

基于归纳的机器学习方法的实现研究

69-74 罗 翊 王克宏 (清华大学计算机系, 北京100084) TP18

摘 要

This paper analyzes the method of inductive learning and emphasizes the importance of knowledge during the inductive processes. From this view, we devise an inductive-based machine learning model LEARNER/TH-1 and an inductive algorithm constructing step by step. In this model, the system uses the domain knowledge on a large scale, and the inducted cases are looked upon as a kind of experienced knowledge. The background knowledge, the experienced knowledge and the inductive conclusion have the same representation form. The learning process is self-drive.

一、引 言

机器学习是解决计算机获取知识问题的方法。在这项研究中,对人类学习方法的模拟是一条重要的途径,而归纳正是人类学习的重要方式之一。

从归纳中获取知识是人类新知识的极其重要的来源,它通过对经验材料的研究,探求新知识^[1]。由于一切科学理论的基本前提最初都是从经验事实中总结出来的,而从经验到一般规律的过程中,归纳推理起着不可替代的作用。开普勒从各大行星的天文观测数据中发现三大运动定律,达尔文从对生物物种的观测中提出自然选择学说,他们都利用了归纳方法。

实现基于归纳的机器学习会产生巨大的影响。当前机器学习的研究取得不少进展,传授学习、示例学习等方法在一些系统中都已投入实用,在帮助获取专家知识方面起到很好的作用。但现有这些方法有着共同的特点,就是它们要求专家对自己的知识具有自觉,即专家已经知道有这样的知识,并能把它形式化。而专家的能力往往在于他多年积累的经验,这些经验可称为“隐式知识”(属

“内涵逻辑”范畴)。这种经验也许难以明白地表达出来,甚至只是一种直觉^[12]。因此提取这种知识是很难的。考虑到这种经验的来源是实践,实际上是一种对大量事例、现象的归纳结论,因此可用归纳学习的方法来获取这种知识。

具有更大意义的是归纳学习可以从大量数据中发现新的规律,而这些规律甚至还不为人知。当然,机器的归纳学习本身不能保证这种规律的正确性,但它至少可以给人提供启迪。这是一种机器发现学习。现在美国的Blum、Ferra等在从事从数据库中 提取知识的研究,并取得了一些进展^[2]。

因此,对归纳学习的研究具有很深远的意义,也有现实的价值。

本文将对归纳学习进行分析,将归纳与演绎相结合,并置于领域知识体系中,按此提出一个基于归纳的机器学习系统模型,并提供一个初步的实现方案。

二、归纳学习分析

2.1 基于归纳的机器学习与示例学习的关系

基于归纳的机器学习从特殊的经验事例

中提取一般规律, 示例学习可说是它的一种特殊情况。在示例学习中, 教师需要提供一系列的正例、反例, 然后由学习系统根据各示例形式上的相似来寻找共同规律。

在示例学习中, 要设计一组恰当的示例需要不少精力。对于数据驱动的示例学习, 教师要考虑示例提交的次序^[3], 在模型驱动学习中可以免去这种麻烦, 但仍存在情形识别等问题^[4]。教师提供的示例好坏直接影响学习的成效。因此示例学习对教师的依赖程度相当大。它只具有归纳的形式而已, 是传授知识的方法而不是主动获取知识的方法。

基于归纳的学习注重从过去的经验积累中发现规律, 乃至更进一步有目的地获取有关事实来形成、检验规律。这些经验是在长期实践中逐渐积累的, 包括系统自己的运行记录和专家的工作案例。它是一种不需要依赖教师设计示例的学习方法。

2.2 归纳的作用

对智能系统而言, 归纳学习有两方面的作用: 1) 减少信息占用的存储空间, 但同时不减少实际拥有的知识量。对人类而言, 这可能并不重要, 因为人的记忆潜力非常大, 但对现在的计算机系统而言, 减少空间占有量具有重大的意义, 在减少空间占用量的同时也意味着时间开销的下降。2) 发现事物间的规律, 并用以指导以后的行动。这时, 归纳结论比前提事例具有更大的蕴涵范围。未被事例覆盖的区域是造成归纳结论可能不真的因素, 但也正是它才给归纳结论以更大的意义, 可以起到发现新知识的作用。

2.3 归纳中知识的作用

当前示例学习的研究中, 重点主要放在对材料提取相似性的方法上, 形式化的有扩张矩阵和决策树^[10], 非形式化的有强化-弱化策略^[4]。这些方法在减少信息占用空间方面还是有效的, 例如可将数十个示例总结为一条规则, 这条规则符合所有示例。但这条规则的含义是什么就难以保证, 即规则不一定具有可理解性。经过学习得到的知识应是

人们所能理解和加以解释的, 这样才能给人类增加新知识^[6]。只由示例间的表面相似而得出的结论, 其可理解性不能由学习过程加以保证。极端情况下, 甚至可以从随机数据中找出一条规律, 如在[6]中提到的学习扑克牌牌点的规则。这种纯形式的学习对于作出假说是非常必要的, 但若只限于此, 则归纳学习系统的功能就不完全, 学习本身就是非智能的。

究其原因, 是只重形式的归纳造成, 科学知识的增长不是单靠归纳或演绎就可得到的, 而应将两者结合起来^[7]。在归纳学习中的演绎实质是一种对已有知识的运用。在归纳过程中, 不仅要对被归纳的若干事例进行形式处理, 还要运用有关的知识进行分析、综合, 获得有意义的结论。否则, 只凭表面的相似性, 很容易得出错误的结论。例如在科学史上曾经流行一时的“热质说”, 也是根据一些经验归纳出来的。但在遇到“摩擦生热”的现象后, 只凭形式的归纳就无法得出正确的解释。这时只有依靠演绎, 运用与各种热现象的原因有关的知识, 如摩擦生热是由于机械运动, 燃烧生热的原因是化学反应即分子的运动, 等等, 形成新的归纳基础, 再从这些事例中总结出规律: 热是一种运动, 从而是能量的一种表现形式。在这个发现实例中, 归纳与演绎的结合不可缺少。

因此, 要进行智能化归纳学习, 必须运用已有知识, 将学习过程置于一个领域知识体系中, 用领域知识来引导学习和评价学习。

在没有教师提供精心设计的示例的情况下, 学习系统的知识库或事例库中的事例很可能并不直接反映本质因素, 因此在这种学习系统中用领域知识处理事例就有着特别重要的意义。例如一个化学实验解释系统, 要从一系列化学实验的现象描述中归纳出其中某种化学物质的性质和组成。这种学习要从表面现象探究深层的原因, 在归纳中必须依靠已有知识。

在机器学习研究中,对如何利用领域知识来帮助归纳学习,加快学习进程,加深学习内涵的研究还不多。SRI的M.P.Georgeff与Monash大学的C.S.Wallace在[8]中提出用知识表达语言本身的某些限制来作为对领域知识的表达,从而使学习结论体现出领域特性。此方法有很大局限性,因为对不同的领域需要不同的描述语言。在[9]中,洪家荣教授描述了在AE5学习系统中的构造性学习模块,该模块可以根据背景知识,从例子集中产生例子中不具有的新概念、属性、变元等,然后产生基于新概念的规则。这是一个利用领域知识的很好的尝试。

总的来说,当前在基于归纳的机器学习中利用知识都还是部分的、局部的,与知识在人类归纳学习中的重要作用不相称。

在知识工程中,人们说知识就是力量,知识库的优劣决定系统性能的好坏。同样地,一个学习系统的学习能力也应随着系统中具有的知识增加而增强。利用系统已有知识是提高学习能力的重要途径。

三、基于归纳的机器学习系统模型

3.1 LEARNER/TH-I的学习过程

根据上文的论述,可将归纳学习形式化地表述为:

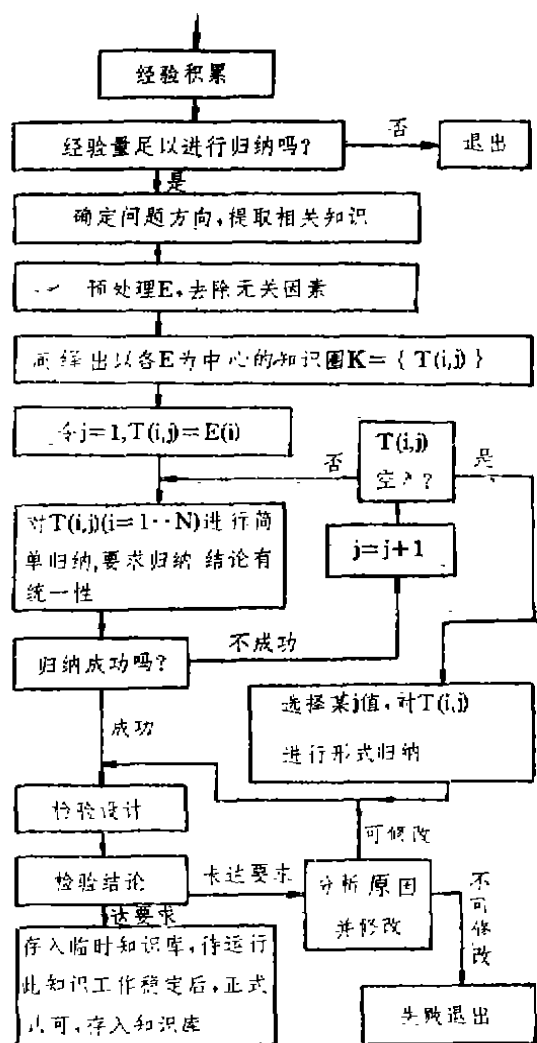
$$K + E \rightarrow K' + R$$

其中K是已有知识,E是*i*个经验事例的集合。经过学习后得到规则R,作为对E的概括,同时由于新的知识会影响到原有知识,对原有知识进行补充、修改,所以K变为K'。

从上文的分析出发,我们提出一个基于归纳的机器学习系统模型LEARNER/TH-I,它的完整的归纳学习过程如下框图所示。

各主要步骤说明如下:

1) **经验积累**: 因为归纳学习不依赖于教师,所以必须自己在运行实践中积累实例,包括运行记录和专家工作实例。这种实例可置于专门的实例库中,也可置于领域知识库中,并与其它知识一样被使用。在某一类相



关事例积累到一定数量时,即可开始归纳过程。

2) **确定方向,提取知识**: 在相关事例中确定归纳对象,即归纳结论中欲说明的主体,从知识库检索与此对象有关的知识以及各事例的背景知识。

3) **预处理**: 根据归纳对象领域特点,将事例中的无关条件删去。这种无关条件主要是不能与归纳对象形成因果链的因素。在有些情况下,这一步骤应取消,例如由统计数据寻找因素间的相关性时,相关因素间也许并没有因果关系。

4) **演绎形成知识圈**: 与E有关的背景知识组成 $K(i)=\{T(i,j)\}$,叫作知识圈。其中

$T(i, j)$ 当 $j=1$ 到 N 增大时, 它与 $E(i)$ 的推理路径是逐渐增长的, 即 $T(i, 1)=E(i)$, $T(i, 2)$ 与 $E(i)$ 直接有联系, $T(i, 3)$ 只与 $T(i, 2)$ 直接联系。以产生式规则为例, 若 $E(i)$ 为 $A, B \rightarrow C$, $T(i, 2)$ 为 $D, E \rightarrow A$, $T(i, 3)$ 则可能是 $F \rightarrow D$ 。由于 $T(i, j)$ 中的知识量可能很多, 所以应对 $T(i, j)$ 的层次 j 的数目加以限制, 以提高效率。

5) **归纳**: 首先, 由于自然界的齐一原则^[11], 我们进行归纳时总是趋向于用一个统一的规则: 相同的事件应有相同的原因, 即使事件的表面现象千差万别。因此, 归纳从最简单的假说形式开始, 即单纯的合取式, 寻找在所有事例中都出现的因素。这一因素可能在 $E(i)$ 中并不表现出来, 而只存在于 $T(i, j)$ 中。例如: 有 $E(1): A \rightarrow R$, $E(2): B \rightarrow R$ 。只由 $E(1), E(2)$ 得不到有意义的结论。但我们还有 $T(1, 2): X, Y, Z \rightarrow A$, $T(2, 2): X, U, V \rightarrow B$ 。这样, 利用 $T(i, j)$ 的知识, 我们可以归纳出 $X \rightarrow R$ 。这就是要从内向外逐层对 $T(i, j)$ 进行归纳的原因。

当然, 利用 K 并不一定能保证得到这样的规则, 因此在利用背景知识的尝试失败后, 说明领域知识还不足, 此时应采取形式化的归纳学习方法从形式的相似性上得到具有一定性能的规则, 用‘与或’式表示。

6) **检验设计**: 因为学到的结论可能是难以直接验证的, 类似于解释学习中的不可操作的规则。此时可利用领域知识, 进行演绎, 推出易于加以检验的推论。推论要具有规则所允许的最大极端性, 以便使规则的正确性得到最严格的考验, 同时也为修改规则提供参考。

7) **分析修改**: 根据推论中的极端性, 逐步缩小规则条件范围。

这个学习过程中大量利用已有领域知识, 因此很适宜于在知识丰富的环境下工作。同时, 由于保留形式归纳, 它也可以工作在知识不多的情况下, 如系统初建之时。

3.2 归纳算法

当前归纳推理研究的主要成就是统计推理, 它是现在已知的纯科学的可实际应用的归纳推理^[11]。应用统计推理能较精确地确定一个归纳结论的可信度。但统计推理并未给出如何构造一个归纳假说的方法。因此, 本文提出一种逐步构造式的归纳算法。

人类的认识规律是由简至繁, 在做出科学发现的过程中, 也都是先做出一些简单的、不完全的假说, 再逐步完善之。在机器归纳学习中, 对这种归纳方式进行模拟, 就是逐步构造式学习。

算法的主要思想是: 先形成简单的假说, 然后用统计方法做出评价, 将达到要求的假说输出到规则集中, 同时淘汰未达标准的假说。若总体性能未达到要求, 则对假说集中的规则进行合并、组合, 形成下一代假说。继续上述过程, 直到规则集的整体性能达到要求。

本算法区分有、无对照例集 NE 的情况, 有略微不同的执行步骤。这里只介绍较复杂的、有 NE 时的步骤。这里的对照例不同于示例学习中的反例。对照例中的主体必须与被归纳对象同属一类。例如, 在考虑乌鸦的属性时, 其对照例的主体可以是其它鸟类, 但不能为非鸟类的其它动物。因为只有同类事物的对比才有意义。在这种意义上, NE 中事例被称为对照例而不是反例。对 NE 作如此规定, 只是为了在系统自己确定归纳主体后再寻找参照时不至于无所适从。

算法的描述如下:

(1) 从事例集 PE , 对照例集 NE 分别概括建立两个假说集 $C1, C2$ 。假说集中的规则具有最简单的条件即只涉及到一个因素, 但分别具有一定高的性能得分; $P1$ 是规则对 PE 的覆盖率, $P2$ 是对 NE 的覆盖率。 $C1$ 具有高的 $P1$ 值, $P1 \geq G1$ 。 $C2$ 具有低的 $P2$ 值, $P2 \leq G2$ 。 $G1, G2$ 是两个预定的阈值。

(2) 将 $C1, C2$ 分别复制到 $S1, S2$ 中, 如果有 $P2 \leq G3$ 的规则, 将它输出到 R 中。

G_3 是一个接近于零的阈值。

(3) C_1 与 S_1 、 C_1 与 S_2 、 C_2 与 S_1 间进行条件的合并,产生新规则。统计各新规则的性能评分,删去其中性能评分未达要求的规则。以 $P_1 \geq G_1$ 的新规则代替 S_1 的旧规则,以 $P_2 \leq G_2$ 的新规则代替 S_2 的旧规则。

(4) 若 S_1 、 S_2 为空,则失败退出。

(5) 对 S_1 、 S_2 中的规则进行局部优化,以优化后的规则代替原规则。

(6) 若 S_1 、 S_2 中没有 $P_2 \leq G_3$ 的规则,则转向(3)。

(7) 将 $P_2 \leq G_2$ 的规则输出到 R 中,检查 R 对 PE 的覆盖率,若未达到要求则转向(3)。

(8) 以 R 作为归纳结论,退出。

在算法中, R 中的各规则间是“或”关系。因此,本算法实际上通过不断对与或式中的或项进行特殊化(增加合取项)来进行学习。而一般化的过程则体现在初始规则集的生成以及规则的局部优化中。

根据归纳结果是单一合取式的可能性的,调整各个阈值 G 的大小,可控制算法运行,提高效率。若是单一合取式的可能性大,则可将 G_1 提高,反之,若“或”项数可能较大时,可将 G_1 以及 G_2 降低。

在对 $T(i, j)$ 进行简单归纳时,因为要求获得合取式,所以令 $G_1 = 1.0$,若考虑噪音干扰, G_1 也可以比1.0略小一点。进行形式归纳时,再将 G_1 降低。

这一算法与遗传学自适应算法^[6]有近似之处,即都采用进化、选择机制。但本算法对规则进行定向进化,因此规则改进速度优于遗传学算法。

四、结论

4.1 比较

本文将归纳看成是在一个相关联的知识体系中进行,不是孤立地寻求相似性。因此在学习过程中大量利用领域知识,这是它不同于其它归纳学习系统的最大特点。由于运用领域知识,很多归纳结论可以具有简单的

表达形式,即用一个单一的与条件表达式来代替直接进行归纳所得到的‘与或’表达式。同时这样得到的归纳结论一般也更能反映归纳对象的本质。

文中提出的基于归纳的机器学习,其归纳基础从系统的知识库中取出,不需教师的特别指导,具有自主的特点。除开从数据库利用统计来提取相关性的学习系统外,现有各归纳学习系统尚无自己寻找事例的功能。而本系统LEARNER/TH-1从知识库中寻找事例,或说将事例作为一种直接的经验知识来使用,这又是数据库学习系统所不具有的。另外数据库学习系统也无法给出事例中的因果关系。

4.2 尚待研究的问题

归纳中,归纳结论可能要修改已有知识,与已有知识发生矛盾,即出现非一致性问题。如何识别这种情形与归纳错误之间的区别,如何对已有知识作出适当的修改、进行一致性处理,是一个很重要的问题。只有解决这个问题才能建立真正的自增殖式学习系统。

对已有知识的利用方式是多种多样的,文中提到的方法只是探讨与试验,如何以更好的方式利用领域知识,提高学习效率,也有待进一步深入研究。

从事例中学习元知识(包括控制学习过程本身的元知识)是提高系统性能的重要途径。这是因为对领域知识的使用方法是体现专家能力的重要方面,而通过学习提高学习的能力将会使系统的独立工作能力大大提高。基于归纳的学习系统还没有这种功能,因为这与知识处理密切相关,需要从两方面作工作,还有待在以后的工作中实现之。

参考文献

- [1] 黄菊丽,归纳法在逻辑中的作用与地位,《归纳逻辑》(北京市逻辑学会第一届归纳逻辑研讨会文集),中国人民大学出版社,1986
- [2] F.M.Ferra, SOCRATES, A System to Extract Rules and Knowledge from

实例学习与解释学习的结合

74-78

唐 雁 邱玉辉 (西南师范大学计算机科学系, 重庆630715) TP18

摘

要

实例学习和基于解释的学习是机器学习领域中具有代表性的、研究得最为深刻的两种学习方法,但由于这两种学习方法都存在一定缺陷,使它们在实际应用中受到较大的限制。在关于人类概念形成的心理学理论的基础上,本文讨论一个基于实例和解释的学习模型,将这两种学习方法有机结合在一起,一方面使用领域知识指导归纳学习过程;另一方面用归纳学习弥补领域知识之不完善。其学习机理更加接近人类学习的认知过程——抽象思维过程,取得了较好的学习效果。

一、引 言

模拟人类学习过程的机器学习,其出路之一就是研制采用综合推理机制的机器学习系统。实例学习从一系列正例和反例归纳出概念定义;基于解释的学习运用领域理论对单一实例进行演绎得出概念定义。由于这两种学习方法都难以依靠自身的力量克服其局限性,我们开始考虑将两种学习方法结合起来,以期获得更好的学习效果和学习方法。

通过对实例学习和基于解释的学习方法的分析和评价,我们认识到这两种学习方法

并不是完全对立的,它们在许多方面相互补充。一方面,由于实例学习只能用统计信息作为搜索的启发性知识,难以区分重要的、次要的和无关的信息,因此常常产生归纳跳步问题;而基于解释的学习以领域理论作为启发性知识,不会发生归纳跳步问题;另一方面,领域理论的不完善限制了基于解释的学习方法在复杂领域的应用,如果用实例学习的结果修正和补充领域理论,则可以使不完善领域理论给学习系统带来的影响大大减少,而且,基于解释的学习只用一个实

Exiting Database, SIGART newsletter, Apr. 1988

- [3] 石纯一,王家敏,实例学习,中国计算机用户,16,1988
- [4] 吴信东,《专家系统设计》,中国科技大学出版社,1990
- [5] R.S.Michalski, A Theory and Methodology of Inductive Learning Machine Learning, Sping-Verla, 1984.
- [6] Richard Forsyth, Expert Systems: Principle and Case Studies, Chapman and Hall, 1984
- [7] 赵承壁,归纳法和归纳法问题,《归纳逻辑》,

中国人民大学出版社,1986

- [8] M.P.Georgeff, C.S.Wallace, A General Selection Criterion for Inductive Learning, ECCAI-85
- [9] 洪家荣,示例式学习及多功能学习系统 AE5,计算机学报,2,1989
- [10] 洪家荣,刘宁,示例式学习的两个方法及其应用,计算机科学,6,1990
- [11] 江天骥,《归纳逻辑导论》,湖南人民出版社,1987
- [12] 戴汝为,王钰,综合各种模型的专家系统设计,模式识别与人工智能,3,1989