，进一步讨论如何以可解释的逻辑规则作为模型与专家之间的协同媒介，使专家知识能够参与样本选择与标注扩展过程。表 1 总结了全文主要符号。

Table 1　Notations of Main Symbols

表 1　主要符号说明

符号	配置环境
X	输入特征空间
Y	标签空间
$C=\|Y\|$	类别数
L^t	第 t 轮已有标注样本集
U^t	第 t 轮未标注样本池
$n_{\text{lab}}, n_{\text{unl}}$	初始标注样本数与未标注样本数，通常 $n_{\text{lab}} \ll n_{\text{unl}}$
$h^t: X \to Y$	第 t 轮基于 L^t 训练得到的分类模型
b	单轮专家标注预算，即查询批大小
$Q_X^t \subseteq U^t$	第 t 轮从未标注池中选出的待查询样本批
Q^t	第 t 轮经专家标注后得到的新标注样本集
$R=\{r_i\}_{i=1}^{m}$	从模型预测行为中归纳得到的候选逻辑规则集合
$r_{\text{condition}}$	逻辑规则 r 的条件部分
$r_{\text{prediction}}$	逻辑规则 r 的预测部分
$Coverage(r)$	满足逻辑规则 r 条件部分的样本集合
$R_{\text{T}} \subseteq R$	专家确认正确的逻辑规则集合
$R_{\text{F}} \subseteq R$	专家确认错误的逻辑规则集合
m	各类别稀疏选择后的交互逻辑规则数上限

3　知识引导的主动学习方法

本文在池式批量主动学习设定基础上，提出知识引导的主动学习方法 KALE。该方法的核心思想是：不再仅从学习器内部状态度量样本价值，而是把当前模型的预测行为转化为领域专家可理解、可审查的逻辑规则，并进一步把专家对逻辑规则的判断转化为标注扩展和样本查询信号。由此，专家不只是为系统提供单个样本标签，而是能够围绕模型已经形成的预测逻辑进行审查，使领域知识以结构化方式进入主动学习循环。

KALE 的总体流程如图 1 所示。首先，系统基于当前模型 h 在未标注池 U 上的预测结果，归纳得到候选逻辑规则集合 R。每条逻辑规则由条件部分和预测部分组成，可表示为“若 $condition_1(x)$ 且 $condition_2(x)$，则模型预测类别 y”，用于刻画模型在某一特征区域内的预测逻辑。其次，领域专家根据任务知识审查这些逻辑规则，并将其中一部分标记为正确逻辑规则 R_{T} 或错误逻辑规则 R_{F}。最后，系统根据专家判断沿 2 条路径更新主动学习过程：正确逻辑规则用于演绎式弱标注扩展，错误逻辑规则用于反绎式分歧查询。

上述流程形成“归纳—演绎—反绎”的知识引导闭环。归纳环节把样本级模型预测提升为逻辑规则级预测逻辑；演绎环节把专家确认正确的逻辑规则作用到其覆盖样本上，生成可信弱标注候选；反绎环节把专家确认错误的逻辑规则视为知识-模型分歧的符号化观测，优先选择最可能解释并修正该分歧的样本进行人工标注。通过这一闭环，KALE 把可解释逻辑规则交互嵌入主动学习过程，使领域知识能够同时参与模型解释、标注利用和查询决策。

3.1　模型预测逻辑的归纳

为了使领域专家能够深度参与主动学习过程，KALE 首先需要把当前模型的预测行为转化为专家可理解、可审查的表达形式。设当前主动学习轮次中的模型为 h，未标注池为 U。与直接要求专家预先给出完整领域规则不同，本文从模型在 U 上的预测结果出发，归纳其已经形成的预测逻辑，再交由专家

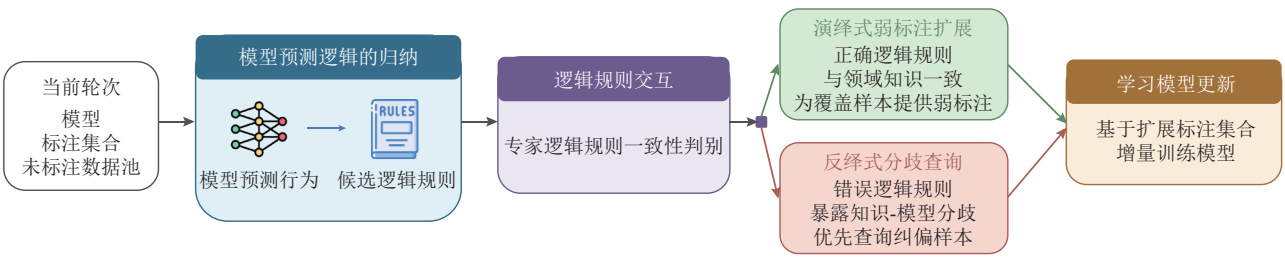


Fig. 1　Overall framework of KALE

图 1　KALE 总体框架

判断这些预测逻辑是否符合领域知识。这里得到的逻辑规则不是预设的真实知识，而是当前模型行为的符号化解释；因此，其作用在于暴露模型“正在如何预测”，为后续专家审查、弱标注扩展和分歧查询提供交互载体。

模型预测逻辑的归纳需要同时满足2个要求。1)忠实性，即生成的逻辑规则应尽可能准确地刻画当前模型 h 在未标注池 U 上的预测行为，而不是另行训练一个与 h 无关的规则分类器。若逻辑规则不能反映模型真实预测模式，专家对规则的判断就难以反馈到模型错误区域。2)简洁性，即逻辑规则数量和表达复杂度必须受到控制。主动学习的目标是降低专家标注成本，如果生成大量冗长规则，专家审查负担可能超过逐样本标注本身，从而削弱人机协同的实际价值。为兼顾这2个要求，KALE采用“类别级逻辑规则归纳、局部预测逻辑规则补充、稀疏逻辑规则选择”的三阶段过程，把样本级模型预测压缩为规模可控的候选逻辑规则集合 R。

1)类别级预测逻辑规则归纳

KALE首先从类别维度归纳模型的主要预测模式。对于每个类别 $c \in Y$，构造二分类数据集 $\{(x, I(h(x)=c)) | x \in U\}$，其中 $I(\cdot)$ 为指示函数，表示当前模型 h 是否把样本 x 预测为类别 c。随后训练决策树 T_c 拟合这一二分类数据集。由于该数据集的标签来自模型预测而非人工真实标签，决策树 T_c 学习到的是模型关于类别 c 的预测行为，而不是任务本身的真实判别规则。

决策树具有天然的逻辑解释形式。树中每个分裂条件都可以看作一个逻辑原子，例如“*age*>20”或“*sex*=Female”；从根节点到某个节点的路径则对应若干逻辑原子的合取，例如“*age*>20且*sex*=Female”。若该路径主要覆盖被模型预测为类别 c 的样本，则可以得到候选逻辑规则：“若样本满足该路径条件，则模型倾向于预测类别 c”。通过遍历决策树节点，KALE可以为每个类别生成一组类别级候选逻辑规则 R_c。这类规则刻画了模型在全局范围内预测某一类别时依赖的主要条件组合，是后续专家理解模型整体行为的基础。需要说明的是，决策树 T_c 在本文中是拟合当前模型输出的后验代理模型，而不是对基学习器类型的限定。候选逻辑规则归纳只需要访问 $h_c(x)$ 或模型预测类别，因此在接口层面可以作用于线性模型、神经网络和其他黑箱分类器。

2)局部预测逻辑规则补充

仅依靠类别级决策树仍可能不足以描述复杂模型的局部预测行为。原因在于，同一类别的样本可能分布在多个异质区域中，单棵全局决策树往往优先拟合覆盖面较大的主导模式，而忽略局部区域中的细粒度条件组合。对于主动学习而言，这些局部模式尤其重要，因为模型错误往往并不均匀分布在整个输入空间，而是集中出现在某些特定特征区域。

为增强局部解释能力，KALE基于已有类别级候选规则构造样本的规则隶属表示。具体地，对每个样本 x 和类别级候选规则集合 R_c，定义 $M(x|R_c) = (r_1(x), r_2(x), \cdots, r_{|R_c|}(x))$，其中 $r_j(x)$ 表示样本 x 是否满足第 j 条规则的条件部分。该表示把原始特征空间中的样本转化为规则满足关系空间中的样本，使具有相似规则覆盖模式的样本更容易被划分到同一区域。随后，KALE在该隶属表示上将样本划分为 g 个局部组 $G_1, G_2, \cdots, G_g$。

对于每个局部组 G_i，KALE进一步构造局部二分类数据，并训练局部决策树以归纳更细粒度的局部规则。直观上，局部组内样本代表模型在某一规则覆盖区域内的相似预测行为，而组外样本或全局样本可作为对照区域。通过在局部范围内重新训练决策树，模型可以发现全局类别级规则难以表达的局部条件组合。这样得到的局部候选逻辑规则补充了类别级规则的不足，使解释集合既能够反映模型的主要全局模式，也能够覆盖对样本选择有价值的局部预测逻辑。

3)稀疏逻辑规则选择与解释压缩

经过类别级归纳和局部补充后，候选逻辑规则数量可能较多，直接交给专家审查会造成较高交互成本。因此，KALE需要进一步从候选集合中选择少量最能代表当前模型预测行为的逻辑规则。本文将这一过程视为稀疏特征选择问题：每条候选逻辑规则都可以看作一个二值特征，表示样本是否满足该规则条件；若少量规则的线性组合能够较好拟合模型预测结果，则这些规则可以被视为当前模型预测逻辑的紧凑解释。

具体而言，令 $r_j(x_i)$ 表示样本 x_i 是否满足第 j 条候选逻辑规则，使用带L1正则的最小二乘目标拟合当前模型关于类别 c 的预测输出，可写为：

$$\min_{w} \sum_{i} \| \sum_{j} w_j r_j(x_i) - h_c(x_i) \|_2^2 + \lambda \sum_{j} |w_j| , \qquad (1)$$

其中 $h_c(x_i)$ 表示模型 h 对样本 x_i 属于类别 c 的预测结果或预测得分，w_j 表示第 j 条逻辑规则的重要性权重，λ 为控制稀疏程度。L1正则会压缩大量不重要规则的权重，使最终解释集中保留少量对模型预测

贡献较大的规则。求解后，KALE 按照权重绝对值由大到小，为每个类别保留至多 m 条逻辑规则作为交互规则。对全部 C 个类别执行上述过程后，单轮得到的交互逻辑规则集合规模不超过 mC。若 $w_j > 0$，则规则 r_j 支持模型预测类别 c；若 $w_j < 0$，则规则 r_j 的否定条件支持模型预测类别 c；若 $w_j = 0$，则规则 r_j 和模型做出的预测类别 c 的行为无关。对所有类别执行上述过程后，即可得到最终候选逻辑规则集合 R。这里采用最小二乘目标的目的是对当前模型的类别预测得分进行稀疏后验近似，以选择最能忠实表达模型行为的少量逻辑规则。当 $h_c(x_i) \in [0,1]$ 为类别概率时，其平方误差可以视为类别级 Brier 型忠实度损失；结合 L1 正则后，目标函数保持凸性，可使用成熟的坐标下降方法高效求解，并通过权重符号直接区分支持类别 c 的规则和支持其否定条件的规则。带 L1 正则的逻辑回归也可用于规则选择，但其通过非线性链接函数拟合对数几率，更适合重新学习分类边界。本文关注的是对既有模型软输出的直接逼近与规则排序，因此采用 Lasso 作为解释压缩的工程实现。

通过上述三阶段归纳，KALE 得到的逻辑规则既来源于当前模型预测，又被压缩到专家可审查的规模。实际交互时，每条逻辑规则可以展示为“条件”“预测类别”“覆盖样本数”“模型置信度”和“规则权重”信息，专家只需判断该规则所表达的预测逻辑是否符合领域知识。若规则符合领域知识，则其覆盖样本可用于后续演绎式弱标注扩展；若规则违反领域知识，则其覆盖区域可作为知识-模型分歧区域，用于后续反绎式分歧查询。这样，模型预测逻辑的归纳不仅提供解释，也为主动学习中的知识引导查询奠定基础。

3.2 逻辑规则交互驱动的样本标注与查询

KALE 中的人机协同包含规则级反馈和样本级反馈 2 个层次。规则级反馈中，专家审核模型归纳出的“条件-预测”逻辑，并给出“正确”或“错误”的知识判断；样本级反馈中，系统根据错误逻辑规则选择最有助于纠正知识-模型分歧的样本，再由专家返回样本真实标签。正确逻辑规则不直接产生真标签，而是用于生成需要进一步筛选的弱标注候选。因此，专家既参与模型预测逻辑的审核，也参与查询样本的真实标注，2 类反馈分别影响训练数据扩展和下一轮样本选择。

具体而言，专家根据领域知识把逻辑规则集合 R 划分为 2 类：正确逻辑规则 R_{T} 和错误逻辑规则 R_{F}。正确逻辑规则表示模型在相应条件区域内的预测逻辑与领域知识一致，可用于演绎式弱标注扩展；错误逻辑规则表示模型在该区域内形成了与领域知识不一致的预测模式，可作为知识-模型分歧的符号化观测，并进一步触发反绎式分歧查询。这样的机制可以避免把不确定知识强行注入训练过程，同时保留后续轮次重新审查的可能性。

1）专家逻辑规则一致性判别

逻辑规则一致性判别是 KALE 中专家知识进入主动学习过程的核心接口。与传统逐样本标注不同，专家面对的不是孤立样本，而是模型从大量未标注样本预测中归纳出的条件-预测结构。设逻辑规则 r 由条件部分 $r_{\mathrm{condition}}$ 和预测部分 $r_{\mathrm{prediction}}$ 组成，其语义可理解为：当样本 x 满足 $r_{\mathrm{condition}}(x) = \mathrm{True}$ 时，当前模型倾向于给出预测 $r_{\mathrm{prediction}}$。专家需要判断的不是某个样本的标签，而是该条件组合是否足以支持模型给出的预测结论。

若专家认为规则条件与预测结论之间符合领域知识，例如该条件组合在任务语义上确实能够推出或强烈支持类别 $r_{\mathrm{prediction}}$，则将 r 标记为正确逻辑规则，并加入 R_{T}。若专家认为该条件组合不应推出模型给出的预测，或者该规则明显违背领域常识、业务约束或已有知识，则将 r 标记为错误逻辑规则，并加入 R_{F}。由此得到的 R_{T} 和 R_{F} 分别对应 2 类不同的知识反馈：R_{T} 提供可利用的正确预测模式，R_{F} 暴露需要纠正的错误预测模式。

2）基于正确逻辑规则的演绎式弱标注扩展

若逻辑规则 $r \in R_{\mathrm{T}}$ 被专家确认为正确，则说明在领域知识层面，满足 $r_{\mathrm{condition}}(x) = \mathrm{True}$ 的样本应具有预测标签 $r_{\mathrm{prediction}}$。基于这一判断，KALE 可以把规则覆盖样本 $Coverage(r) = \{x \in U | r_{\mathrm{condition}}(x) = \mathrm{True}\}$ 作为弱标注候选区域，并为其中样本生成候选标签 $r_{\mathrm{prediction}}$。该过程具有演绎式推理特征：给定“逻辑规则 r 正确”和“样本 x 满足规则条件”这 2 个前提，系统推出该样本的候选标签。

需要强调的是，本文将这类标签称为弱标注候选，而不是无噪声真标签。规则层面的正确判断不能保证其覆盖的每个样本均被正确标注。本文离线实验以规则覆盖样本上的隐藏真值一致率作为模拟专家判断依据，仅将一致率不低于 0.95 的规则划入正确逻辑规则集合；对其覆盖样本，还要求当前模型预测与规则结论一致后才生成弱标注。

3）基于错误逻辑规则的反绎式分歧查询

若逻辑规则 $r \in R_{\mathrm{F}}$ 被专家确认为错误，则说明当前模型在满足 $r_{\mathrm{condition}}$ 的区域内可能形成了系统性偏

差。与正确规则不同,错误规则不适合直接用于标注扩展;它的价值在于指出模型预测逻辑与领域知识发生冲突的位置。KALE把这种错误逻辑规则视为知识-模型分歧的符号化观测,并据此构造反绎式查询信号:如果某个未标注样本x满足错误规则条件,且当前模型仍按照该错误规则给出预测$r_{\text{prediction}}$,则该样本可能是解释并修正该分歧的关键样本。

为度量样本与错误逻辑规则之间的关系,KALE首先定义样本x关于错误逻辑规则集合R_{F}的不一致性分数。设$I(\cdot)$为指示函数,$r_{\text{condition}}(x)$表示样本x是否满足逻辑规则r的条件部分,$r_{\text{prediction}}$表示规则r给出的预测标签,则有:

$$Inconsistency(x|R_{\text{F}})=\sum_{\{r\in R_{\text{F}}\}}\frac{I[r_{\text{condition}}(x)\wedge h(x)=r_{\text{prediction}}]}{\sum_{\{r\in R_{\text{F}}\}}I[r_{\text{condition}}(x)]}。\tag{2}$$

当样本x不被任何错误逻辑规则覆盖时,令$Inconsistency(x|R_{\text{F}})=0$。该分数刻画的是:在所有覆盖样本$x$的错误逻辑规则中,当前模型仍沿用这些错误规则给出相同预测的比例。分数越高,说明样本x越可能位于模型错误预测逻辑集中的区域,其人工标注越可能为模型提供直接纠偏信号。

然而,仅依据不一致性分数进行查询可能偏向覆盖范围较大的错误规则,而忽略不同样本的模型预测状态。为此,KALE进一步引入模型对错误规则预测标签的概率,定义概率修正不一致分数:

$$P\text{-}I(x|R_{\text{F}})=\sum_{\{r\in R_{\text{F}}\}}\frac{I[r_{\text{condition}}(x)]\cdot\wedge h(y=r_{\text{prediction}}|x)]}{\sum_{\{r\in R_{\text{F}}\}}I[r_{\text{condition}}(x)]},\tag{3}$$

其中,$h(y=r_{\text{prediction}}|x)$表示当前模型把样本$x$预测为错误逻辑规则$r_{\text{prediction}}$的概率。当样本被多条错误逻辑规则覆盖时,*P-I*分数相当于对这些错误规则对应预测概率的归一化聚合。分数越高,说明模型不仅处于专家已否定的错误规则覆盖区域内,而且仍以较高概率支持这些错误预测。因此,选择*P-I*分数较高的样本进行人工标注,有助于优先修正模型中较稳定、较集中的错误预测模式。

最终,KALE可根据*P-I*从未标注池U中选择查询样本,并将其交由专家进行真实标注。该过程体现了反绎推理思想:系统不是从正确规则直接推出标签,而是从“专家知识与模型预测发生冲突”这一观测出发,寻找最可能解释并修正该冲突的待验证样本。通过这种方式,错误逻辑规则被转化为主动学习中的纠错查询信号,使专家知识能够直接影响下一轮样本选择。

与传统不确定性采样相比,*P-I*分数并不单纯关注模型对某个样本“是否犹豫”,而是优先关注模型是否在专家已否定的逻辑区域内仍保持错误预测倾向;与仅基于错误逻辑规则覆盖的查询相比,*P-I*分数又进一步区分了同一错误规则覆盖区域内不同样本的模型预测强度。因此,该分数同时包含知识约束和模型状态2类信息:前者保证查询样本来自知识-模型分歧区域,后者刻画当前模型对错误预测模式的依赖程度。被选中的样本经专家标注后,不仅可以纠正单个样本的标签,也可能改变模型在同类区域内的预测边界,从而提升后续轮次中逻辑规则归纳和样本查询的质量。通过这种设计,KALE把专家对错误逻辑规则的一次判断转化为面向区域性错误的查询策略,使主动学习从“寻找难判样本”进一步扩展为“寻找最有助于修正知识冲突的样本”。KALE知识引导的主动学习算法总体的过程如算法1所示。

算法 1. 知识引导主动学习方法KALE。

输入:初始标注集L^0,未标注池U^0,主动学习轮数T,单轮查询预算b,类别数C,局部分组数g,单类保留规则数上限m;

输出:最终分类模型h^T。

① for t=0, 1, …, T−1 do;
② 基于当前标注集L^t训练分类模型h^t;
③ 初始化候选逻辑规则集合$R^t=\varnothing$;
④ for c=1, 2, …, C do
⑤ 构造类别级预测数据$D_c=\{(x,I(h^{t(x)}=c))|x\in U^t\}$;
⑥ 训练决策树T_c拟合D_c,并将树节点路径转化为类别级候选逻辑规则R_c;
⑦ 计算每个样本关于R_c的规则隶属表示$M(x|R_c)$;
⑧ 基于$M(x|R_c)$将U^t划分为g个局部组$G_1, G_2, \cdots, G_g$;
⑨ for i=1, 2, …, g do
⑩ 以局部组G_i中的模型预测行为构造局部二分类数据,并训练局部决策树;
⑪ 将局部决策树节点路径转化为局部候选逻辑规则,并加入R_c;
⑫ end for
⑬ 以候选逻辑规则为二值特征,求解带L1正则的稀疏选择目标;
⑭ 选取权重绝对值最大的m条逻辑规则加入R^t;

⑮ end for

⑯ 专家审查R^t，将其划分为正确逻辑规则R_{T}^t、错误逻辑规则R_{F}^t；

⑰ 根据R_{T}^t生成演绎式弱标注候选，并将高置信弱标注样本加入L^t；

⑱ 根据R_{F}^t计算未标注样本的概率修正不一致分数$P\text{-}I(x|R_{\mathrm{F}}^t)$；

⑲ 从U^t中选择得分最高的b个样本Q_X^t交由专家标注，得到Q^t；

⑳ 更新标注集和未标注池：$L^{\{t+1\}} = L^t \cup Q^t$，$U^{\{t+1\}} = U^t \backslash Q_X^t$；

㉑ end for

㉒ 返回最终模型h^T。

算法 1 中，第⑤~⑭步对应模型预测逻辑的归纳式规则化过程。该过程首先在类别级别拟合模型关于每个类别的预测行为，再通过规则隶属表示划分局部区域，补充全局规则难以刻画的细粒度预测逻辑，最后利用 L1 稀疏选择压缩候选规则规模。第⑯~⑳步对应逻辑规则交互驱动的样本标注与查询过程：正确逻辑规则用于扩展高置信弱标注候选，错误逻辑规则用于构造知识-模型分歧信号，并通过概率修正不一致分数选择最值得专家标注的样本。整个算法把模型解释、专家知识输入和主动学习查询统一在同一轮次循环中，实现“归纳—演绎—反绎”的知识引导主动学习闭环。

3.3 交互实现与复杂度

在实际系统中，KALE 的交互粒度以逻辑规则而不是单个样本为单位。每轮主动学习结束后，系统基于当前模型生成候选逻辑规则，并为每条规则计算规则权重、覆盖样本数、模型预测置信度等辅助信息。随后，系统将各类别经稀疏选择后保留的逻辑规则汇总，形成至多mC条交互规则并展示给专家。专家对这些规则逐条给出“正确”或“错误”的判断。每条逻辑规则以“条件—预测类别—覆盖规模—置信信息”的形式呈现，专家只需判断该规则所表达的预测逻辑是否符合领域知识，并给出“正确”或“错误”的反馈。

专家反馈后，KALE 按照 2 条路径更新主动学习流程。首先，系统使用正确逻辑规则R_{T}进行演绎式弱标注扩展：对满足规则条件且模型预测与规则预测一致的高置信样本，生成弱标注候选，并在置信阈值或少量抽检约束下加入训练集。其次，系统使用错误逻辑规则R_{F}计算未标注样本的概率修正不一致分数，选择位于知识-模型分歧区域且具有较高纠错价值的样本交由专家真实标注。这样的执行顺序使正确规则主要提升标注利用率，错误规则主要提升查询针对性，从而把专家的一次规则审查转化为对多个样本的间接影响。

当一条逻辑规则具有较高覆盖度且语义能够被专家快速判断时，规则级审查可能低于逐一标注其覆盖样本的成本，但这一关系并非无条件成立。若初始模型存在严重系统性偏差，候选逻辑规则可能大量被否决，此时规则交互会形成额外成本。因此，实际系统需要同时报告规则审核数量、规则接受率和样本标注数量，而不能仅以样本标注预算衡量全部人机交互成本。为降低错误传播和局部聚焦风险，实际应用还可结合规则权重、模型置信度和覆盖规模设置筛选阈值，并在每轮保留少量随机采样样本或多样性采样样本。

复杂度方面，KALE 的额外开销主要来自候选逻辑规则生成、规则隶属矩阵计算、局部规则补充和稀疏规则选择。设未标注样本数为n_{unl}、候选逻辑规则数为$|R_{\mathrm{c}}|$，则规则隶属矩阵规模为$O(n_{\mathrm{unl}}|R_{\mathrm{c}}|)$，后续局部分组和规则覆盖计算均可在该矩阵上完成。稀疏选择阶段可使用坐标下降等成熟 Lasso 求解器，在$|R_{\mathrm{c}}|$受规则生成深度和交互预算控制时计算开销可控。正确规则的弱标注扩展和错误规则的查询打分分别与R_{T}、R_{F}的覆盖计算相关，整体可写为$O(n_{\mathrm{unl}}(|R_{\mathrm{T}}|+|R_{\mathrm{F}}|))$。由于逻辑规则抽取只在主动学习轮次之间执行，且主动学习轮数通常远小于模型训练迭代次数，该额外成本在标注成本较高的领域任务中是可以接受的。

4 实验与分析

4.1 实验设置

本文实验包含 9 个分类任务，覆盖人工逻辑数据和真实应用数据。Monk-1 和 Monk-3 来自 UCI 机器学习库[31]，用于检验逻辑规则任务；真实应用任务来自 UCI 和 OpenML-CC18（OpenML Curated Classification benchmarking suite 2018），包括金融信贷审批任务 Credit-Approval、软件缺陷预测任务 PC1 和 PC4、生物化学属性预测任务 QSAR-biodeg 和 MiceProtein、棋盘胜负判断任务 Tic-Tac-Toe 以及空气质量预测任务 Ozone-level-8hr。所有任务均随机划分 30% 实例作为测试集，其余实例作为主动学习训练池。

比较方法包括随机采样（RAN）、不确定性采样（Margin）、聚类采样（Clusters）、贝叶斯主动学习（BALD）[21]、核心集方法（CoreSet）[8]、信息核心集 [32]

以及知识驱动主动学习(KAL)[20]。所有方法使用同一主动学习流水线和同一基学习器(GBDT)[33], 并在相同损失函数、学习率和其他超参数下训练。

评价指标包括预算曲线下面积(AUBC)和最终准确率(F-ACC)。AUBC 是准确率随标注预算变化的学习曲线面积, 刻画主动学习全过程的整体收益; F-ACC 是标注预算耗尽后的最终准确率, 刻画最后一轮训练后的性能。表 2~4 中所有结果均为 10 次随机划分的平均值和标准差。

本文采用批量主动学习设置, 主动学习轮数固定为 10, 并考察 2 类标注预算: 小预算设置中, 初始标注集大小为$|L|$=30, 每轮查询预算 b=10; 中等预算设置中, 初始标注集大小为$|L|$=100, 每轮查询预算 b=50。2 类预算分别对应标注极度稀缺和标注资源相对充足的应用阶段。

本文规模化基准实验未引入真实领域专家在线参与, 而是使用主动学习训练池中对学习器隐藏的数据集真值构造模拟专家预言机。测试集标签仅用于最终性能评价, 不参与逻辑规则判别、弱标注扩展或样本查询。具体地, 在第 t 轮主动学习中, 对候选逻辑规则 r, 记其在当前未标注池上的覆盖集合为

$$C_t(r)=\{x_i\in U_t \mid r_{\text{condition}}(x_i)=\text{True}\}。$$

根据覆盖样本的隐藏真实标签, 定义规则一致率

$$A_t(r)=1/|C_t(r)|\sum_{x_i\in C_t(r)} I[y_i=r_{\text{prediction}}]。$$

当$A_t(r)\geqslant\tau_r$时, 将该规则判定为正确逻辑规则并加入R_{T}; 否则判定为错误逻辑规则并加入R_{F}。本文统一设置$\tau_r=0.95$, 覆盖集合为空的规则直接舍弃。模拟预言机只向算法返回规则类别, 不向模型开放覆盖样本的真实标签; 只有被查询策略选中的样本,

Table 2 AUBC and F-ACC on Monk-1, Monk-3, and Credit-Approval Tasks

表 2 Monk-1, Monk-3, 和 Credit-Approval 任务上的 AUBC 和 F-ACC %

方法	Monk-1				Monk-3				Credit-Approval			
	小预算		中等预算		小预算		中等预算		小预算		中等预算	
	AUBC	F-ACC	AUBC	F-ACC	AUBC	F-ACC	AUBC	F-ACC	AUBC	F-ACC	AUBC	F-ACC
RAN	93.6±4.38	98.7±1.73	99.7±0.25	99.8±0.54	96.6±2.52	96.9±2.77	98.7±0.41	**99.0±0.40**	83.2±2.27	85.3±2.27	86.7±2.52	87.4±2.81
Margin	96.0±3.35	98.9±3.25	99.8±0.19	99.8±0.36	97.2±1.37	**98.7±0.60**	**98.8±0.41**	98.9±0.47	84.2±2.38	86.8±2.38	87.0±2.38	87.3±2.37
Clusters	93.5±2.83	99.6±0.48	99.5±0.71	99.5±1.26	96.4±1.96	97.8±1.40	98.7±0.51	98.8±0.56	82.1±2.42	85.2±2.65	86.6±2.41	87.1±2.04
CoreSet	96.1±1.28	99.5±0.75	99.8±0.25	99.8±0.36	97.4±1.36	98.0±1.25	98.6±0.45	98.9±0.47	83.0±2.43	86.3±2.43	86.5±2.57	87.2±2.42
BALD	95.4±1.56	98.4±2.02	99.8±0.28	99.8±0.36	96.9±1.47	98.2±1.21	98.7±0.41	98.9±0.47	83.0±3.33	84.9±3.33	86.6±2.64	87.2±2.42
InCoreSet	96.8±3.57	98.9±3.25	99.8±0.24	99.8±0.36	97.7±1.23	98.3±0.96	**98.8±0.45**	98.9±0.47	84.5±3.08	85.6±3.08	87.1±2.40	87.1±2.53
KAL	96.0±3.35	98.9±3.25	99.8±0.19	99.8±0.36	97.2±1.37	98.7±0.60	98.8±0.41	98.9±0.47	84.2±0.30	86.8±0.23	87.0±2.38	87.3±2.37
KALE（本文）	**98.7±0.78**	**99.8±0.36**	**99.9±0.09**	**100±0.00**	**98.2±1.01**	**98.4±1.08**	**98.8±0.40**	**98.9±0.52**	**85.3±1.96**	**87.2±2.55**	**87.0±2.50**	**87.3±2.41**

注：小预算的$|L|$=30，b=10；中等预算的$|L|$=100，b=50。数值格式为平均值±标准差，黑体数值标记最优或并列最优结果。

Table 3 AUBC and F-ACC on PC1, PC4, and Tic-Tac-Toe Tasks

表 3 PC1, PC4, 和 Tic-Tac-Toe 任务上的 AUBC 和 F-ACC %

方法	PC1				PC4				Tic-Tac-Toe			
	小预算		中等预算		小预算		中等预算		小预算		中等预算	
	AUBC	F-ACC	AUBC	F-ACC	AUBC	F-ACC	AUBC	F-ACC	AUBC	F-ACC	AUBC	F-ACC
RAN	83.6±3.60	89.0±1.72	90.4±0.76	93.0±0.01	83.9±1.36	87.7±1.62	88.5±0.74	90.1±0.94	74.0±1.84	82.4±1.76	92.5±1.20	97.2±1.06
Margin	81.5±6.68	90.2±3.92	91.5±0.96	93.2±1.38	84.1±2.36	88.6±2.02	89.7±0.79	91.0±0.94	74.8±2.52	88.0±3.36	95.0±1.12	97.6±0.57
Clusters	79.6±3.20	84.1±1.92	89.6±1.03	92.3±1.17	83.2±2.38	85.6±2.31	88.1±1.15	89.9±1.18	70.0±2.45	76.3±2.40	91.9±1.63	97.3±1.10
CoreSet	77.6±7.80	83.7±5.02	90.3±1.08	93.2±1.35	82.1±2.34	87.7±1.43	89.5±0.97	**91.5±1.09**	77.0±2.53	86.8±3.60	93.1±1.56	97.5±0.77
BALD	83.7±4.28	89.5±2.75	90.2±1.45	93.2±1.35	84.2±1.80	87.5±1.85	88.2±0.83	89.8±0.90	73.9±2.18	82.1±3.91	91.0±1.73	97.3±1.01
InCoreSet	83.0±4.94	88.7±3.27	91.5±0.99	93.1±1.37	85.0±1.77	89.3±1.67	89.9±0.69	90.8±1.03	75.0±3.12	84.1±4.08	93.1±1.72	**97.8±0.82**
KAL	81.4±6.78	90.6±3.15	91.5±0.96	93.2±1.38	84.0±2.39	88.6±1.83	89.7±0.79	91.0±0.94	74.8±2.52	88.0±3.36	95.0±1.12	97.6±0.57
KALE	**91.4±1.47**	**93.0±1.27**	**92.5±1.25**	**97.6±0.83**	**88.4±1.35**	**90.0±0.57**	**90.1±0.58**	**90.8±0.89**	**78.6±2.66**	**93.0±3.35**	**95.4±1.02**	**97.6±0.83**

注：小预算的$|L|$=30，b=10；中等预算的$|L|$=100，b=50。数值格式为平均值±标准差，黑体数值标记最优或并列最优结果。

Table 4 AUBC and F-ACC on QSAR-biodeg, Protein, and Ozone-level-8hr Tasks

表 4 QSAR-biodeg、Protein 和 Ozone-level-8hr 任务上的 AUBC 和 F-ACC %

方法	QSAR-biodeg				Mice Protein				Ozone-level-8hr			
	小预算		中等预算		小预算		中等预算		小预算		中等预算	
	AUBC	F-ACC	AUBC	F-ACC	AUBC	F-ACC	AUBC	F-ACC	AUBC	F-ACC	AUBC	F-ACC
RAN	76.5±2.29	81.4±2.19	83.9±1.07	86.1±1.22	64.7±1.82	79.7±2.88	91.4±1.43	95.7±1.53	83.7±2.52	90.3±1.49	90.6±1.13	93.1±0.72
Margin	76.3±1.69	81.4±2.42	85.1±1.45	86.5±1.63	62.4±2.63	81.8±3.25	94.2±1.39	97.1±2.10	85.9±3.21	93.2±1.12	92.3±0.81	**93.9±0.91**
Clusters	74.6±3.60	78.5±3.33	76.6±2.17	81.7±2.76	58.6±4.17	72.0±5.21	89.3±1.90	92.9±2.56	81.6±2.83	87.1±2.41	89.8±1.10	92.1±1.24
CoreSet	74.6±4.32	77.3±4.11	84.0±1.60	86.4±1.56	71.7±2.18	89.2±2.26	94.6±1.45	97.0±2.16	75.9±4.29	79.4±3.96	84.7±2.82	87.8±2.39
BALD	76.5±2.88	82.0±2.04	84.1±1.42	85.4±1.60	66.5±3.41	83.7±3.65	92.9±1.63	97.0±2.16	85.4±2.99	92.6±1.22	90.7±1.10	93.0±1.06
InCoreSet	75.3±2.40	79.0±2.85	84.0±1.40	85.9±1.76	70.9±2.01	90.0±4.79	94.3±1.48	**97.2±2.13**	84.5±2.76	92.9±1.35	92.1±0.90	**93.9±0.83**
KAL	76.6±2.17	81.7±2.76	85.1±1.45	86.5±1.63	60.9±3.18	81.1±6.82	94.1±1.54	97.0±2.07	85.8±3.20	93.4±1.20	92.3±0.81	**93.9±0.91**
KALE	**77.9±2.44**	**82.8±1.49**	**85.2±1.52**	**86.6±1.49**	**74.9±4.04**	**94.9±2.74**	**94.7±0.11**	**97.1±1.40**	**92.1±0.58**	**93.6±0.67**	**93.1±0.75**	**93.9±0.76**

注：小预算的$|L|$=30，b=10；中等预算的$|L|$=100，b=50。数值格式为平均值±标准差，黑体数值标记最优或并列最优结果。

其隐藏真值才作为专家标注结果加入标注集。规则判断次数与样本标注预算 b 分开计量。每轮主动学习中，系统针对 C 个类别分别保留至多 m 条经稀疏选择的逻辑规则，形成至多 mC 条交互规则；模拟预言机对这些规则逐条返回正确或错误判断。因此，单次 10 轮主动学习实验最多产生 $10mC$ 次规则级判断。规则级反馈用于判断模型预测逻辑是否与领域知识一致，样本级反馈用于返回查询样本的真实标签。表 2~4 给出了相同样本标签预算下的性能比较。

4.2 主要结果

表 2~4 完整列出了 KALE 与 7 类代表性主动学习基线在 9 个任务、2 类标注预算和 2 个评价指标上的结果。AUBC 反映的是 10 轮主动学习全过程中准确率随标注预算增加的整体表现，而 F-ACC 反映预算耗尽后的最终准确率。整体来看，KALE 在 18 个数据集-预算组合的 AUBC 指标上均达到最优或并列最优水平，说明其优势不是来自某一个 F-ACC 点，而是贯穿主动学习迭代过程。这一点对主动学习尤为重要，因为真实标注过程往往受到预算、时间和专家可用性的动态限制，能够在早期和中期持续提升模型的策略更具实际价值。

与随机采样 RAN 相比，KALE 在多数任务上取得明显提升，尤其在小预算设置下更为突出。例如，在 PC1 小预算设置下，RAN 的 AUBC 为 83.6，KALE 达到 91.4；在 Mice Protein 小预算设置下，RAN 的 AUBC 为 64.7，KALE 提升到 74.9；在 Ozone-level-8hr 小预算设置下，RAN 的 AUBC 为 83.7，KALE 达到 92.1。这些结果表明，当初始标注样本较少、模型预测逻辑尚不稳定时，KALE 能够通过逻辑规则交互更快定位模型存在系统性偏差的区域，从而提高有限标注预算的使用效率。

不同基线方法的表现体现出主动学习策略的跨任务不稳定性。Margin 和 KAL 整体相对稳定，但前者主要依赖模型不确定性，后者依赖预设知识关系；CoreSet、BALD 和 InCoreSet 在部分任务上表现较好，但在另一些任务中退化明显。例如，CoreSet 在 PC1 小预算下 AUBC 为 77.6，低于 RAN 的 83.6 和 KALE 的 91.4；在 Ozone-level-8hr 小预算下，CoreSet 的 AUBC 为 75.9，也低于 RAN 的 83.7。BALD 在 Mice Protein 等任务上有一定优势，但在 Tic-Tac-Toe 和 PC4 等任务中表现并不稳定。这说明，仅依赖多样性、模型分歧或核心集覆盖，并不能保证在不同任务和预算阶段都选到真正有助于纠正模型错误的样本。

KALE 的稳定性来自其知识引导机制。该方法并不把领域知识作为静态规则库直接套用，而是先从当前模型预测行为中归纳出逻辑规则，再由专家判断这些规则是否符合领域知识。正确逻辑规则用于演绎式弱标注扩展，错误逻辑规则用于反绎式分歧查询。这样，专家知识被转化为可解释、可操作的区域级样本选择信号，使主动学习不再只依赖学习器内部状态，而能够围绕知识-模型分歧进行针对性纠偏。上述结果表明，逻辑规则交互能够在部分任务中把模拟专家判断转化为区域级样本选择信号，使查询过程同时参考学习器状态和知识-模型分歧。然而，这一机制的收益仍会受到初始模型质量、规则解释忠实度和标注预算的影响，不能据此推断 KALE 在所有任务和预算下均稳定优于其他方法。

从真实应用任务看，KALE 的优势更为明显。PC1

和 PC4 属于软件缺陷预测任务，模型容易受到少量初始样本和局部特征组合影响。KALE 在 PC1 小预算下 AUBC 达到 91.4，显著高于 Margin 的 81.5、KAL 的 81.4 和 BALD 的 83.7；在 PC1 中等预算下 F-ACC 达到 97.6，也明显高于多数基线。Mice Protein 任务中，KALE 在小预算下 F-ACC 达到 94.9，而 RAN 为 79.7，KAL 为 81.1，InCoreSet 为 90.0，说明逻辑规则交互能够在标注极少时显著改善最终模型性能。Ozone-level-8hr 任务中，KALE 小预算 AUBC 达到 92.1，高于 Margin 的 85.9 和 KAL 的 85.8，表明在特征关系复杂、初始模型偏差较强的任务中，错误逻辑规则驱动的查询能够提供有效纠偏信号。

在逻辑结构较强的任务上，KALE 同样表现出优势。Monk-1 和 Monk-3 本身具有较明确的逻辑结构，多数方法在中等预算下已经接近饱和，因此 F-ACC 差异不大。即便如此，KALE 在 Monk-1 小预算下的 AUBC 达到 98.7，高于 InCoreSet 的 96.8 和 CoreSet 的 96.1，说明在规则结构较清晰的任务中，KALE 能够更快利用专家确认的逻辑规则提升学习效率。Tic-Tac-Toe 任务也体现了这一点：KALE 小预算的 F-ACC 达到 93.0，高于 Margin 和 KAL 的 88.0，也高于 CoreSet 的 86.8，说明逻辑规则不仅具有解释作用，也能实际转化为有效的样本选择依据。

从 F-ACC 看，KALE 在多数设置中达到最高或接近最高结果，但也存在少数 F-ACC 不占优的情形。例如，PC4 中等预算下 CoreSet 的 F-ACC 为 91.5，高于 KALE 的 90.8；Tic-Tac-Toe 中等预算下 InCoreSet 的 F-ACC 达到 97.8，KALE 为 97.6；Mice Protein 中等预算下 InCoreSet 的 F-ACC 为 97.2，KALE 为 97.1。这些差异较小，且不少情况下标准差范围存在重叠，说明不同方法在预算充分后可能收敛到相近的最终性能。相比之下，AUBC 更能反映主动学习过程中每一轮标注样本的实际利用效率。KALE 在这些设置下通常仍保持 AUBC 优势或接近最优，说明其主要贡献不只是追求最后一轮的单点 F-ACC，而是在预算逐步消耗过程中更早、更稳定地提升模型。对于真实应用中可能随时受限的专家标注流程而言，这种全过程收益比单纯最终精度更具实践意义。

综合来看，表 2~4 支持本文的核心判断：传统主动学习策略若只从学习器内部状态或数据分布出发，容易受到任务、模型和预算阶段影响；KALE 通过“归纳式规则化—专家逻辑规则判别—演绎式弱标注扩展—反绎式分歧查询”的闭环，把领域知识转化为可解释的人机协同信号。正确逻辑规则提高标注利用率，使专家一次规则判断能够影响多个候选样本；错误逻辑规则则帮助定位知识-模型分歧区域，使查询过程更集中于模型需要纠偏的位置。2 类规则反馈共同作用，使 KALE 在小预算和跨任务设置下获得更稳定的性能提升。这也说明，本文中的可解释人机协同不是附加的解释展示模块，而是直接参与主动学习样本选择和训练数据扩展的关键机制，从而把专家知识、模型解释和数据选择统一到同一个主动学习闭环中。

4.3 跨任务改进概率分析

主结果能够说明不同方法在已给定实验任务上的平均性能差异，但主动学习方法在实际部署中还需要面对另一个更关键的问题：当任务、数据分布和标注预算发生变化时，方法是否仍然能够稳定优于简单而稳健的随机采样。已有主动学习经验研究表明，一些主动学习策略虽然在个别数据集上取得较高结果，但其性能增益在不同任务之间并不一致，甚至可能低于随机标注基线。由于主动学习中的标注成本通常不可回收，一旦选择了不合适的查询策略，已经消耗的专家标注预算难以重新利用。因此，仅报告若干数据集上的平均准确率还不足以刻画方法的实用价值，还需要进一步评估其跨任务泛化稳健性。

为此，本文在 AUBC 指标上引入改进概率（probability of improvement）来衡量一种主动学习方法在未知数据集或未知任务上超过基线方法的可能性。与只比较均值或 F-ACC 不同，改进概率关注的是“方法 X 是否更可能优于方法 Y”这一排序关系，因而更适合用于多任务、多随机划分条件下的方法稳健性比较。设 X 和 Y 分别表示 2 个主动学习方法的标量性能指标，且指标越大表示性能越好。记 X_d 和 Y_d 为二者在数据集 d 上的性能结果，假设在该数据集上分别观测到 n_x 个 X_d 样本和 n_y 个 Y_d 样本，则可用 Mann-Whitney U 统计量[34]估计方法 X 在数据集 d 上优于方法 Y 的概率：

$$P(X_d > Y_d) = \frac{1}{n_x}\frac{1}{n_y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y} S(x_{d,i}, y_{d,j}),$$

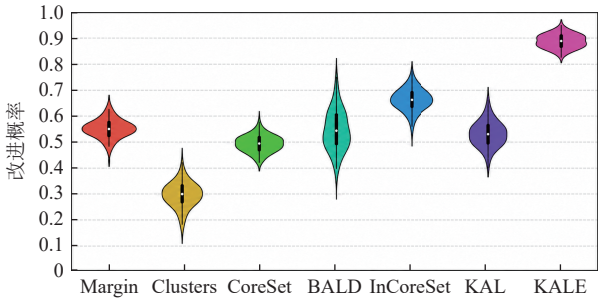
$$S(x,y) = \begin{cases} 1, & y < x, \\ \dfrac{1}{2}, & y = x, \\ 0, & y > x。 \end{cases}$$

$$P(X \text{ outperforms } Y) = \frac{1}{D}\sum_{d=1}^{D} P(X_d > Y_d),$$

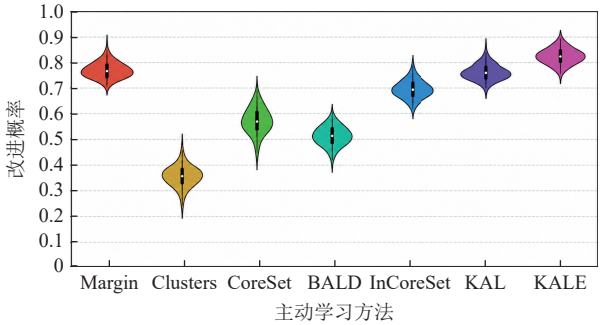
其中，$x_{\{d,i\}}$ 表示方法 X 在数据集 d 上第 i 次随机划分

或重复实验中的 AUBC 结果，$y_{\{d,j\}}$表示方法 Y 对应的 AUBC 结果，D 表示参与统计的数据集数量，函数 $S(x,y)$用于比较一对实验结果。由此得到的 $P(X_d > Y_d)$ 可以理解为在数据集 d 上随机抽取一次 X 的实验结果和一次 Y 的实验结果时，X 优于 Y 的概率；进一步对所有数据集求平均，得到方法 X 在跨任务条件下优于方法 Y 的总体改进概率。

具体实现时，本文以随机采样 RAN 作为参照基线，估计各主动学习方法在 AUBC 指标上超过 RAN 的概率。为获得更稳定的分布估计，本文对每个数据集上的$X_{\{d,1:\,n_x\}}$和$Y_{\{d,1:\,n_y\}}$进行分层 bootstrap 重采样，每个数据集重复重采样 200 次，并绘制 Mann-Whitney U 统计量的 bootstrap 分布。根据 Neyman-Pearson 统计检验准则，当置信区间上界高于 0.75 阈值时，可认为该方法相对于基线具有统计意义上的改进倾向。图 2 分别给出了小预算和中等预算设置下，不同主动学习方法相对于 RAN 基线的 AUBC 改进概率分布。



(a) 小预算下的改进概率分析



(b) 中等预算下的改进概率分析

Fig. 2 Probability of improvement over random sampling under small and medium budgets

图 2 小预算和中等预算下相对随机采样的改进概率

从图 2 可以看出，KALE 在 2 类预算设置下均具有最高的改进概率，说明其不仅在表 2~4 所列任务上取得较优 AUBC 结果，而且在跨任务比较中更可能稳定超过随机采样。尤其在小预算设置下，主动学习方法的价值最为突出，因为此时每一轮查询样本的选择都会显著影响后续模型训练方向。许多传统方法的改进概率分布集中在较低区间，其四分位范围未能达到 0.75 阈值，表明这些方法在未知任务上并不总能稳定优于随机采样。相比之下，KALE 的改进概率整体明显高于 0.50，且四分位区间高于 0.75，说明其在标注极为稀缺时仍具有较强的稳定收益。

需要指出的是，图 2 中的改进概率以 AUBC 为评价指标、以随机采样 RAN 为参照，反映的是一种方法在完整标注过程中取得更高预算曲线面积的可能性，而不等同于其 F-ACC 在所有任务上均优于其他基线。结合表 2~4 可以看到，在部分中等预算任务中，KALE 的 F-ACC 与 CoreSet 或 InCoreSet 已经较为接近，甚至略低。这说明随着标注数据增加，基学习器逐渐收敛，逻辑规则所提供的信息与已有标注产生更多重叠，其边际收益会相应减弱。因此，图 2 主要支持 KALE 在跨任务全过程的学习效率，特别是在标注稀缺阶段具有较稳定的改进倾向，而不意味着其在任意任务、指标和预算阶段均持续领先。

这一结果与本文的方法设计是一致的。传统不确定性、代表性或分歧驱动的查询策略主要从学习器自身状态出发，其有效性容易受到模型校准程度、初始标注集分布、任务特征空间和预算规模的影响。当模型在早期阶段尚未形成可靠决策边界时，单纯依赖学习器内部信号可能放大局部偏差。KALE 则通过逻辑规则把模型预测行为显式呈现给领域专家，并将专家的判断转化为 2 类互补信号：正确逻辑规则用于演绎式弱标注扩展，错误逻辑规则用于反绎式分歧查询。因此，样本选择不再只依赖当前模型的不确定性或几何覆盖关系，而是同时受到领域知识约束和知识-模型分歧信号的引导。

在中等预算设置下，各方法整体改进概率相比小预算有所提升，说明更多初始标注和更大的查询批量能够缓解部分策略的不稳定性。然而，KALE 仍保持最高或接近最高的改进概率分布，表明其优势并非仅来自小样本偶然增益，而是来自逻辑规则交互机制对主动学习过程的持续约束。换言之，随着标注预算增加，KALE 既能利用正确逻辑规则扩大可信训练信息，又能持续通过错误逻辑规则定位模型尚未修正的系统性偏差，从而在不同预算范围内保持较好的泛化稳健性。综上，改进概率分析从跨任务统计角度进一步验证了 KALE 的有效性。与只在个别数据集上取得较好结果的主动学习策略相比，KALE 更强调在未知任务中的稳定收益。这一点对面向可解释人机协同的主动学习尤为重要：在真实

应用中，用户往往无法预先知道某一查询准则是否适合当前任务，而 KALE 通过“模型归纳逻辑规则—专家审核逻辑规则—系统利用规则反馈更新查询”的闭环，把领域知识引入样本选择过程，从而降低了主动学习策略选择对具体任务和预算设置的敏感性。

4.4 逻辑规则归纳分析

逻辑规则交互能否有效引导主动学习，首先取决于模型解释模块是否能够以有限数量的逻辑规则忠实刻画当前模型的预测行为。如果抽取出的逻辑规则与模型行为不一致，专家对逻辑规则正确性的判断就难以转化为可靠的样本选择信号；如果逻辑规则数量过多、条件组合过碎，专家审查成本就会抵消主动学习减少标注的收益。因此，本节围绕逻辑规则解释能力展开辅助实验，考察 KALE 中归纳式规则化模块能否在解释忠实性、覆盖能力和交互成本之间取得平衡，并进一步分析该模块对可解释人机协同闭环的支撑作用。

对比方法选择代表性的规则型可解释学习方法。BoostedRule[35] 和 SLIPPER[36] 通过 boosting 构造逻辑规则集合，重点提高困难样本上的预测精度；RuleFit 将树模型生成的逻辑规则作为线性模型特征，可以视为不包含局部预测逻辑补充的规则化基线。与这些方法相比，KALE 的目标并不是单纯训练一个规则分类器，而是在保持模型预测忠实性的同时，生成规模受控、结构清晰、便于专家审核的交互逻辑规则。换言之，KALE 更关注逻辑规则能否作为模型与专家之间的中间表示，而不仅是作为最终分类器本身。

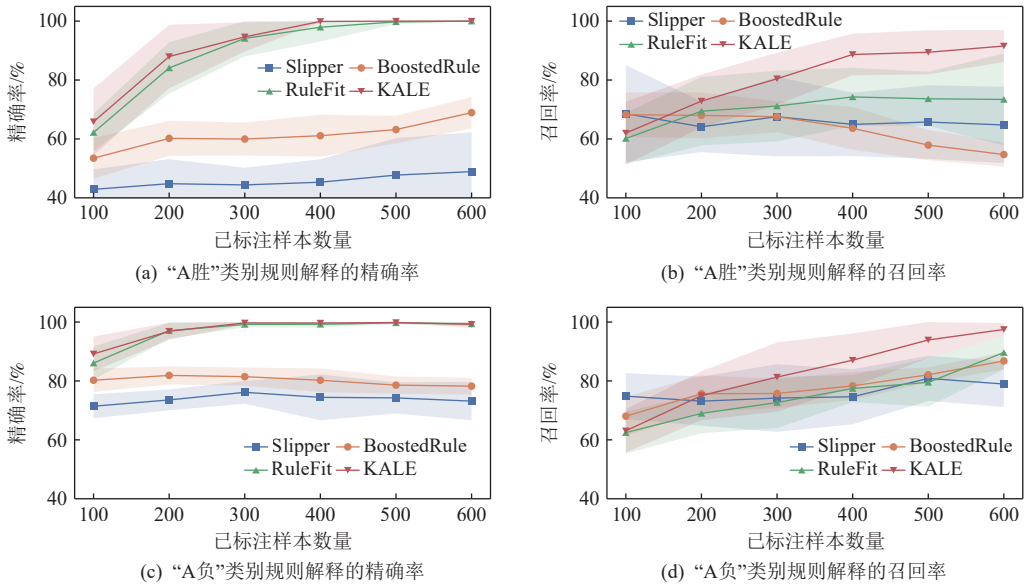
实验在 Tic-Tac-Toe 任务上进行。该任务具有清晰的棋盘状态语义和可理解的胜负条件，适合检验逻辑规则是否能够表达模型关于“A 胜”和“A 负”的预测逻辑。在不同标注预算下，各方法分别从当前模型或训练数据中生成逻辑规则，并以当前模型预测作为参照，计算规则预测相对于模型预测的精确率(*Precision*)和召回率(*Recall*)。对于类别 c，记 TP_c 为规则预测为 c 且模型预测也为 c 的样本数，FP_c 为规则预测为 c 但模型预测不为 c 的样本数，FN_c 为模型预测为 c 但未被相应规则覆盖的样本数，则有：

$$Precision_c = \frac{TP_c}{TP_c + FP_c},$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c},$$

其中，精确率反映逻辑规则给出的类别判断是否与模型预测保持一致，召回率反映逻辑规则能够覆盖多少模型在该类别上的预测区域。二者分别对应解释的可靠性和覆盖性：前者关系到专家是否可以信任规则所揭示的模型逻辑，后者关系到系统能否充分暴露模型中可被专家审查的预测模式。

图 3 给出了不同标注数量下“A 胜”和“A 负”两类预测的规则精确率与召回率。可以看到，KALE 在 2 类预测上均优于 RuleFit、Slipper 和 BoostedRule 等规则解释基线，并且在标注样本较少时仍能保持较高精确率。该结果说明，类别级逻辑规则能够捕捉模型在主要类别上的稳定模式，局部逻辑规则补充能够覆盖全局规则难以表达的局部预测区域，而稀疏选择机制则限制了最终展示给专家的逻辑规则数



(a) “A胜”类别规则解释的精确率　(b) “A胜”类别规则解释的召回率

(c) “A负”类别规则解释的精确率　(d) “A负”类别规则解释的召回率

Fig. 3 Precision and recall of rule explanations on Tic-Tac-Toe tasks

图 3 Tic-Tac-Toe 任务上的规则解释精确率和召回率

量，使解释结果在忠实性和交互成本之间取得平衡。

从本文“归纳-演绎-反绎”的方法主线看，逻辑规则解释能力不是附加的可视化指标，而是整个人机协同机制成立的前提。较高精确率意味着专家看到的逻辑规则更接近模型真实预测逻辑，因而能够更可靠地区分正确逻辑规则和错误逻辑规则；较高召回率意味着系统能够覆盖更多模型预测区域，从而既发现可用于演绎式弱标注扩展的可信模式，也暴露可用于反绎式分歧查询的知识-模型冲突。换言之，规则解释质量直接影响专家反馈能否从“解释阅读”转化为“样本选择”，并决定领域知识能否真正进入主动学习循环。本实验还表明，逻辑规则解释能力需要同时兼顾忠实性、覆盖性和交互可用性。若逻辑规则过少，系统可能只能呈现模型最显著的全局预测模式，遗漏对样本选择具有价值的局部错误区域；若逻辑规则过多或条件组合过碎，则会增加专家审查负担，使逻辑规则交互接近传统规则工程过程。因而，在主动学习场景中，解释模块不能只追求规则分类精度，还需要生成数量受控、语义清晰且能够覆盖关键预测区域的逻辑规则。KALE 通过类别级规则归纳、局部规则补充和稀疏规则选择，将模型预测行为压缩为适合专家审查的逻辑规则集合，从而为后续演绎式弱标注扩展和反绎式分歧查询提供更可靠的知识接口。

5 局限性讨论

KALE 的有效性受到初始模型质量和标注预算阶段的影响。在标注稀缺阶段，逻辑规则能够补充有限样本尚未充分表达的区域信息，知识反馈的作用较为明显；随着标注数据增加，基学习器的决策边界逐渐稳定，逻辑规则覆盖的信息与已有标注数据产生更多重叠，其边际收益也相应减弱。另一方面，如果初始模型存在严重的系统性偏差，由其预测行为归纳出的候选逻辑规则也可能大量无效。专家否决虽然可以阻止错误逻辑规则进入弱标注分支，但无法发现模型尚未表达或规则抽取未覆盖的偏差区域。因此，逻辑规则交互是对学习器视角的补充，而非独立于学习器的无偏知识来源，也不能保证在标注预算充足时突破基学习器自身的性能上限。

演绎式弱标注扩展还可能带来错误传播风险。本文仅接受覆盖样本真实标签一致率不低于 0.95 的逻辑规则，并要求模型预测与逻辑规则结论一致后再生成弱标注，以降低噪声进入训练集的可能性。然而，规则层面的高一致率并不意味着其覆盖的每个样本都能被正确标注。当逻辑规则覆盖范围较大、不同规则相互冲突或弱标注在训练集中占比较高时，少量错误仍可能在后续迭代中被放大。实际部署时，可以限制单轮弱标注扩展规模，剔除被冲突规则覆盖的样本，并通过少量人工抽检动态评估弱标注质量。由于本文未开展专门的弱标注噪声敏感性实验，相关结论仍应限定在当前严格的模拟专家协议下。

在方法适用范围方面，KALE 的规则归纳接口原则上不依赖特定类型的基学习器。决策树在本文中用于拟合当前模型的预测输出，而不是替代基学习器完成分类训练，因此线性模型、神经网络和其他黑箱分类器均可采用相同思路归纳代理逻辑规则。但是，逻辑规则能否成为有效的人机协同媒介，还取决于输入特征是否具有可理解的语义。对于视觉、自然语言或深度潜在表征，直接生成的逻辑规则可能难以被领域专家理解，需要进一步结合概念级表征、多模态解释或大语言模型辅助规则生成。

6 结　论

本文围绕主动学习在标注成本高、查询策略跨任务不稳定以及专家知识难以有效进入学习过程等问题，提出面向可解释人机协同的知识引导主动学习方法 KALE。该方法以逻辑规则作为模型与领域专家之间的交互媒介，首先从当前模型在未标注样本上的预测行为中归纳类别级和局部预测逻辑，并通过稀疏选择压缩为便于专家审核的逻辑规则集合；随后依据专家对逻辑规则正确性的判别，将正确逻辑规则用于演绎式弱标注扩展，将错误逻辑规则视为知识-模型分歧的符号化观测并用于反绎式分歧查询。由此，KALE 形成了由归纳式规则化、专家逻辑规则判别、演绎式弱标注扩展和反绎式分歧查询组成的闭环，使专家反馈不再停留于单个样本标签，而能够以可解释方式同时影响训练数据扩展和后续查询策略。实验结果表明，在多个分类任务和不同标注预算下，KALE 在 AUBC 和 F-ACC 上整体优于或接近代表性主动学习基线，尤其在小预算场景中能够更早提升模型性能；跨任务改进概率分析进一步说明，该方法相对于随机采样具有更稳定的收益；逻辑规则归纳实验则表明，KALE 能够在解释忠实性、覆盖能力和交互成本之间取得较好平衡，为可解释人机协同提供有效知识接口。

总体而言，本文工作说明，将模型预测逻辑显式

规则化并引入专家判别，可以把领域知识、模型解释和样本选择统一到主动学习过程之中。未来还可进一步研究更复杂数据类型上的逻辑规则表达、多模态或大语言模型辅助规则生成，以及面向真实专家交互场景的成本建模与自适应规则展示机制。

作者贡献声明：邵杰晶负责整体方案设计与实现，并撰写论文；杨铭和杨骁文负责提供部分算法思路；李宇峰提出指导建议并修改论文。

参 考 文 献

[1] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436−444

[2] Portugal I, Alencar P S C, Cowan D D. The use of machine learning algorithms in recommender systems: A systematic review[J]. Expert Systems with Applications, 2018, 97: 205−227

[3] Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare[J]. Nature Medicine, 2019, 25(1): 24−29

[4] Lewis D D, Gale W A. A sequential algorithm for training text classifiers[C]//Proc of the 17th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 1994: 3−12

[5] Tong S, Koller D. Support vector machine active learning with applications to text classification[J]. Journal of Machine Learning Research, 2001, 2: 45−66

[6] Dasgupta S, Hsu D J. Hierarchical sampling for active learning[C]// Proc of the 25th Int Conf on Machine Learning. New York: ACM, 2008: 208−215

[7] Brinker K. Incorporating diversity in active learning with support vector machines[C]//Proc of the 20th Int Conf on Machine Learning. Menlo Park, CA: AAAI, 2003: 59−66

[8] Sener O, Savarese S. Active learning for convolutional neural networks: A core-set approach[C/OL]//Proc of the 6th Int Conf on Learning Representations. OpenReview, 2018[2026-07-15]. https://openreview.net/forum?id=H1aIuk-RW

[9] Seung H S, Opper M, Sompolinsky H. Query by committee[C]//Proc of the 5th Annual Workshop on Computational Learning Theory. New York: ACM, 1992: 287−294

[10] Zhan Xueying, Liu Huan, Li Qing, et al. A comparative survey: Benchmarking for pool-based active learning[C]//Proc of the 30th Int Joint Conf on Artificial Intelligence. Marina del Rey, CA: IJCAI Organization, 2021: 4679−4686

[11] Bahri D, Jiang H, Schuster T, et al. Is margin all you need? An extensive empirical study of active learning on tabular data[J/OL]. CoRR, 2022, abs/2210.03822[2026-06-04]. https://doi.org/10.48550/arXiv.2210.03822

[12] Hacohen G, Dekel A, Weinshall D. Active learning on a budget: Opposite strategies suit high and low budgets[C]//Proc of the 39th Int Conf on Machine Learning. Cambridge, MA: PMLR, 2022: 8175−8195

[13] Zhou Zhi-Hua. Abductive learning: Towards bridging machine learning and logical reasoning[J]. Science China Information Sciences, 2019, 62(7): 076101

[14] Dai Wangzhou, Xu Qiuling, Yu Yang, et al. Bridging machine learning and logical reasoning by abductive learning[C]//Advances in Neural Information Processing Systems 32. Red Hook, NY: Curran Associates, Inc., 2019: 2811−2822

[15] Fischer M, Balunovic M, Drachsler-Cohen D, et al. DL2: Training and querying neural networks with logic[C]//Proc of the 36th Int Conf on Machine Learning. Cambridge, MA: PMLR, 2019: 1931−1941

[16] Xu Jingyi, Zhang Zilu, Friedman T, et al. A semantic loss function for deep learning with symbolic knowledge[C]//Proc of the 35th Int Conf on Machine Learning. Cambridge, MA: PMLR, 2018: 5502−5511

[17] Yang Zhun, Lee J, Park C. Injecting logical constraints into neural networks via straight-through estimators[C]//Proc of the 39th Int Conf on Machine Learning. Cambridge, MA: PMLR, 2022: 25096−25122

[18] Badreddine S, d'Avila Garcez A S, Serafini L, et al. Logic tensor networks[J]. Artificial Intelligence, 2022, 303: 103649

[19] Yang Zhun, Ishay A, Lee J. NeurASP: Embracing neural networks into answer set programming[C]//Proc of the 29th Int Joint Conf on Artificial Intelligence. Marina del Rey, CA: IJCAI Organization, 2020: 1755−1762

[20] Ciravegna G, Precioso F, Betti A, et al. Knowledge-driven active learning[J/OL]. arXiv, 2021, abs/2110.08265[2026-06-04]. https://arxiv.org/abs/2110.08265

[21] Gal Y, Islam R, Ghahramani Z. Deep Bayesian active learning with image data[C]//Proc of the 34th Int Conf on Machine Learning. Cambridge, MA: PMLR, 2017: 1183−1192

[22] Liu Mingyang, Wang Ruomei, Zhou Fan, et al. Video question answering scheme base on multimodal knowledge active learning[J]. Journal of Computer Research and Development, 2024, 61(4): 889−902 (in Chinese)
(刘明阳，王若梅，周凡，等. 基于多模态知识主动学习的视频问答方案[J]. 计算机研究与发展，2024，61(4)：889−902)

[23] Ding Zixin, Chen Si, Jia Ruoxi, et al. Learning to rank for active learning via multi-task bilevel optimization[C]//Proc of the 40th Conf on Uncertainty in Artificial Intelligence. Cambridge, MA: PMLR, 2024: 1112−1128

[24] Werner T, Burchert J, Stubbemann M, et al. A cross-domain benchmark for active learning[C]//Advances in Neural Information Processing Systems 37. Red Hook, NY: Curran Associates, Inc., 2024: 62875−62911

[25] Bayer M, Lutz J, Reuter C. ActiveLLM: Large language model-based active learning for textual few-shot scenarios[J]. Transactions of the Association for Computational Linguistics, 2026, 14: 1−22

[26] Ribeiro M T, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier[C]//Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1135−1144

[27] Lundberg S M, Lee S I. A unified approach to interpreting model predictions[C]//Advances in Neural Information Processing Systems 30. Red Hook, NY: Curran Associates, Inc., 2017: 4765−4774

[28] Ribeiro M T, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations[C]//Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 1527−1535

[29] Friedman J H, Popescu B E. Predictive learning via rule ensembles[J]. The Annals of Applied Statistics, 2008, 2(3): 916−954

[30] Pei Zhongyi, Liu Lin, Wang Chen, et al. An explainability-centric requirements analysis framework for machine learning applications[J]. Journal of Computer Research and Development, 2024, 61(4): 983−1002 (in Chinese)
(裴忠一, 刘璘, 王晨, 等. 面向机器学习应用的可解释性需求分析框架[J]. 计算机研究与发展, 2024, 61(4): 983−1002)

[31] Dua D, Graff C. UCI machine learning repository[DB/OL]. Irvine: University of California, 2017[2026-06-04]. http://archive.ics.uci.edu/ml

[32] Shao Jiejing, Xu Yunlu, Cheng Zhanzhan, et al. Active model adaptation under unknown shift[C]//Proc of the 28th ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2022: 1558−1566

[33] Friedman J H. Greedy function approximation: A gradient boosting machine[J]. The Annals of Statistics, 2001, 29(5): 1189−1232

[34] Agarwal R, Schwarzer M, Castro P S, et al. Deep reinforcement learning at the edge of the statistical precipice[C]//Advances in Neural Information Processing Systems 34. Red Hook, NY: Curran Associates, Inc., 2021: 29304−29320

[35] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997, 55(1): 119−139

[36] Cohen W W, Singer Y. A simple, fast, and effective rule learner[C]// Proc of the 16th National Conf on Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conf. Menlo Park, CA: AAAI, 1999: 335−342



Shao Jiejing, born in 1997. PhD. His main research interests include neuro-symbolic learning, interpretable machine learning, and active learning.
邵杰晶, 1997 年生。博士。主要研究方向为神经符号学习、可解释机器学习、主动学习。



Yang Ming, born in 2002. Master candidate. His main research interests include neuro-symbolic computation and multimodal learning.
杨　铭, 2002 年生。硕士研究生。主要研究方向为神经符号计算、多模态学习。



Yang Xiaowen, born in 2000. PhD candidate. His main research interests include neuro-symbolic learning and interpretable machine learning.
杨骁文, 2000 年生。博士研究生。主要研究方向为神经符号学习、可解释机器学习。



Li Yufeng, born in 1983. PhD, professor, PhD supervisor. Distinguished member of CCF. His main research interests include machine learning, data mining, semi-supervised and weakly supervised learning, and statistical learning and optimization and their applications.
李宇峰, 1983 年生。博士, 教授, 博士生导师。CCF 杰出会员。主要研究方向为机器学习、数据挖掘、半监督学习与弱监督学习、统计学习与优化及其应用。