人工智能 知识获取 数据库

44-47

计算机科学1993 Vol.20№.5

TP18

知识获取和数据库学习系统*)

李德毅

(中国人民解放军总参谋部第61研究所 北京100039)

摘 要

The difficulties faced in the study of expert systems, such as knowledge acquisition, knowledge representation and elementary Knowledge absencing, are analysed in this paper. Considering the fact that the database technique and its applications have been very fruitful at present, the machine learning from databases by means of the representation and operation tools of relational databases could be a good idea as a starting point of learning. This paper introduces the essence of this new strategy.

一、专家系统的困惑

从人工智能的角度看，专家系统实质上是一个问题求解系统，目前的主要理论工具是基于一阶谓词演算的机器定理证明技术——一阶演绎系统。领域专家长期以来面向一个特定领域的经验世界，通过人脑的思维活动积累了大量的有用信息。在研制一个专家系统时，知识工程师首先要从领域专家那里获取知识，这一过程实质上是个归纳过程，是非常复杂的个人到个人之间的交互过程，有很强的个性和随机性，没有统一的办法。因此，知识获取成为专家系统研究中公认的瓶颈问题。其次，知识工程师在整理表达从领域专家那里获得的知识时，用if-then等类的规则表达，约束性太大，用常规数理逻辑来表达社会现象和人的思维活动太局限，也太习难，勉强抽象出来的规则有很强的工艺色彩，差异性极大，知识表示又成为一大难题。此外，即使某个领域知识通过一定的手段获取了，表达了，但是这样做成的专家系统对常识和百科知识出奇地贫乏，而人类专家的知识是以拥有大量的常识作为基础的。人工智能学家Feigenbaum估计，一般人拥有的常识存入计算机大约有100万条事实和抽象经验法则，离开常识的专家系统有时会比傻子还傻。例如战场指挥员会根据"在某地发现一只刚死的波斯猫"的情报很快断定敌高级指挥所的位置，而再好的军事专家系统也难以顾全到如此的信息。

以上这三大难题大大限制了专家系统的应用，使得专家系统目前还停留在构造诸如发动机故障诊断一类的水平上。

*)国家自然科学基金资助项目课题。李德毅 博士，高级工程师。

法加入对CN进行修补(扩张或缩小)，引入逐步求精过程等。AILM的设计方法是既考虑单个学习模块的封装（独立性），又考虑整个系统的一致性。目前正在Sun工作站上实现，进展顺利。

五、结束语

集成各种学习方法的系统必将在更大程度上提高机器的智能，集成学习工具比单一学习方法具有更广泛的应用前景，所以这是一个很值得研究的课题。我们提出的AILM代表了一种集成系统的设计思路，期望起到抛砖引玉的作用。（参考文献共9篇略）

二、用数据库作为知识源的新策略

既然由人工智能工程师从领域专家那里获取知识的道路不具备普遍意义，而要找到一种完美的、理论上严密的知识获取技术还有相当长的道路，甚至根本找不到。那么，对于要研究的一个特定的客观世界（领域）来说，能否把知识获取分为两大步走：先利用一种成熟的技术和方法搜集、存贮并管理该领域的大量信息，然后在这基础上发现并获取知识。由于第一步的工作已经把问题形式化、规范化了，这就限制了知识获取的范围，也大大提高了知识获取的起点。当前广泛应用的数据库正是实现前一步的最好选择。

(1) 关系数据库具有归一化的组织结构形式，关系之间的平等性、属性之间的平等性，便于知识发现过程中的并行运算。在知识发现过程中需要对数据的操作全都是读操作，而发现方法和发现工具箱的使用也都可以同时进行，这就为在网络计算 (Network Computing) 环境和并行机上发现知识提供了广阔前景，为解决机器学习效率找到了出路。

(2) SQL具有第四代语言的重要特征，从对数据库进行随机查询到数据库的管理和程序设计，几乎所无不能。在把数据库作为知识源进行知识发现研究时可以充分应用，甚至可以充分借用SQL语言能力。例如借用关系表表达有概念提升能力的自上而下的概念树，向面向对象拓广，还可以沿用关系操作符对知识段进行广义运算。

(3)关系数据库和一阶谓词逻辑具有极好的对应性。业已证明[1]，用产生式系统表示关系数据库，用一阶谓词逻辑运算可以很好地表达各种关系运算，而对数据库的查询可以认为是一个特定推理的演绎证明。这就为知识发现过程中实现以归纳为主，归纳和演绎相结合打下了基础，用演绎的方法对发现的知识进行正确性证明。

(4) 在类似于数据基的知识基中，如果把元组看成是抽样后的一个个示例，示例中有正例、反例、特例和噪声，则构成示例学习；如果把属性看成是一个个特征，用以确定它们之间的相关度，则可构成概念与发现学习；对元组（或属性）进行事例间的类比，又构成了类比学习。

(5) 借用数据字典形式、E-R图手段和人-机界面技术，描述待发现的知识的常识或有关概念之间的背景，也显得十分自然和贴切。

由此看来，专家系统研究中的三大困惑在把数据库作为知识源的新策略中找到了回答。如果这样的知识发现系统能够在给定的数据库基础上，合理而有效地给出该数据库中虽不直接包含、但和这一数据库总情况和总趋势相吻合的新的数据解答，那么就自然地构成了一个数据库学习系统。

三、归纳状态空间理论和发现工具箱

以数据库中的事实为依据，从大量的特殊性或个别性中抽象出一般性、规律性的东西，并揭示事物的内在本质及其联系，这样一个归纳过程想一次性完成是注定要失败的，企图把这个过程通过建立形式系统实现，追求理论上的完备性，也是行不通的，想从同一数据库中归纳发现永恒的、对大家都一致的所有知识，也是不现实的。其原因是它们都违反了人类认识论的过程，违反了人类借助自然语言来表达思维的过程，违反了不同人出于不同目的对相同客观世界在归纳问题上的差异性。

基于以上认识，我们首先要求明确待学习的任务，从数据库中有针对性地获取相关知识。提出学习任务，反映了提问者的立场。问题常常由问式和题设两部分组成。提问者虽然没有断定什么，但题设中常常隐含着一些预设，它大大压缩了知识获取的范围，可以借用SQL语言和关系运算，把仅仅和题设相关的数据集中到一个视图中来，称之为最初始的由元组组成的数据基。同时还要借助数据字典，充分获取题设中涉及到的

概念和背景知识。

其次，在数据基上进行抽样生成知识基。抽样的原则就是要保持数据基所反映事物的整体和本质的特征，抽样的方法有：纯随机抽样（简单随机抽样）、机械抽样（等距抽样）、类型抽样、整群抽样、分层抽样等。这些方法可以单独使用，也可以结合使用，从而大大提高知识获取的效率。这种经过高度压缩而又能反映整体特征的由宏元组组成的知识基仍然以二维表的形式存在，成为知识发现的基础。

具有平面特性的这种知识基随着归纳的深入（抽象级的提高）不停地交替、浓缩，逐步由微观走向中观、宏观，每一个中间状态的知识基都对应着一个归纳空间表，它是在这一抽象级别上知识和知识属性的几何表示。整个归纳过程形成归纳状态空间。

假设当前归纳状态空间表共有n个知识属性，第i个属性的允许取值个数为$v_i(1\leqslant i\leqslant h)$，实际取值个数为$d_i$，则由定义全部知识空间值为：

$$P_{entire}=\prod_{i=1}^{n}v_i$$

又设现有各元组占用空间值为P_{occupy}，则定义归纳状态空间表的稀疏度S为：

$$S=P_{entire}-P_{occupy}$$

我们还定义归纳状态空间表的杂复度C为：

$$C=\sum_{i=1}^{n}d_i$$

注意，式中d_i必须小于v_i。若实际情况是$d_i=v_i$，则意味着该域的实际值已经包含了所有可能值，那么这一知识特征对进一步归纳将不再有贡献，此时强制置$d_i=0$。

以上稀疏度和复杂度的定义还可以用于归纳状态空间表的部分表——子空间表，从而比较并选择归纳的最佳入口点和归并方式，即知识融合（Knowledge fusion）方式。

低的稀疏度意味着每条知识内部的高相似性和凝聚力，也意味着不同知识之间的低相似性和耦合力；低复杂度意味着这条知识可以用一个简单的形式表示，因此我们总是追求低稀疏度和低复杂度的最好配合。

这样一来，归纳知识的抽象度η定义为：

$$\eta=f(S,\ C)$$

最简单的方法可以令$\eta=S+C$；也可以简单加权定义$\eta=W_1S+W_2C$，$(W_1+W_2=1)$它反映了从数据中发现知识的质量准则。建立了这样的准则之后，我们可以确定归纳状态空间表沿着什么样的机制跳跃，并得到适可而止的宏观知识。

在对归纳空间表进行概念提升、匹配、替换、剪枝等操作时，要用到很多工具，这些工具统统存放在发现工具箱中，比较典型的工具有：统计、归格化、过滤、排序、归约、分类、聚类、找同、找差、找反、找异常、找种子（典型、中心）、找相关、找派生、找因果、找方程、找趋势（变化）等。其中，特别是统计和聚类工具显得更为重要。这里还要强调两点：

1．并不是在每一条知识的形成中所有工具都必然要用到，归纳状态空间表中的知识属性有的是文字型，有的是数字型。文字型属性有的可以进行定序、定距或定比分析，有的也许只能作定类比较，即仅起标称作用，实质上它们是无序的。

2．要尽可能借用现成的工具，或者稍加改造即可，充分利用软件模块化设计功能，使之能方便地加入工具箱。

四、数据库学习系统构成

数据库学习系统的结构组成见图1。由图可以看出，通过题设对数据字典和实数据库的访问生成数据基，经抽样后形成知识基，随着对归纳状态空间表的浓缩，映射出新的结构关系形式，即概括出带有一般性的“知识”来，这时这些“知识”可称为假设，通过知识验证系统，对假设进行演绎推理，在实数据库的支持下得到证明，或者修改假设再进一步证明，得到归纳后的知识。如果没

有演绎推理的配合，就不可能实现认识的归纳过程。

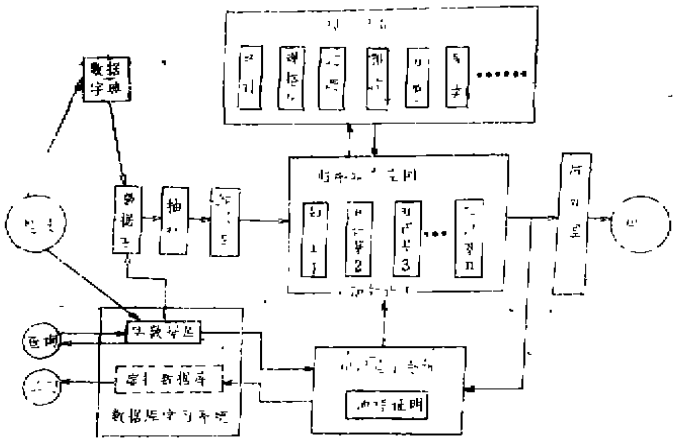

图1

利用知识验证系统，还可以针对特定查询，依靠归纳获得的知识，演绎生成虚拟数据库，即和实数据库整体相吻合但根本不存在的数据，形成数据库学习系统。

在从数据库中发现知识的过程中，不仅要归纳、提练某一类对象是否具有给定的性质（或关系），而且要判断在某一范围内具有给定性质的对象的全体具有什么样的量的特征，由全称量词（∀）和存在量词（∃）组成的一阶谓词逻辑是一种典型的量化逻辑，但人类的实际思维过程中远远超过了这两种量化形式，为了能判断并论证论域D中具有性质P的对象有多少，是多数还是少数，还是许许多多，相当多或极少极少，在归纳和演绎过程中，我们建立了复量化逻辑体系，引入模糊语言值，如"少部分"、"基本上（原则上）"、"大多数"、"几乎都"等，进行似然推理。语言方法的引入是对思维和感知中不精确性的普遍承认，在定量基础上的定性归纳能深刻地反映问题的本质，能用较少的代价传送足够的信息，对复杂事物作出高效率的判断和推理。利用它对发现知识进行后处理，语言值的引入增加了知识的弹性，不但使从数据库中最终获得的知识更加健壮，而且更易被人们理解。

五、结束语

数据库技术已经成为信息处理的基石，人们面临着信息爆炸的挑战，而当前数据库最实质的应用仅仅是检索。为了避免信息资源的浪费日趋严重的现象，也为了克服专家系统研究中遇到的三大难题，努力把数据库的应用提高到决策支持的高度正是机器学习的最佳突破口。一个按照本文思路的实验系统正在进行之中，在局域网上的数据库服务器中输入几百个典型犯罪案例，希望能产生若干刑法条文，并据此对新案例进行自动判刑；还希望能通过一个地区的人口数据库得出有助于计划生育政策的种种决策；通过一个商品数据库发现有利于价格调整的知识，等等。总之，应用前景是极其广阔和迷人的，我们将在追求阶段成果局部实用的情况下不断深化研究。

参考文献

[1] Deyi Li, «A PROLOG Database System», 1984, RSP, England

[2] W.J.Frawley, Knowledge Discovery in Databases, An Overview, AAAI, Press/The MIT Press, London, England, 1991

[3] 李德毅、杨雪南，关系数据库中的知识发现研究，小型微型计算机系统，Vol.13, No.4, 1992

[4] Proceedings of 1991 AAAI Workshop on Knowledge Discovery in Databases, Anaheim, CA, July, 4-15, 1991

[5] Deyi Li, «A Fuzzy PROLOG Database system», RSP, England, 1990