计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330547
61(7): 1791−1798, 2024

带拒绝推理的反绎学习方法

黄宇轩　姜　远
（计算机软件新技术全国重点实验室（南京大学）　南京　210023）
（huangyx@lamda.nju.edu.cn）

Abductive Learning Method with Rejection Reasoning

Huang Yuxuan and Jiang Yuan
（*National Key Laboratory for Novel Software Technology* (*Nanjing University*), *Nanjing* 210023）

Abstract　In recent years, many research efforts have focused on combining data-driven machine learning with knowledge-driven logical reasoning, aiming to improve the performance of machine learning systems. Many works have attempted to use abductive reasoning to integrate machine learning and logical reasoning into a unified framework. These methods typically first generate pseudo-labels through machine learning models, and then use abductive reasoning to revise the inconsistent pseudo-labels, which serves to update the machine learning model and the above routine repeats. However, there may be incorrect labels in the abduced labels, which could have a negative impact on the training of the machine learning model and are challenging to detect. We propose abductive learning with rejection reasoning, a method that takes into account both model uncertainty and reasoning uncertainty of the abduced labels. This strategy comprehensively evaluates the reliability of abductive results from the perspective of both the data and knowledge levels, and avoids the negative impact of unreliable abduced labels on model training by rejecting some of the abductive reasoning results. Empirical studies show that the proposed method can reduce the proportion of incorrect abduced labels. In turn, this facilitates a faster training process for abductive learning and improves its overall performance.

Key words　abductive learning; machine learning; logical reasoning; abductive reasoning; neuro-symbolic learning

摘　要　近年来，许多研究工作致力于将数据驱动的机器学习和知识驱动的逻辑推理相结合，以提高机器学习的性能. 其中，不少工作尝试利用反绎推理，将机器学习与逻辑推理融合到一个框架中. 这些方法通过机器学习模型生成伪标记，然后利用反绎推理来修正不一致的伪标记，以更新机器学习模型并多次迭代. 然而，反绎中可能会存在错误标记，这些标记会对模型训练产生负面影响且难以被发现. 因此提出一种带拒绝推理的反绎学习方法，它同时考虑反绎标记的模型不确定性和推理不确定性，从数据层面和知识层面综合评估反绎结果的可靠性，并通过拒绝部分反绎推理结果来避免不可靠的反绎标记对模型训练的负面影响. 实验表明，提出的方法可以减少错误反绎标记的比例、加速反绎学习的训练并带来更好的性能.

关键词　反绎学习；机器学习；逻辑推理；反绎推理；神经符号学习

中图法分类号　TP181

融合机器学习与逻辑推理一直是人工智能（artificial intelligence, AI）领域长期存在且备受关注的问题. 为了克服当前机器学习方法的局限性，许多研究者倡导下一代人工智能需要将数据驱动的机器学

收稿日期：2023-06-24；**修回日期**：2023-11-06
基金项目：国家自然科学基金项目（62176117）
This work was supported by the National Natural Science Foundation of China (62176117).
通信作者：姜远（jiangy@lamda.nju.edu.cn）

习与知识驱动的推理(例如逻辑推理)相结合[1]. 神经符号学习(neuro-symbolic learning, NeSy)[2-3]和统计关系人工智能(statistical relational AI)[4]是近年来这个方向的代表性工作，它们尝试基于一阶逻辑表示的领域知识构建神经网络结构或概率图模型. 另一方面，概率逻辑程序(probabilistic logic program, PLP)[5]则尝试扩展一阶逻辑以容纳概率逻辑事实，从而实现概率推理.

反绎学习(abductive learning, ABL)[6-7]是一种将机器学习模型与逻辑推理模型结合的新框架. 该框架可以同时保留双方的完整表达能力：机器学习模型将原始未标记输入数据(例如图像、文本)转换为符号表示的逻辑事实，称为伪标记(pseudo-label)，这些伪标记被用作逻辑推理部分的输入. 而一阶逻辑推理模型试图对伪标记进行反绎推理(abductive reasoning)，以修正逻辑事实的真值，并用于更新机器学习模型. 由于反绎学习保留了完整的逻辑推理能力，因此它可以直接利用一阶逻辑规则表示的背景知识库，并减少对大量标记数据的需求.

在现有的反绎学习方法中，根据最小化不一致性的原则，会为所有未标记样本生成相应的反绎标记(abduced label). 具体来说，由于反绎推理是非确定性的，对于每个未标记的样本，可能存在多个反绎标记. 反绎学习基于最小化反绎标记和符号背景知识之间的不一致性的原则来选择最佳的反绎标记. 随后，这些反绎标记会被当作未标记样本的真实标记，与输入示例一起用于更新机器学习模型. 由于机器学习模型预测的伪标记可能存在错误，反绎推理也存在不确定性，因此得到的反绎标记也具有不确定性，不一定与真实标记相同.

反绎标记中可能存在不正确的标记，这可能损害机器学习模型的性能，并且很难被发现. 例如，在手写数字等式识别任务中，若将一个手写数字图片“1”错误地修改为“6”并用于训练机器学习模型. 由于这个错误标记的误导，更新后的模型可能会将很多与数字“1”相似的图片都识别为数字“6”. 这不仅会导致反绎学习的训练速度变慢，还可能使得其中的机器学习模型陷入局部最优解[8-9]，无法收敛到最优. 由于输入的样本没有真实标记，难以发现隐藏在不确定反绎标记中的错误，这给学习带来了不少困难.

为了解决上述问题，本文定义并设计了针对反绎标记的不确定性. 由于反绎标记同时受到机器学习模型的伪标记以及逻辑推理的影响，我们同时考虑机器学习模型生成伪标记的模型不确定性以及反绎推理的推理不确定性对它的影响. 其中，模型不确定性由分类器输出的置信度(confidence)计算而来，而推理不确定性通过反绎逻辑推理得到的结果集合的大小进行估算. 基于上述反绎标记的不确定性，我们提出了一种带拒绝推理的反绎学习(abductive learning with rejection reasoning, $\mathrm{ABL_r}$)方法. 该方法尝试过滤反绎学习中具有较大不确定性的反绎标记，即若遇到不确定的样本，$\mathrm{ABL_r}$会拒绝反绎推理的结果并抛弃该样本. 我们在若干数据集上进行了实验，验证了本文提出的方法可以拒绝错误的反绎标记，不仅加速反绎学习训练，还能同时带来更好的性能.

1 相关工作

神经符号学习[2-3]提出将符号推理与神经网络相结合，以学习从环境中感知并对已感知的信息进行推理. 大多数 NeSy 方法采用端到端的深度神经网络来建模这一过程，其中使用符号表示的领域知识库构建神经网络结构[2,10-11]. 然而，这类方法大多用分布式表示替代符号表示，并采用模糊算子近似逻辑推理. 因此，它们通常需要大量的标记训练数据.

PLP[5]和统计关系学习(statistical relational learning, SRL)[4,12]试图在保留逻辑公式的同时为逻辑模型提供了概率语义. PLP 通过扩展一阶逻辑使其能进行概率逻辑推理；SRL 利用领域知识构建概率图模型结构进行统计推理. 这些方法通常需要直接的语义级别输入，难以应用于如图像等原始数据的输入.

近期，一些研究者提出了建立混合模型的想法，这些方法通常由一个用于机器学习的感知模型和一个用于逻辑推理的推理模型组成. 代表性框架包括 DeepProbLog[13]、反绎学习[6-7]和神经语法符号模型(neural-grammar-symbolic model, NGS)[14]. 这些方法中的感知模型负责学习将原始数据转换为原始逻辑事实，作为符号推理的输入；而推理模型则尝试基于给定的知识库推理并修改所感知到的逻辑事实的真值，以更新感知模型. 这 2 个系统的集成是通过反绎来实现的.

近年来，研究者提出了许多基于反绎学习的方法. ABL[15]是第 1 个基于反绎学习框架实现的方法. SS-ABL[16]将半监督学习与反绎学习结合，并成功应用于法律文书判决任务. GABL[8]提出了在知识库为具体事实库(ground knowledge base)时的反绎策略及对应的反绎学习方法. $\mathrm{Meta_{Abd}}$[9]基于反绎学习，提出了同时学习机器学习模型和一阶逻辑知识库的方法.

ABLSim[17]提出使用集束搜索(beam search)结合相似度加速反绎学习中的一致性优化过程. ABL_{nc}[18]提出在开放环境出现新概念时，可对反绎学习的知识库进行知识精化(knowledge refinement). ABL-KG[19]提出可以从大规模知识图谱中提取规则库作为反绎学习的知识库.

上述文献[8−9, 16−19]所述的反绎学习相关方法分别关注不同的问题设定，例如在知识库不完备的情况下，如何利用知识图谱或数据等作为知识库的来源. 一般而言，这些方法通过反绎推理修改标记后，会使用所有的反绎标记更新机器学习模型. 与上述方法不同，本文方法假设知识库已知，并选择性地挑选一部分反绎标记用于模型的更新，以提升训练效率和模型性能.

2 基础知识

本文提出的方法主要基于反绎学习，下面对相关概念和基本知识进行介绍.

2.1 反绎推理

反绎推理是逻辑推理的一种基本形式，它尝试通过逻辑蕴含关系寻求一个假设，以解释观察到的现象. 为了更清晰地阐述，本文将逻辑符号表示为："∧"表示合取(逻辑与)；"∨"表示析取(逻辑或)；"←"表示蕴涵，即若"←"右侧的前提成立，则左侧的结论成立. 例如，考虑以下命题逻辑规则：

$$\text{草地湿} \leftarrow \text{昨晚下雨} \vee \text{洒水器开启}, \tag{1}$$

$$\text{鞋子湿} \leftarrow \text{草地湿}, \tag{2}$$

$$\text{假} \leftarrow \text{昨晚下雨} \wedge \text{洒水器开启}, \tag{3}$$

其中前2条规则(式(1)和式(2))阐明了草地和鞋子湿的原因，后一条规则(式(3))表明昨晚下雨和洒水器开启这2个命题不能同时为真. 当我们观察到鞋子湿时，式(2)表明草地湿也应该为真. 延续这个过程，根据式(1)，昨晚下雨和洒水器开启都是2种可能的解释. 假如我们还观察到昨晚没有下雨，根据式(3)，洒水器开启便是鞋子湿这个现象的唯一的解释. 上述例子描述了根据鞋子湿这一现象，反绎推理得到洒水器开启这一解释/假设的推理过程.

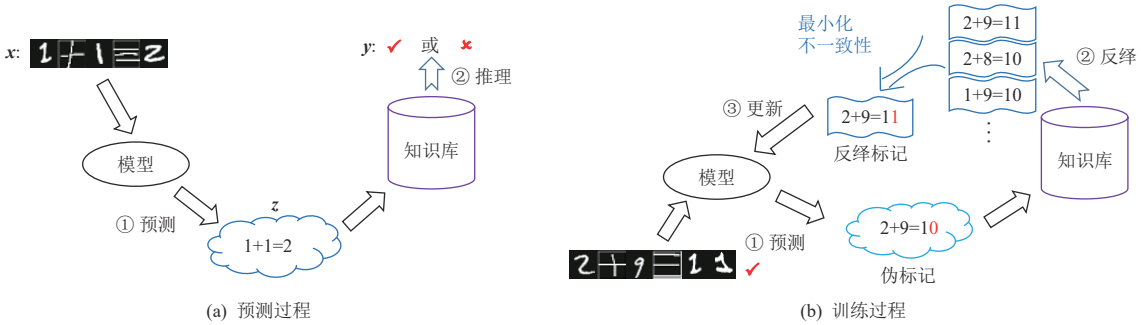
2.2 反绎学习框架

反绎学习[6-7]是一个融合感知模型和推理模型的框架. 感知模型负责将原始输入数据$\boldsymbol{x} \in \mathcal{X}$映射到离散符号$\boldsymbol{z} \subseteq \mathcal{Z}$，其中$\mathcal{X}$为输入空间，$\mathcal{Z}$为感知模型的输出空间；$\boldsymbol{x}$和$\boldsymbol{z}$分别为输入示例和提取的符号组成的向量，其中$\boldsymbol{x}=(x_1,x_2,\cdots)$和$\boldsymbol{z}=(z_1,z_2,\cdots)$分别表示$\boldsymbol{x}$中的各个示例和相应的符号. 以图1手写等式识别任务为例，输入的等式图像是$\boldsymbol{x}$，各个分割好的手写数字图像是x_i；预测的等式符号序列是$\boldsymbol{z}$，其中每个符号分别为z_i；等式是否满足知识库表示为$y=\text{True}$或$y=\text{False}$.

推理模型包含一个一阶逻辑规则知识库KB，它可以接收符号序列$\boldsymbol{z}$并通过逻辑推理得到最终输出$y \in \mathcal{Y}$. 例如，给定十进制整数的加法规则，当输入事实$\boldsymbol{z}=(1,+,1,=,2)$(记为"1+1=2")时，推理模型将输出具体逻辑事实(ground fact)$y=\text{True}$，表示这是一个正确的等式；而当输入$\boldsymbol{z}$为"2+9=10"时，其输出具体逻辑事实$y=\text{False}$，表明这是一个不正确的等式. 图1(a)展示了在手写等式破解任务[6]中预测过程的一个例子.

反绎学习的训练过程可以形式化为：给定未标记数据集X、知识库KB和最终期望的输出集合Y，反绎学习旨在学习一个感知模型$f:\mathcal{X}\mapsto\mathcal{Z}$，它能准确预测输入实例的标记，并且这些标记可以与$KB$一起逻辑蕴涵(entail)得到$y$. 由于我们没有$\boldsymbol{z}$的监督信息，因此类似于弱监督学习[20]，我们将$\boldsymbol{z}$称为伪标记.

反绎学习训练包括3个步骤：预测伪标记、反绎推理标记和更新模型. 首先，对输入数据$\boldsymbol{x} \in X$，反绎学习使用感知模型f获取预测伪标记$\boldsymbol{z}=f(\boldsymbol{x})$. 然后，推理模型接收符号$\boldsymbol{z}$并判断它们与一阶逻辑知识库



(a) 预测过程 (b) 训练过程

Fig. 1 The predicting process and training process of abductive learning framework

图1 反绎学习框架的预测过程和训练过程

KB的一致性，若不一致，则通过反绎推理将伪标记z修改为$\bar{z}$. 通常来说，存在几个与KB一致的候选$\bar{z}$. 推理模型基于最小化数据和知识库之间不一致性的原则，推理出最有可能正确的伪标记$\bar{z}$. 最后，反绎学习将$\bar{z}$视为真实标记来更新感知模型，上述过程不断迭代重复. 图 1(b)展示了一个反绎学习例子，其中输入的具体事实真实标记是“2+9=11”且y = True. 感知模型f预测了错误的伪标记z =“2+9=10”. 反绎推理得到“2+9=11”“2+8=10”“1+9=10”等可能的结果，在最小化不一致性之后，选择$\bar{z}$ =“2+9=11”作为最终反绎标记并用于更新模型f.

3 带拒绝推理的反绎学习方法 $\mathbf{ABL_r}$

本文提出了带拒绝推理的反绎学习方法 ABL，其基本思想是在反绎学习中拒绝高不确定性的反绎标记，只用剩余较为确定的反绎标记及对应样本更新模型. 在本节，我们首先提出反绎标记的不确定性这个概念，并将其分为模型不确定性和推理不确定性，给出相应的设计. 随后介绍基于模型不确定性和推理不确定性的拒绝推理策略，该策略同时考虑反绎标记的模型不确定性和推理不确定性，并拒绝模型和推理不确定性均较高的样本. 最后介绍本文提出的 $\mathrm{ABL_r}$ 方法，该方法在反绎学习中利用了拒绝推理策略，并动态调节参数以控制拒绝样本的比例，使得训练样本数量保持在预设范围之内.

3.1 反绎标记的模型不确定性

由于反绎标记是由机器学习模型生成的伪标记修改而得，因此可以通过模型的输出计算得到反绎标记的模型不确定性. 具体而言，我们在计算模型不确定性时，用到机器学习模型输出的置信度. 机器学习模型的置信度是指模型对某个预测结果的可信程度或者预测的准确度，其取值范围为 [0,1]. 在分类问题中，模型的置信度通常指模型对某个样本属于某个类别的概率估计值. 例如，对于一个二分类问题，模型对于某个样本预测其属于正类的概率为 0.8，那么我们可以认为这个模型的置信度较高，因为它对预测结果比较自信. 而如果模型对于另一个样本预测其属于正类的概率为 0.5，那么我们就认为这个模型的置信度比较低.

我们定义反绎标记$\bar{z}$的模型不确定性为：

$$ModelUncertainty(\bar{z}) = 1 - \frac{1}{|\bar{z}|}\sum_{\bar{z}_i\text{为}\bar{z}\text{的元素}} Conf(x_i,\bar{z}_i), \quad (4)$$

其中，$Conf(x_i,\bar{z}_i)$指的是模型f输出的关于样本x_i属于类别$\bar{z}_i$的置信度，$|\bar{z}|$是输入示例的个数. 式(4)等号右侧的$\frac{1}{|\bar{z}|}\sum_{\bar{z}_i\text{为}\bar{z}\text{的元素}} Conf(x_i,\bar{z}_i)$计算的是机器学习模型对各个示例对应的反绎标记的平均置信度，若反绎标记的平均置信度越高，说明模型对对应的标记越自信，因此$ModelUncertainty(\bar{z})$也越低. 通常情况下，若机器学习模型具有一定的预测能力，那么一个准确的反绎标记应该有较高的置信度以及较低的模型不确定性. 因此，可以根据模型不确定性的大小评估反绎标记的质量，并对反绎标记进行筛选，以确保反绎标记的可靠性.

3.2 反绎标记的推理不确定性

在引言中提到，反绎推理具有不确定性，推理模型可以通过逻辑推理得到若干反绎标记，因此，除了模型不确定性外，我们还可以从推理的角度描述反绎标记的不确定性，即推理不确定性. 例如，对于一个手写等式，如果现有的知识库可以反绎推理得到的反绎候选标记集合为{“1+1=2”}，由于集合内只有 1 个元素，说明在这种情况下，只有这个修改后的标记与知识库一致，也就是说，推理部分对推理结果的确定性非常高，因此可以直接将该标记作为反绎标记更新模型. 如果知识库反绎推理得到的反绎候选标记集合为{“1+1=2”,“1+2=3”,⋯}，共包含 10 个候选反绎标记，而在最小化不一致性时需要从中选择 1 个作为最终反绎标记，此时能选到样本对应真实标记的概率，与仅有 1 个候选标记的情况相比大大降低. 因此，我们可以通过反绎候选标记的集合的大小衡量推理部分的不确定性.

形式化地说，对一个训练样本，给定最终输出y和知识库KB，令A为该样本反绎推理得到的所有与知识库一致的候选标记的集合，即$A = \{\bar{z} \mid KB \cup \bar{z} \vDash y\}$，我们可以定义反绎标记$\bar{z}$的推理不确定性为：

$$ReasoningUncertainty(\bar{z}) = 1 - \frac{1}{|A|}, \quad (5)$$

其中，$|A|$指的是该样本的反绎候选标记集合内的元素个数，而$\frac{1}{|A|}$表示从集合A中随机选 1 个反绎候选标记的概率. 与模型不确定性相同，推理不确定性的取值也是在 [0,1] 之间. 反绎推理得到的可能标记个数越多，从中能选中真实标记的概率越低，因此推理不确定性也就越高.

3.3 基于模型不确定性和推理不确定性的拒绝推理策略

在反绎学习中，我们需要同时考虑模型不确定性和推理不确定性的影响，以评估反绎推理结果的可靠性. 如果模型不确定性和推理不确定性都很低，

则反绎标记是真实标记的可能性较高；如果模型不确定性和推理不确定性都较高，那么反绎结果很可能不可靠，给后续训练带来噪声，使得机器学习模型收敛速度变慢同时性能降低. 为了减少不可靠反绎标记对反绎学习的负面影响，我们需要根据反绎标记的不确定性，拒绝一部分推理得到的反绎标记.

我们设计了一种拒绝推理策略，即根据反绎标记的不确定性来决定是否接受该标记. 在这种策略中，我们根据模型不确定性和推理不确定性的程度分别定义 2 个阈值θ_{m}和θ_{r}，如果某个训练样本对应的反绎标记的模型不确定性和推理不确定性均大于该阈值，则将其拒绝，并将其从反绎学习本轮循环的训练数据集中移除. 阈值θ_{m}和θ_{r}的值越高，则被拒绝的样本越少，反之，其值越低，则被拒绝的样本数量越多. 具体来说，令$\bar{Z}$为所有样本的反绎标记$\bar{z}$的集合，则需要拒绝的反绎标记集合为

$$\bar{Z}_{\mathrm{rej}}=\{\bar{z}\in\bar{Z}\mid ModelUncertainty(\bar{z})>\theta_{\mathrm{m}}\wedge ReasoningUncertainty(\bar{z})>\theta_{\mathrm{r}}\},\tag{6}$$

剩下的反绎标记用于训练，即

$$\bar{Z}_{\mathrm{use}}=\bar{Z}\setminus\bar{Z}_{\mathrm{rej}}=\{\bar{z}\in\bar{Z}\mid ModelUncertainty(\bar{z})\leqslant\theta_{\mathrm{m}}\vee ReasoningUncertainty(\bar{z})\leqslant\theta_{\mathrm{r}}\}.\tag{7}$$

这个拒绝推理策略的思想是：当一个样本的模型不确定性和推理不确定性都高于对应阈值时，那么这个标记很可能是错误的或者不可靠的，因此我们拒绝该样本. 例如，考虑图 1(b)中的手写等式任务，若反绎候选标记只有 1 个(“2+9=11”)，说明推理部分的不确定性很低，这很可能是正确的标记，因此即便此时模型不确定性较高，也能用于更新模型；反之，若反绎标记的平均置信度很高(如 0.9)，说明模型对这个反绎标记非常自信，同样此时这很可能是正确的标记，因此即便此时推理不确定性较高，也可以相信这个伪标记；如果模型不确定性和推理不确定性都较高，此时该标记很可能包含噪声，因此拒绝该标记.

3.4 带拒绝推理的反绎学习

图 2 展示了带拒绝推理的反绎学习方法 ABL$_{\mathrm{r}}$ 的基本流程. 从图 2 可以看到，在反绎推理得到样本的反绎标记后，ABL$_{\mathrm{r}}$ 会基于模型不确定性和推理不确定性的拒绝推理策略，来判断是否应该拒绝该反绎标记. 如果决定拒绝，那么在本轮循环暂不使用该样本；如果没有拒绝，那么将其视为真实标记更新机器学习模型.

在方法的具体实现中，ABL$_{\mathrm{r}}$ 还引入了动态自适应阈值机制. 当训练刚开始时，通常此时机器学习模型输出的置信度较低. 如果直接将 3.3 节的拒绝推理策略应用于反绎学习，则大多数样本的模型不确定性都会较高，从而可能导致训练样本过少的问题. 若直接将模型不确定性阈值θ_{m}固定为一个较高的值，那么绝大部分反绎标记样本都会被接受，失去了拒绝的作用. 为了解决这个问题，ABL$_{\mathrm{r}}$ 方法动态调整模型不确定性阈值θ_{m}，确保反绎样本的数量不会过少：给定训练样本的最低占比ρ，如果根据阈值θ_{m}得到反绎标记后，其数量$|\bar{Z}_{\mathrm{use}}|$占总数$|\bar{Z}|$的比例小于ρ，那么将θ_{m}动态调整为能使$|\bar{Z}_{\mathrm{use}}|\geqslant\rho|\bar{Z}|$成立的最小阈值$\theta'_{\mathrm{m}}$，即$\theta'_{\mathrm{m}}=\arg\min_{\theta_{\mathrm{m}}}|\bar{Z}_{\mathrm{use}}|/|\bar{Z}|\geqslant\rho$，其中$\bar{Z}_{\mathrm{use}}$的定义为式(7). 一般来说，超参数$\rho$的值越大，调整后模型不确定性$\theta'_{\mathrm{m}}$的值越高. 通过此做法，ABL$_{\mathrm{r}}$ 可以在训练开始时自动提高模型不确定性阈值θ_{m}，避免训练样本过少的问题.

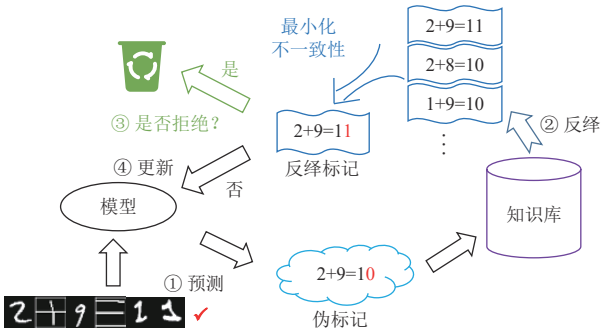


Fig. 2 Training process of ABL$_{\mathrm{r}}$

图 2 ABL$_{\mathrm{r}}$ 方法的训练过程

ABL$_{\mathrm{r}}$ 算法如算法 1 所示. 容易计算出，拒绝推理部分的时间复杂度为$O(|\boldsymbol{z}|)$，空间复杂度为$O(1)$，而不使用拒绝推理方法的时间和空间复杂度均为$O(1)$. 尽管采用拒绝推理导致时间复杂度增加，但得益于训练数据中噪声的减少，实际运行速度却有所提升. 值得注意的是，由于反绎学习是一个通用的框架，具有足够的灵活性，因此其中的机器学习模型f可以是任何模型，如神经网络、决策树、随机森林等.

算法 1. 带拒绝推理的反绎学习方法 ABL$_{\mathrm{r}}$.

输入：未标记数据X，最终期望的输出Y，知识库KB，机器学习模型f，阈值θ_{m}，θ_{r}，ρ，迭代轮数T；

输出：机器学习模型f.

① for $t=1$ to T do

② $\quad Z=f(X)$；

③ $\quad \bar{Z}=Abduce(KB,Z,Y)$；

④ $\quad \bar{Z}_{\mathrm{use}}=Filter\left(\bar{Z},\theta_{\mathrm{m}},\theta_{\mathrm{r}}\right)$；/*根据式(7)计算*/

⑤ $\quad$ if $|\bar{Z}_{\mathrm{use}}|<\rho|\bar{Z}|$ then

⑥ $\quad\quad \bar{Z}_{\mathrm{use}}=Filter\left(\bar{Z},\theta'_{\mathrm{m}},\theta_{\mathrm{r}}\right)$；/*$\theta'_{\mathrm{m}}$为能使$|\bar{Z}_{\mathrm{use}}|\geqslant\rho|\bar{Z}|$成立的最小阈值*/

⑦ end if

⑧ $f = Update\left(f, X_{use}, \bar{Z}_{use}\right)$;

⑨ end for

4 实 验

在本节中, 我们通过在若干个数据集上进行实验, 将本文提出的 ABL_r 方法与多种方法进行对比和分析, 以验证本文所提出的方法的有效性.

4.1 实验数据

我们在神经符号学习领域的 3 个公开数据集上进行实验, 包括 Addition, HWF, HED. 表 1 提供了实验数据集的详细信息.

Table 1 Basic Information of the Datasets in the Experiments

表 1 实验中所用数据集的基本信息

数据集	训练样本数量	测试样本数量	类别数量	领域知识库
Addition	30 000	10 000	10	个位数加法规则
HWF	9 000	2 000	13	加减乘除运算规则
HED	20 000	2 000	12	多位数加法规则

Addition[13] 数据集最初在 DeepProbLog[13] 中被提出, 其输入是一对 MNIST 图像, 最终输出是它们的总和. 该数据集要求根据原始输入的手写数字图片和最终计算结果, 学习一个能识别未见过手写数字的机器学习模型. 同时, 我们还有关于个位数加法规则的一阶逻辑领域知识库.

HWF[14] 数据集包括手写的十进制式子以及它们的最终计算结果, 如“1+4/3”. 与此同时, 我们还有关于十进制数字的运算规则. 我们希望通过同时利用输入数据和领域知识, 学习一个能准确识别手写数字/符号的分类器. 值得注意的是, 由于原始论文中的 HWF 数据集包括长度为 1 的式子用作预训练模型, 我们在实验中删除了这些式子. 因此, 我们的运行结果与原文[14] 中的结果有所不同.

HED[15] 数据集源自于反绎学习[15] 论文中的手写等式破解任务. 该数据集的示例如图 1 所示, 其输入包含十进制等式算术的图像, 涵盖 12 个符号(0,1,…, 9,+,=), 最终输出为其正确性标记. 此外, 知识库包含关于如何进行十进制加法运算的符号规则, 用于训练的等式的正确性标记(True 或 False)也以逻辑事实的形式提供. 与前 2 个数据集一样, 我们希望通过同时利用数据和知识库学习一个能准确识别手写图像的分类器. 由于该任务较难, 我们为每一类别提供 2 张带有标记的图片供模型预训练.

4.2 评价指标及基准模型

在本文研究中, 由于主要目标是利用一阶逻辑领域知识库辅助学习机器学习模型, 我们采用常用的评价指标准确率(accuracy)来评估机器学习模型的性能. 同时, 为了验证本文方法对训练效率的影响, 我们还统计了不同方法的运行时间. 在此, 我们将运行时间定义为模型测试准确率达到目标值(Addition 和 HWF 数据集为 98%, HED 数据集为 97%)所需的时间.

我们将本文方法 ABL_r 与 DeepProbLog[13], NGS[14], ABL[15] 进行比较. DeepProbLog 通过梯度下降方法融合概率逻辑推理和神经网络. NGS 采用类似反绎的思想, 将上下文无关语法作为知识库, 并使用马尔可夫链蒙特卡罗采样, 根据后验概率修改伪标记并更新模型. ABL 通过反绎推理并最小化不一致性的思想修改标记, 与 ABL_r 不同的是, 这里的 ABL 没有采用拒绝推理策略.

在实验中, 我们通过对训练数据进行了机器学习中广泛接受和使用的 10 折交叉验证来确定本文方法的超参数. 我们设定阈值 $\theta_m = 0.4, \rho = 0.8, \theta_r = 0.86$. 所有方法均设置了 30 min 的限时, 若 30 min 内没有收敛, 则判定为超时. 我们在一台配置有 Intel Xeon Gold 6248R CPU 和 Nvidia Tesla V100S GPU 的服务器上进行了 10 次重复实验, 并统计了平均值和标准差. 所有方法都共享相同的知识库和随机初始化的感知模型.

4.3 实验结果与分析

表 2 展示了 ABL_r 方法与其他对比方法在不同数据集上的准确率和训练时间对比. 可以看出, ABL_r 方法在 3 个数据集上都取得了最高的准确率, 并且相比其他方法用了更短的时间收敛.

Table 2 Perception Accuracy and Training Time of Different Methods in the Experiments

表 2 实验中不同方法的感知准确率和训练时间

评价指标	数据集	DeepProbLog	NGS	ABL	ABL_r (本文)
准确率/%	Addition	96.5±0.5	98.5±0.7	90.7±9.6	**98.9±0.4**
	HWF	32.2±0.6	99.5±0.3	99.7±0.1	**99.9±0.1**
	HED	83.7±6.4	96.7±5.4	97.0±0.2	**97.5±0.1**
训练时间/s	Addition	596±6	203±5	501±50	**140±14**
	HWF	超时	180±8	171±15	**132±10**
	HED	超时	606±68	608±83	**557±61**

注: 黑体数值为最优值. “±”前后分别表示平均值和标准差.

在 Addition 数据集中, 4 种方法都能达到 90% 以上的准确率. DeepProbLog 能正常收敛, 但却耗时最长, 这是因为其对概率逻辑程序的梯度计算非常耗

时. NGS 具有与 ABL 类似的反绎思想，最终模型达到了较高的准确率，但由于其随机游走采样过程较耗时，因此比 ABL_r 方法花费的时间更长. ABL 方法的性能较低，主要原因是有几次实验 ABL 陷入了局部最优且跳出较慢. 在陷入局部最优时，由于 ABL 将所有反绎标记用于训练，其对训练集中的部分错误反绎标记产生了过拟合，因此收敛较慢甚至偶尔一直停留在局部最优点. ABL_r 方法通过拒绝部分不可靠的反绎结果，使得模型接受了较少的标记噪声，因此收敛更快且最终性能更好.

在 HWF 数据集的实验中，可以看到除了 DeepProbLog 之外的方法都收敛且达到了很高的准确率. 这是由于 HWF 数据集相比其他数据集更加简单，因此学习较为容易. DeepProbLog 性能较低的原因主要是其推理效率太低，而 HWF 的知识库较为复杂，因此到达设定限时的时候，其仍在收敛过程中. 由于数据集较为简单，因此 NGS, ABL, ABL_r 在收敛速度上差别不大，ABL_r 稍微比其他方法收敛快一些，同样说明拒绝部分样本后虽然数据集变小，但减少了错误标记的影响后反而提升了性能和速度.

HED 是实验中难度最大的一个数据集，因此整体训练时间比其他任务长且最终准确率比其他任务低. 同样，我们的 ABL_r 方法在 HED 数据集上也取得了最高的准确率和最快的收敛速度. DeepProbLog 在该数据集上表现欠佳，主要是由于数据集的复杂性导致模型的学习难度较大. 由于使用全部反绎标记，NGS 和 ABL 方法也出现了过拟合错误反绎标记的问题，导致收敛较慢. ABL_r 方法通过拒绝部分不可靠的反绎结果，更快地跳出局部最优，并达到更好的性能. 总体而言，ABL_r 方法在这 3 个数据集上都表现出色，并且相比其他方法具有更好的准确率和训练效率.

4.4 消融实验分析

ABL_r 方法综合考虑了模型不确定性和推理不确定性，并加入了动态自适应阈值机制. 为了深入研究这 3 个部分对最终效果的影响，我们设计了 ABL_r 方法中关于模型不确定性(model uncertainty, MU)、推理不确定性(reasoning uncertainty, RU)和动态阈值(dynamic threshold, DT)的消融实验. 在消融实验中，我们分别移除 ABL_r 方法的模型不确定性(w/o MU)、推理不确定性(w/o RU)和动态阈值(w/o DT)部分，以研究这 3 个部分对最终效果的贡献.

在消融实验中，我们比较了 5 种方法：

1)ABL 基准方法(ABL). 使用原始的 ABL 方法，即接受所有反绎标记和样本.

2)ABL_r 基准方法(ABL_r). 使用完整的 ABL_r 方法，即包括模型不确定性、推理不确定性和动态阈值.

3)移除模型不确定性(w/o MU). 从 ABL_r 方法中移除模型不确定性部分，只保留推理不确定性和动态阈值部分.

4)移除推理不确定性(w/o RU). 从 ABL_r 方法中移除推理不确定性部分，只保留模型不确定性和动态阈值部分.

5)移除动态阈值(w/o DT). 从 ABL_r 方法中移除动态阈值部分，只保留模型不确定性和推理不确定性部分.

消融实验的结果如表 3 所示. 在本次实验中，我们统计了各种方法在测试准确率和反绎准确率 2 个方面的表现. 测试准确率衡量了机器学习模型在测试集上的最终预测能力，而反绎准确率则关注在第 1 轮中用于训练的反绎标记的准确率. 实验结果表明，与 ABL 方法相比，ABL_r 在拒绝部分不可靠标记后，在 3 个数据集上的反绎准确率均有所提高，这说明我们的方法在减少错误反绎标记的比例方面取得了明显的效果. 此外，在分别去除模型不确定性和推理不确定性后，测试准确率都出现了一定程度的降低，这进一步说明，模型不确定性和推理不确定性部分对 ABL_r 方法的性能发挥了积极作用. 与此同时，大部分相应的反绎准确率也得到了提高，这进一步证实了拒绝样本能够提高真实标记的比例.

Table 3 Ablation Study: Testing Accuracy and Abduced Accuracy of Different Methods

表 3 消融实验：不同方法的测试准确率和反绎准确率 %

准确率	数据集	ABL	ABL_r (本文)	w/o MU	w/o RU	w/o DT
测试	Addition	90.7±9.6	98.9±0.4	98.5±0.2	98.6±0.2	86.6±30.3
	HWF	99.7±0.3	99.9±0.1	99.8±0.1	99.8±0.1	99.8±0.1
	HED	97.0±0.2	97.5±0.1	97.1±0.3	82.4±20.1	84.5±24.0
反绎	Addition	56.5±13.0	59.8±6.9	61.7±6.7	55.7±8.2	57.7±18.4
	HWF	91.9±0.8	93.5±0.9	96.4±0.6	91.7±1.3	96.0±0.5
	HED	77.9±5.0	79.9±5.0	77.4±9.2	72.8±6.2	70.9±4.6

注：“±”前后分别表示平均值和标准差.

我们还发现，在去除动态自适应阈值机制后，Addition 和 HED 数据集的测试准确率明显下降. 经过进一步分析，我们发现这是由于在训练初期，模型输出的置信度普遍较低，从而导致大量样本被拒绝，这会使模型陷入局部最优，且性能无法得到有效提升. 这些消融实验充分说明了 ABL_r 方法中的 3 个关键部分，即模型不确定性、推理不确定性和动态自适应阈值机制，都对整体性能产生了积极的影响.

5 总 结

本文提出了一种反绎学习的提升方法,这个方法同时从模型不确定性和推理不确定性来衡量反绎标记的可靠性,并拒绝可能不可靠的推理结果.实验结果表明,我们的方法能够有效地拒绝错误的反绎标记,从而加速反绎学习的训练过程,并提升机器学习模型的性能.此外,这个方法具有良好的通用性,同样可以应用于其他基于反绎的方法.尽管本文工作在评估反绎标记的模型不确定性和推理不确定性方面取得了一定的成功,但所采用的方法相对简单.如何更精确地刻画反绎标记的不确定性,以便更准确地识别不可靠的反绎标记,仍是一个亟待进一步研究的问题.

作者贡献声明: 黄宇轩负责提出方法、完成实验、撰写论文;姜远负责写作指导和修改审定.

参 考 文 献

[1] Bengio Y. The consciousness prior [J]. arXiv preprint, arXiv: 1709.08568, 2017

[2] Garcez A, Gori M, Lamb L C, et al. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning[J]. Journal of Applied Logics, 2019, 6(4): 611−632

[3] De Raedt L, Dumancic S, Manhaeve R, et al. From statistical relational to neuro-symbolic artificial intelligence [C] //Proc of the 29th Int Joint Conf on Artificial Intelligence. New York: Curran Associates, 2020: 4943−4950

[4] De Raedt L, Kersting K, Natarajan S, et al. Statistical Relational Artificial Intelligence: Logic, Probability, and Computation [M]. California: Morgan & Claypool Publishers, 2016

[5] De Raedt L, Kimmig A. Probabilistic (logic) programming concepts[J]. Machine Learning, 2015, 100(1): 5−47

[6] Zhou Zhihua. Abductive learning: Towards bridging machine learning and logical reasoning [J]. Science China Information Sciences, 2019, 62(7): 76101: 1−76101: 3

[7] Zhou Zhihua, Huang Yuxuan. Abductive learning [M] //Neuro-Symbolic Artificial Intelligence: The State of the Art. Amsterdam: IOP Press, 2022: 353−379

[8] Cai Lewen, Dai Wangzhou, Huang Yuxuan, et al. Abductive learning with ground knowledge base [C] //Proc of the 30th Int Joint Conf on Artificial Intelligence. New York: Curran Associates, 2021: 1815−1821

[9] Dai Wangzhou, Muggleton S. Abductive knowledge induction from raw data [C] //Proc of the 30th Int Joint Conf on Artificial Intelligence. New York: Curran Associates, 2021: 1845−1851

[10] Trask A, Hill F, Reed S, et al. Neural arithmetic logic units [C] //Proc of the 31st Neural Information Processing Systems. Cambridge, MA: MIT, 2018: 8035−8044

[11] Grover A, Wang E, Zweig A, et al. Stochastic optimization of sorting networks via continuous relaxations [C/OL] //Proc of the 6th Int Conf on Learning Representations. 2018[2023-09-10]. https://openreview.net/pdf?id=H1eSS3CcKX

[12] Koller D, Friedman N, Džeroski S, et al. Introduction to Statistical Relational Learning [M]. Cambridge, MA: MIT Press, 2007

[13] Manhaeve R, Dumancic S, Kimmig A, et al. DeepProbLog: Neural probabilistic logic programming [C] //Proc of the 31st Neural Information Processing Systems. Cambridge, MA: MIT, 2018: 3749−3759

[14] Li Qing, Huang Siyuan, Hong Yining, et al. Closed loop neural-symbolic learning via integrating neural perception, grammar parsing, and symbolic reasoning [C] //Proc of the 37th Int Conf on Machine Learning. New York: ACM, 2020: 5884−5894

[15] Dai Wangzhou, Xu Qiuling, Yu Yang, et al. Bridging machine learning and logical reasoning by abductive learning [C] //Proc of the 32nd Neural Information Processing Systems. Cambridge, MA: MIT, 2019: 2811−2822

[16] Huang Yuxuan, Dai Wangzhou, Yang Jian, et al. Semi-supervised abductive learning and its application to theft judicial sentencing [C] //Proc of the 20th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2020: 1070−1075

[17] Huang Yuxuan, Dai Wangzhou, Cai Lewen, et al. Fast abductive learning by similarity-based consistency optimization [C] //Proc of the 34th Neural Information Processing Systems. Cambridge, MA: MIT, 2021: 26574−26584

[18] Huang Yuxuan, Dai Wangzhou, Jiang Yuan, et al. Enabling knowledge refinement upon new concepts in abductive learning [C] // Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 7928−7935

[19] Huang Yuxuan, Sun Zequn, Li Guangyao, et al. Enabling abductive learning to exploit knowledge graph [C] //Proc of the 32nd Int Joint Conf on Artificial Intelligence. New York: Curran Associates, 2023: 3839−3847

[20] Zhou Zhihua. A brief introduction to weakly supervised learning[J]. National Science Review, 2018, 5(1): 44−53

Huang Yuxuan, born in 1996. PhD. His main research interests include abductive learning, knowledge and data dual-driven learning, and machine learning.

黄宇轩, 1996年生. 博士. 主要研究方向为反绎学习、知识与数据双驱动学习、机器学习.

Jiang Yuan, born in 1976. PhD, professor, PhD supervisor. Her main research interests include artificial intelligence, machine learning and data mining.

姜 远, 1976年生. 博士, 教授, 博士生导师. 主要研究方向为人工智能、机器学习与数据挖掘.