4—6 知识获取 商品化 产业化 (2)

计算机科学1993Vol.20№.3

促进知识自动获取的商品化与产业化

洪家荣 (哈尔滨工业大学计算机科学与工程系 哈尔滨150006) TP18

摘 要

Recently, machine learning is the most active area of artificial intelligence, and research on automated knowledge acquisiton from empirical data is a central focus in machine learning. However, few existing knowledge acquisition systems are applicable to the real-world problems. In this paper, various automated knowledge acquisition systems implemented at H. I. T. are presented. All these systems are applicable to many real-world problems, are already put into the marketplace, and are going to form the first automated knowledge acquisiton industry in the world

一、引 言

当人工智能研究正处于徘徊不定的局面时，机器学习的研究仍方兴未艾，而基于机器学习方法的知识自动获取的研究正向纵深发展。然而，当前知识获取的研究基本上仍处于实验室阶段。究其原因，有的是由于方法本身固有的困难，例如类比学习中类比源的寻找依赖于灵感机制的解决，联接学习的突破有待于神经元机理的发现以及好的学习算法的探求；有的是由于方法尚不成熟，缺乏实践的考验，例如解释学习面临专家系统类似的困难，如领域知识的不完备等，机器发现或者方法过于简单，如科学发现，或者刚刚问世，如知识发现，观察学习中的概念聚类与概念形成尚未能应用于大规模问题，遗传算法正在摸索自身发展的道路，而Valiant的PAC学习理论仍然与符号学习及实际应用脱节。在所有学习方法中，只有示例学习发展较为成熟，并且应用价值大，其中知名的方法有决策树归纳ID3[1]、决策规则归纳AQ15[2,3]、与扩张矩阵方法AE5[4,5]。ID3的特点是训练与测试速度快，AQ15的特点是产生的决策规则较为优化并有很好的不精确推理机制，AE5理论基础坚实。然而，它们距实用化仍有一段距离。例如ID3缺乏一个较好的近似匹配策略，AQ15与AE5对连续取值的属性尚无一个有效的离散化方法等。另一方面，现存学习方法都是单一的学习策略，解决的问题很有限，而人类知识获取一般需要多种学习策略的协同工作。

本文介绍作者及其研究小组近年来在知识自动获取的方法、实用化与商品化方面的一系列工作。其中已实用化的较为重要的工作有：基于认识论的通用智能软件包THOUGHT[6]，通用学习系统AQ15的实用化版本AQ19，AE5的改进AE9，正例学习系统LPC[7]，扩张矩阵的模拟退火实现，遗传算法的改进[8]等；在应用领域取得突破性进展的实用系统有基于学习的自由手写数字识别系统HWNR[9]（包括数据表格自动录入卡和邮码识别系统），脱机手写与印刷体汉字识别系统HWCR[10]、布尔函数极小化实用系统COMB[11]等。

同国际上现有学习系统相比，我们的工

洪家荣 教授，研究方向：机器学习、专家系统、模式识别、并行计算、计算几何。

作在方法的先进性、集成化的规模、实用化与商品化的程度、以及应用的广泛性等诸方面均具有明显的优势。我们并期望在近两三年内能形成国际上第一家知识自动获取产业。

二、集成化知识自动获取

人类智能的获取需要多种智能机制的协同工作。对于大规模困难问题，人工智能的获取也需要多种智能方法的联合。这里，我们着重介绍两个集成化智能系统THOUGHT与AQ19。

2.1 专家系统实时化工具THOUGHT

THOUGHT[6]是一个多功能通用智能软件包，由900多个通用子程序和函数组成，共有4万多PASCAL语句行。由这些子程序与函数的适当组合可以构造出多种多样的智能系统，如专家系统、分类器系统与遗传算法、知识发现系统、以及各种知识自动获取系统等等。THOUGHT的核心由专家系统EST、经验获取系统EXACQ、概念聚类系统LEOBS、示例学习系统GS[12]、以及实时专家系统EXPERT组成。

THOUGHT各子系统协同工作过程如下：专家系统EST对数据库中的已知数据用知识库中的产生式规则推理，产生问题的求解经验，并由EXACQ搜集。一条经验由经验规则和经验路径组成。一条经验规则以已知的事实为前提，所得的解为结论；一条经验路径是该次求解过程所用到的必要规则编号的序列。概念聚类系统LEOBS将经验规则的集合按概念的相似性分类，然后由示例学习系统GS产生每一类排除其余类的概念描述。这个过程递归进行下去，产生一棵决策树状的结构。在这个结构中，每一结点是经验规则（以及所有规则号集合）的一个子集。父结点所含经验是其子结点所含经验的并集。父结点到其子结点的弧上的索引是该子结点的经验规则集的概念描述。叶结点由同一个解的经验规则及相应经验路径组成。新产生的专家系统EXPERT有三种问题求解方式：对经验过的问题（其标志为从根到某叶结点的所有索引全部被满足）实现在决策树上的检索；对同以往经验类似的问题（其标志为只满足从根结点到某一非叶结点的子路径的全部索引），用在该非叶结点上的规则号码所代表的子知识库进行推理求解；对完全陌生的问题（即不满足根以下的任何索引）则用原专家系统EST进行求解。第一种方式是快速的、实时的，相应于专家直觉求解；第二种是较快的，相应于专家联想类比求解；第三种是最慢的，相当于生手推理。

目前THOUGHT已经商品化。在鞍钢冷轧厂酸洗自动线建造了一个实时故障诊断系统。对具有500多条产生式规则、200多目标（故障）的专家系统，在98次求解中测试，EXPERT的求解速度比EST快一百多倍。

2.2 新一代知识自动获取工具AQ19

AQ19旨在将当前国际上最有影响的两个学习系统ID3[11]与AQ15[8]结合起来，互相取长补短，并且增加了实用化所必需的功能：连续取值属性的定性化。AQ19的工作原理是，对大规模数据库，先用ID3进行分类，化成一些小数据库，然后再用AQ15在每一个小数据库上产生相应的决策规则。AQ19用Bayes分类器和K-N近邻估计[13]将连续属性值离散化，这种方法的优点是具有最小风险。目前AQ19已应用于大庆录井数据综合解释，并进行了现场测试，结果比我们建造的具有四千多条规则的大庆油、气、水录井综合解释专家系统HD.1给出的结果还要准确。更重要的是，只要提供有关数据库，AQ19就可以自动产生相应的决策规则；然而如果用HD.1就须根据层号、井号、地区等的不同不断地修改知识库，而对如此庞大的知识库的修改是相当困难的。

AQ19还用于安徽省计算中心承担的国家863项目农业专家系统的知识自动获取。

三、知识获取在模式识别方面的实际应用

我们已将知识获取方法成功地应用于模

式识别领域中的一些困难问题，如手写数字识别与手写汉字识别。

3.1 手写数字识别系统HWNR

自由手写数字识别要求书写无限制、识别速度快、识别率高，因此难度较大。目前国内外基于传统模式识别方法的手写数字识别都对书写有较严格限制，识别率低（误识率大于5%）。为了降低误识率，不得不扩大拒识率（一般可达20%），并采用人机交互式，因而降低了识别效率。

我们将机器学习方法引入手写数字识别。具体做法是从十多万样本中挑选两万多做为训练样本，并将另外一万多样本用来测试。然后用AQ19在SUN4工作站上产生每类样本的描述规则，并用这些规则进行测试。测试结果表明系统HWNR误识率低于0.1%，拒识率低于3%，对书写无限制并具有自学习功能，达到了实用化水平。

目前我们同大庆开源应用技术开发公司合作，已做成数据表格自动录入卡（包括表格、印刷汉字与数字、手写数字与符号识别等部分），即将投入市场。我们现在正进一步提高HWNR的识别率，以使之用于信函自动分拣。

3.2 脱机手写汉字识别系统HWCR

脱机手写体汉字识别由于汉字多、变体多、写法因人而异，因此难度极大。十多年来国内集聚一个庞大的科研队伍攻坚，取得了许多重大成果。但92年经国家科委评比测试表明，离实用水平尚有一段距离，看来这个问题完全靠传统模式识别的方法可能很难解决。我们的HWCR系统就是企图探索一条新路。整个系统将模式识别同机器学习与自然语言理解结合起来，机器学习在四个层次上进行，决策规则与决策树归纳协同使用，采用传统的四角号码检字法做分类，后台处理基于语法与语义分析。可以证明，这种识别方法对训练样本的识别率为100%。目前该系统已经基本完成。

四、在其他领域的应用

我们研制的知识自动获取系统在医学[3]，大规模集成电路等领域有重要应用，在布尔函数化简方面，我们将机器学习方法同传统方法相结合，达到了国际上报导的最好结果[14]。

五、结论

机器学习是人工智能中的核心领域和带头学科，基于机器学习的知识自动获取将推动人工智能实用化的进程。我们在知识获取方面一系列研究工作和商品化努力，展示了知识获取首先在中国商品化与产业化的美好前景。（参考文献共13篇略）

（接第49页）

参考文献

[1] O. M. Nierrstrasz"An Object-oriented System", Office Automation System, 1986

[2] P. Dadam, et al., A DBMS prototype to support extanded NF² relation: An integrated view on flat tables and hierarchies, Proc. of ACM-SIGMOD 86 Intl. Conf. on Management of data, Washinton, D. C, 1986

[3] Jack A. et al.,PROBE Spatial Data Modeling and Query Processing in An Image Database Appiication, IEEE Transactions on Software Engineering, Vol. 14, No. 5, 1988

[4] 施伯乐等，数据库理论及新领域，高等教育出版社，1990

[5] 冯玉才，汉字关系数据库管理系统CRDS，软件产业，1988

[6] "第三代数据库宣言"，计算机科学，1991，1