基于归纳逻辑程序设计的学习方法及其实现的研究*

刘贵全 陈恩红 蔡庆生

(中国科学技术大学计算机系 合肥 230027)

摘要 归纳逻辑程序设计是机器学习领域中的一个新方法,它研究的是从实例和背景知识进行逻辑程序(新知识)的构造.本文介绍了归纳逻辑程序设计的基本理论和方法,并介绍了这种学习方法在专家系统中的应用情况.

关键词 归纳逻辑程序设计,机器学习,专家系统,有噪声数据.

中图法分类号 TP18

在机器学习中归纳学习较为成熟,应用也最广泛.近年来归纳学习的研究集中在从实例和背景知识构造目标概念的一阶逻辑定义,被称为归纳逻辑程序设计 ILP(inductive logic programming). ILP 发展很快,就其原因有 3 点.第 1,ILP 提供了从例子和相关背景知识中学习关系描述的较好方法;第 2,ILP 比属性-值表示具有更强的描述能力,属性-值学习只能使用固定的描述语言(其描述符是例子和背景知识中出现过的),而 ILP 则可以突破这个限制;第 3,ILP 具有较坚实的理论基础.在国外 ILP 已有不少成功的应用.[1~4]

早期 ILP 系统,象 MIS 和 CIGOL 都没有考虑数据有噪声的情况.后来的 FOIL 能处理有噪声数据,它是通过引入编码长度的启发式信息来实现的.[2]由于现在的 ILP 学习系统是面向实际应用的,而现实数据往往不是绝对准确,因此需要引入处理有噪声数据的机制.

本文介绍了 ILP 方法在专家系统的知识库维护与获取中的应用情况.在系统中,考虑到实际应用的复杂性,引入了对有噪声数据处理的机制.

1 归纳逻辑程序设计

所有的 ILP 问题都可描述如下[1]:给定

- 背景知识 B
- 概念描述语言 $\mathscr{L}$

* 本文研究得到国家自然科学基金和国家教委博士点基金资助.作者刘贵全,1970 年生,博士生,主要研究领域为机器发现,人工智能.陈恩红,1968 年生,博士,讲师,主要研究领域为机器学习,遗传算法,约束满足问题等.蔡庆生,1938 年生,教授,博士导师,主要研究领域为人工智能,机器学习,专家系统.

本文通讯联系人:刘贵全,合肥 230027,中国科学技术大学计算机系 Email:gqliu@cs.ustc.edu.cn

本文 1996-11-18 收到修改稿

·实例集 $E = E^+ \cup E^-$,其中 E^+ 和 E^- 分别表示正例和反例,满足条件:$B \wedge E \not\models \square$,$B \not\models E^+$要找猜想(规则或知识)$H$,$H$ 用概念描述语言 $\mathcal{L}$ 描述,满足约束条件:

①$B \wedge H \wedge E \not\models \square$($\square$表示失败);②$B \wedge H \not\models E^+$,即 $B \wedge H$ 能解释 E^+;③$B \wedge H \not\models E^-$,即 $B \wedge H$ 不能解释 E^-.

概念描述语言 $\mathcal{L}$ 又称为猜想语言,它是程序规则语言的子集.

2 学习系统

2.1 实例集的预处理

对实例集的预处理包括两个内容:①当反例没有显式给出时,生成反例集;②对实例的数据不完全的情况进行处理(比如在属性-值表示的情况下,缺少某个属性的值).

2.2 猜想的构造

完成了实例集的预处理之后,系统开始构造猜想.这里采用类 GOLEM 算法.

系统首先随机取若干对正例,对于每个这样的对,GOLEM 计算两个正例的共同特性,根据这些特性构造出一条对于两个例子都正确的规则.对所选的所有对都构造一条规则之后,找出其中最好的规则;规则好坏的标准是:①尽量少地覆盖反例(小于某给定值);②尽量多地覆盖正例;继续采用上述办法对未被覆盖的正例进行学习,直到结果不能再改进了.

2.3 有噪声数据的处理和学习结果的后处理

ILP 系统中规则的精确性作为一种启发式信息可以用于 3 个方面:①指导对规则的搜索;②作为结束的标准;③用于结果的后处理.

规则的精确性是根据规则对当前实例的分类精确性来进行估计的.FOIL 引入规则的编码长度作为启发式信息,其思想是将规则的长度限制在描述它所覆盖的正训练例所需要的编码的长度之内;但这种方法可能会使 FOIL 在没有噪声的情况下不能构造出完全的规则,而在有噪声的情况下构造出很特殊的规则.[2]比如,假设在有噪声的情况下有 1 023 个反例和 1 个正例,FOIL 将构造长度为 $\log_2(1\ 024) + \log_2(C(1\ 024,1)) = 20$ 的规则($C(m, n)$为从 m 中选 n 的组合数),但这种规则的用处往往不大.

鉴于上述原因,这里精确性的计算采用了概率估计的方法.对概率的估计可以有多种方法,象相关频率法、Laplace 估计法、m-估计法等.选择概率估计方法作为启发式信息对学习结果的影响比通常的精确方法的影响要大.因此要尽量采用简单可靠的方法.

下面给出用于 ILP 的度量规则性能的启发式信息.一种简单的度量规则 $C=T\leftarrow Q$ 性能的标准是其期望分类精确性($A(C)$),定义为一个正例被规则 C 覆盖的概率;

$$A(C) = p(+|C)$$

期望分类错误率定义为 $1-A(C)$.

由概率论知道,对 $A(C)$的估计越合理,学习结果也就越好.

由于相关频率法和 Laplace 估计法都要求例子是均匀分布的,但在一般情况下,这很难保证,故本文采用 m-估计法计算 $p(+|C)$:

$$p(+|C)=\frac{n^+(C)+m\times p_a(+)}{n(C)+m}$$

$p_a(+)$是先验概率,可用初始训练实例集中正例出现的频率 n^+/n 来估计,其中 n 是所

有训练实例的个数，n^+是训练实例中所有正例的个数. 这种方法的参数 m 是可调的，它的值根据噪声的大小而定；噪声越大，m 就越大.

采用 m-估计法的优点是能提供以贝叶斯分析为背景的扎实的理论基础，同时学习结果在对新的例子进行分类的时候能够更精确.

假设在系统中采用了一个精确性(覆盖的正例和反例的个数. 规则的复杂度等等)评价函数 A；

无关文字—有噪声数据情况：文字 L 称为规则 C 中无关文字，若 $A(C)\leqslant A(C-\{L\})$.

无关规则—有噪声数据情况：规则 C 称为 H 中无关规则，若 $A(H)\leqslant A(H-\{C\})$.

对学习结果的后处理就是在不影响结果精确性的前提下去掉无关规则和无关文字.

3 应用实例

3.1 在专家系统中的应用

专家系统中的知识库维护和知识获取一直是专家系统发展中的瓶颈问题，结合 ILP 与其它学习方法能较好地解决这个问题. 下面结合一个实际系统对此加以讨论. 本子系统着眼于提高知识库的效率及构造新规则(知识获取).

提高知识库的效率实际上是一个规则选择的问题，如果每次都能选择一条合适的规则进行执行以解决当前的子问题，那么系统在执行过程中不用进行回溯就可以得到正确的证明，而且不会得出错误的结果，这时知识库的效率将达到最高. 具体的作法是：①首先采用基于解释的学习方法运用知识库对训练实例进行解释(分析)，如果规则的执行导致不必要的回溯或错误的结论，则建立该规则的控制实例. 规则的一个正的控制实例集是整个执行过程中相应的规则必须实现的子目标. 相应地可以建立规则的反的控制实例集. ②应用 ILP 方法对所选规则的控制实例集进行学习，学习得到的控制规则将用于对该规则的控制. 假设规则 D 的控制实例表达的概念是 C，则控制规则的意思是"当 C 满足时应用 D". ③运用学到的控制规则对初始知识库进行修改.

如果规则改动不是很大的话，就可以直接对规则进行修改，然后保留最好的修改；如果需要做较大的修改或当知识库不能解释正例时，应该重新构造规则. 改动是不是很大可根据具体情况给出标准. 一般情况下应有下面两个评价标准：① 规则中前提的个数，即规则的复杂度；② 规则覆盖正例的个数. 构造规则采用的是前面提到的 ILP 的方法.

1988 年中国石化总公司下达了重点科技攻关项目(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》. 该项目已于 1995 年通过部级鉴定.

作为专家系统工具中的一个学习子系统——知识求精子系统，其目的是改进知识库的性能，提高它们对未来数据进行预报的能力. 知识求精子系统应用了上述的思想和方法. 该学习子系统可使"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"[5]；而无学习能力的系统的性能没有改善. 目前，专家系统及工具已在洛阳炼油厂投入使用，取得了比较好的效果.

3.2 对有噪声数据的处理

为了检验噪声对学习结果的影响，不同数量的噪声被加到了上面的数据中. 在下面，变

元A的$x\%$的噪声是指$x\%$的例子的变元A的值被替换为随机的值.当类的值引入噪声是它被看作附加的变元.引入的噪声的百分比取5%到80%的如下几个值:5%,10%,15%,20%,30%,50%和80%.背景知识中没有引入噪声.图1是实验结果.其中横轴表示噪声的百分比,纵轴表示知识库诊断的精确性(也用百分比表示).在噪声比较低的时候,结果是比较好的;而无噪声处理机制的系统的结果则要差得多.可见,这种方法能较好地处理有噪声数据.

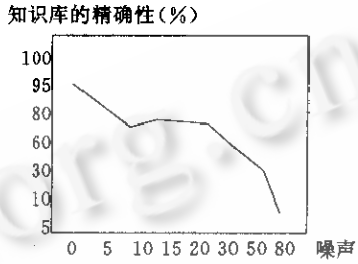

图1 例子中加入噪声的实验结果

4 结　论

本文介绍了一个学习系统及其在专家系统中的应用情况.通过应用可以看出:①系统能较好的解决知识库中的知识维护和知识获取问题.②采用了有噪声数据的处理方法后能较好地处理有噪声数据.这说明学习系统完全可以应用于现实数据中.这对数据库中的知识发现的研究和机器学习以及其它领域的科学研究都是一个良好的开端.

参考文献

1 Stephen M. Inductive logic programming: derivations, successes and shortcomings. SIGART Bulletin, 1995,5(1):5～11.
2 Saso D, Nada L. Inductive learning in deductive databases. IEEE Trans. Knowledge and Data Engineering, 1993, 5(6):939～949.
3 Ivan B, Ross K. Applications of inductive logic programming. SIGART Bulletin, 1995,5(1):43～49.
4 Ross D K, Steppen M, Richard A L *et al*. Drug design by machine learning: the use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. Proc. Natl. Acad. Sci., USA, 1992,89:11322～11326.
5 中国石化总公司等.《催化裂化反再故障先兆诊断专家系统》研制技术报告.1995.

RESEARCH ON INDUCTIVE LOGIC PROGRAMMING BASED LEARNING METHOD AND ITS REALIZATION

LIU Guiquan　CHEN Enhong　CAI Qingsheng

(*Department of Computer Science and Technology　University of Science and Technology of China　Hefei*　230027)

Abstract　ILP(inductive logic programming) is a newly developed field within machine learning. It focuses on the problem of constructing concept definitions from the examples and the background knowledge. This paper studies the learning method based on ILP and its application in expert systems.

Key words　Inductive logic programming, machine learning, expert system, noise data.
Class number　TP18