计算机科学1997 Vol. 24 №. 3

一种集成学习模型 MIIE*)

刘贵全 陈小平 顾振梅 蔡庆生
(中国科学技术大学计算机系 合肥 230027)

摘 要 This paper gives an overview of theories of Inductive Logic Programming(ILP) and some limitations are identified. Then we construct a model for Integrating ILP and EBL(MIIE). The application of the Integrating Learning model is discussed in the end of the paper.

关键词 Inductive Logic Programming(ILP), Explanation Based Learning(EBL), Integrated Learning

归纳逻辑程序设计,以下简称 ILP,是将归纳学习的理论与逻辑程序设计的方法相结合的一种机器学习方法,ILP 以一阶逻辑为基础[1],从实例和背景知识出发进行机器学习.和归纳学习一样,ILP 的基本方法有一般化方法和特殊化方法.其一般化方法有:①从子句中去掉一个文字;②对子句施加逆置换(或进行逆归结).特殊化方法有:①往子句添加文字;②对子句施加置换.

基于解释的学习(Explanation Based Learning,简称 EBL)是另外一种机器学习方法.这种方法通过应用背景知识对单个实例的解释,得到一个高效的概念描述;其任务可描述为:应用单个实例和背景知识,将一个抽象的概念描述转换为一个有用的、可操作的概念描述.新的描述以宏规则的形式加到系统中,这往往会产生双重影响:一方面可以提高系统的(分类)效率;但另一方面,过多的具体规则加入又会降低系统的性能.传统 EBL 系统还有一个最大的问题就是,它要求提供给系统的背景知识是正确和完备的.这个条件对于一个学习系统来说显然太苛刻了,一般情况下很难保证.

1 问题的提出

ILP 的不足之处主要在于[1,2]:

1)ILP 学习的完备性不容易保证,即不能保证学习结果完全与实例吻合.

2)ILP 也可以看作是在概念描述空间进行搜索的问题[Mitchell 1982].但是猜想空间(概念描述空间)的描述语言比较难选择.选择很一般的语言会使空间太大而导致不可学习;选择具体的语言又可能使空间太小,又得不到正确的概念描述.

3)错误的排除也很困难.当有反例被覆盖时,如何对子句进行修改?当有正例未被覆盖时,是构造新规则(子句),还是修改已有规则?如果是选择已有规则进行修改,那么如何选择及如何修改?

4)现有 ILP 系统一般只能进行选择性归纳(即不能引入新的概念描述符),这限制了 ILP 的学习能力.从长远考虑需要进行构造性归纳学习.

ILP 方法的优点是它们从例子中归纳出概念描述的能力较强,但它们运用背景知识来指导对猜想空间进行有效搜索的能力较弱;基于解释的学习采用解释(实际上是演绎)的方法来学习概念和问题求解知识的能力较强,但不能处理不完备或有错误的背景(领域)知识.我们希望通过集成这两种学习方法,能充分发挥两者的优点,克服它们单个的不足.

集成 ILP 与 EBL 可从两方面进行研究:一是概念学习方面,运用背景知识来指导对猜想空间进行有效搜索;二是理论修正(也可称为知识求精)方面,运用例子对领域理论作(尽量少)的修改,提高其精确度.本文研究的主要是第二个方面,即理论修正方面.

2 ILP/EBL 用于理论(知识)求精

一般的 ILP 应用系统都只考虑了往系统知识中加入新规则,以覆盖正例而不覆盖反例.理论求精允许背景知识是不正确的,可以对其中的规则进行一般化,特殊化和删除等操作.

定义1(理论求精) 一般的理论求精问题可描述如下:给定一个领域理论 T,一个正例集 E^+和一

*)本文的研究得到了国家自然科学基金和863计划的支持.刘贵全 博士生;陈小平 博士;顾振梅 硕士;蔡庆生 教授,博士生导师.

个反例集 E^-；确定 T 的一个修正(或求精)T_r，满足 $T_r|=E^+$，但 $T_r|\neq E^-$。

修改最初的理论的时候不能改动其正确的地方，也就是说应尽量少改动原有理论。最小的语义改变是将不正确的例子看作理论的例外情况，因此很多系统把注意力集中在最小的语法改动上。这种方法常用的一个概念是"语法距离"，两个理论的"语法距离"可定义为将一个理论变换为另一个理论所需的基本操作的个数。但是当修正一个理论时这种方法是不实用的，因为对一个理论进行变换，结果有很多，然后还需要对这些结果进行搜索和比较。

对这个问题的解决可以运用 EBL 的方法。用初始理论 T 对训练实例进行解释，当 T 不能解释正例或错误地解释了反例时，可以得到一棵(完整的或不完整的)解释树，根据这棵解释树就能定出需要进行修改的规则的范围。

我们先来看看初始领域理论 T 错误地解释了反例 e 的情况。这时得到一棵关于 e 的证明树，这棵树上的规则集 $RULE_{max}$ 是将要被修改的规则的候选集，因为很显然需要被修改的最小规则集 $RULE_{min}$ 是 $RULE_{max}$ 的子集。

现在要做的就是确定 $RULE_{min}$ 并加以修改。具体的方法是：对 $RULE_{max}$ 中的某些子句进行特殊化，逐渐找出最好的修改结果。

当领域理论 T 不能解释正例 e 时，也会得到一棵解释树，但这棵树在某些地方与 e 中断了，即和 e 不匹配(图1)：

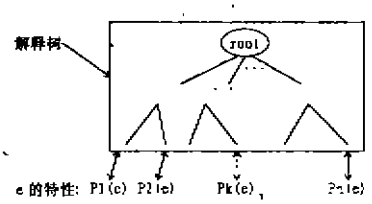

图1 中断的解释树

图中：$P_i(i=1,\cdots,n)$表示 e 的特性，实线箭头表示匹配，虚线箭头表示不匹配。可以看出，解释被中断的原因是解释树的叶子节点和实例 e 的某些特性不匹配。如果这样的特性不是很多或改动不很大的话，就可以对执行特性所对应的子树中的规则进行修改；否则应该重新构造规则。对规则的修改采用一般化算子，保留最好的修改。改动是不是很大，可根据具体情况给出标准。

前面所说的特殊化和一般化都是 ILP 中的方法。

当正例未被解释，可能实例只有少数的特性与解释树的叶节点不匹配，为了算法描述的方便给出下面的定义：

定义2(不正确路径) 设 T 是实例 e 的解释树，$P_i(e)$是 e 的某个特性，$P_i(e)$与 T 的叶子节点 $Leaf_i(T)$不匹配，从 $Leaf_i(T)$到 T 的根节点 Root(T)这条路径就称为 P(e)相对于 T 的不正确路径。

如果对错误路径以外的规则进行修改，由于这些规则里面没有与 P(e)有关的描述(即没有文字或项与 P(e)匹配)，所以只会对不同于 P(e)的其它特性(属性)产生影响，修改过后 P(e)仍然不能被匹配，而知识库则仍然不能解释 e；这与要求的最小语法修改是不一致的。由此就可以得到下面的命题：

命题 设正例 e 不能被知识库解释，T 是其解释树，P(e)是一个未被匹配的特性(属性)。如果要在知识库中进行修改以使 P(e)得到匹配，则只需对 P(e)相对于 T 的错误路径上的规则进行修改即可。

还需要说明的是，在上面的修正过程中有可能会有多个修改结果，所以都得选择最好的修改结果予以保留。对修改结果的评价有这样几个标准：①子句中前提的个数，即子句的复杂度；②子句覆盖正例的个数；③子句覆盖反例的个数。

如果要考虑学习结果和实例的一致性，那么应该要求每个子句所覆盖的反例个数为0。但如果我们要考虑学习的复杂性和有噪声数据的影响，就需要把子句覆盖反例的个数也作为评价标准加了进去。

下面给出生成解释树的具体算法：

```
输入：知识库 B 和某个实例 e
for B 中描述目标概念的每条规则
    运用 EBL 方法用 B 解释 e，这样就得到一棵树 T，如果
    树的所有叶节点和 e 的某个特性相匹配，则解释了 e。
    if e 被解释
       if e∈E⁻
       then 返回需要修改标志和解释树 T
       if e∈E⁺
       then 返回不需要修改标志
    end if
    if e∈E⁺
    then 将 e 相对于 T 的不正确的路径记录下来
endfor
if e∈E⁺
then 返回需要修改标志和所有不正确路径
```

解释完成后，运用 ILP 方法对需要修改的规则按上面的方法进行修改，就可逐渐改进理论。

3 一种集成学习模型 MIIE

通过上面的讨论可以得到集成 ILP 和 EBL 的一种模型，我们称之为 MIIE(Model for Integrating ILP and EBL)。这个模型可用下面的图2形象地表示出来。需要指出的是，在 MIIE 中，知识库的修改往往需要进行多次反复的修改，这点在图中没有明显地

标识出来。

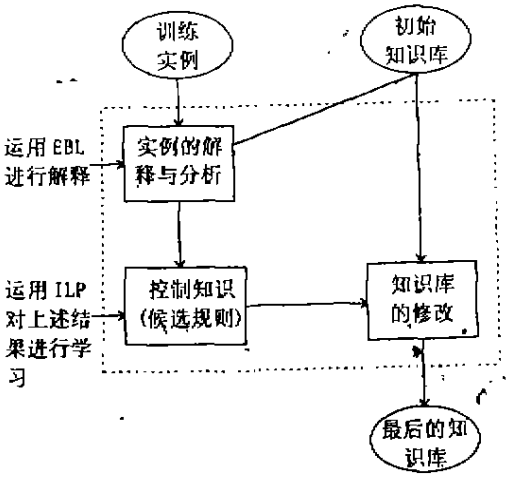

图2 ILP 和 EBL 的集成学习模型 MIIE

4 集成学习模型的应用

催化裂化(FCC)是石油炼制、石油化工中一个重要的过程，它不仅能生产高辛烷值汽油，而且能提供大量的各类芳烃。近几年来，由于各国对环境保护的要求日益提高，大大加快了汽油低铅化和无铅化进程，同时随着石油化工的迅速发展，国内外市场对苯类产品的需求增涨，催化裂化的地位得到日益提高。对炼油厂而言，催化裂化还能够提供纯度很高的富氢气体和液化气等副产品。但同时催化裂化的工艺是最复杂的，因操作异常或设备发生故障而造成的经济损失也是巨大的。为此中国石化总公司下达了重点科技攻关项目(编号389017，由洛阳石化公司负责，中国科学技术大学为主要参加单位)：《石油化工过程故障诊断专家系统工具》及《催化裂化反再故障先兆诊断专家系统》(以下简称 FCC 专家系统)，希望把专家系统的方法引入催化裂化领域，用计算机监视催化裂化过程，能够及时预报出故障，从而避免可能造成的巨大损失。

在 FCC 专家系统中用于反向推理的知识库共有11个，分别对应着11类故障。专家系统工具中的一个学习子系统——知识求精子系统，用于改进知识库的性能，提高它们对未来数据进行预报的能力。知识库和知识求精子系统的关系如图3所示，其中知识求精子系统应用了 MIIE 的思想和方法。

专家系统中的知识是用产生式规则表示的。现场数据则表示为"仪表名称　单位　仪表号　稳态值　瞬时值"，共有92个仪表，因此一组现场数据的数据量很大，限于篇幅在此只能对学习子系统的应用情况作一简要描述。

当某知识库需要修改时，知识求精子系统运用该知识库对上述数据进行解释，得到一棵不完全的解释树，然后对候选规则集进行修改。因为在这里，一般的叶子节点是比较式，是直接修改规则还是重新构造规则由对不等式的改动的大小和上面的标准共同决定的，比较式是某一变量(实际上是仪表号，如 T388是仪表"进料总管"的仪表号)的阈值，这是由专家给出的，如果需要对比较式进行太大的改动，则不能直接对规则进行修改，例如本来对 T388给的变化范围是150到300之间，假如为了能覆盖某个正例需要把比较式 T388＞150改为 T388＞300，显然这是不大可能的，这时应重新构造规则。

对知识库作了修改之后，知识库的性能得到了提高，从图4的比较中就可以看出这点。可见采用学习子系统可使"系统不断地进行学习，从而不断地自我完善，不断地提高诊断水平"(中国石化总公司鉴定语)。

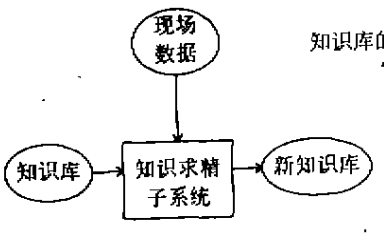

图3

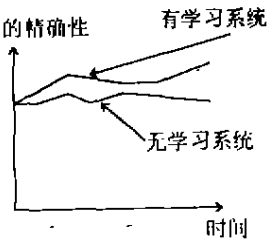

图4

参考文献

[1] Stephen Muggleton, Inductive Logic Programming: derivations, successes and shortcomings, SIGART Bulletin, 5(1)1995

[2] Saso Dzeroski and Nada Lavrac, Inductive Learning in Deductive Databases, IEEE Trans. on Knowledge and Data Engineering, 5(6)1993

[3] Ivan Bratko and Ross King, Applications of Inductive Logic Programming, Same to [1]

[4] Stephen Muggleton and Wray Buntine, Machine Invention of First-order Predicates by Inverting Resolution, Proc. 5th Intl. Conf. on ML, 1988

[5]《催化裂化反再故障先兆诊断专家系统》研制技术报告，中国石化总公司等，1995.11