初中几何专家系统中的知识获取及实现

常建鹏,赵 克,亿珍珍,万棋顺
(西安电子科技大学 机电工程学院,陕西 西安 710071)

摘 要:描述了初中几何专家系统中知识获取及实现的一般过程,指出了知识获取及实现中的难点和重点。由于研究问题的复杂性,专家系统规则库中规则量往往十分庞大,这给规则库的管理和维护带来很大不便。专家系统知识库的冗余性是影响系统运行效率和知识库维护的一个重要方面,针对一个具体的专家系统——平面几何智能解题系统,分析了关于知识库规则生成时效率低的问题,然后利用基于粗糙集的约简理论来消除和减少规则库的冗余,使得系统规则库中的规则精炼、简洁,易于维护,同时大大提高了系统的效率。

关键词:专家系统; 知识获取; 粗糙集; 冗余性
中图分类号:TP182 **文献标识码**:A **文章编号**:1673-629X(2008)07-0156-04

KA and Realization in Junior Middle School Geometry Expert System

CHANG Jian-peng, ZHAO Ke, YI Zhen-zhen, WAN Qi-shun
(School of Electromechanical Engineering, Xidian University, Xi'an 710071, China)

Abstract: Describes the generic process of KA and realization in the junior middle school geometry expert system, indicates the difficulty and stress of the process of KA and realization. Because of the complexity of research problems, rule bases on the expert system are usually huge, which brings much inconvenience for rule bases in its management and the maintenance. Redundancy of knowledgebase (KB) makes an important effect on running an expert system (ES). Aiming at a special ES, solving system for plane geomelry, the generation of KB is analyzed on its low efficiency, then eliminate and reduce redundancy of the rule base using reduction theory based on the rough set theory, it makes rule of expert system become refined, compact and tractable, and boosts the efficiency of expert system greatly at the same time.

Key words: expert system; KA; rough set theory; redundancy

0 引 言

专家系统是一种模拟人类专家解决领域问题的计算机程序系统。近年来,专家系统在各个领域获得了广泛的应用[1]。在设计专家系统时,知识获取是将大量的底层数据经过分类与抽象而得到上层信息,目标在于将专家系统中感兴趣的问题编成知识体。负责知识获取的知识工程师的任务就是使计算机尽可能模拟人类专家解决某些实际问题的决策和工作过程。但是,目前还没有人提出一种成熟的知识获取策略[2]。

文中所开发的系统为一个初中平面几何智能辅导专家系统,该智能辅导系统主要用于初中学生的课后辅导。文中将探讨平面几何领域内的专家系统的知识获取方法以及基于粗糙集理论(Rough set theory,简称RST)的规则冗余性化简。

收稿日期:2007-10-24
基金项目:科技部科技型企业创新基金(01c26226111002)
作者简介:常建鹏(1983-),男,陕西榆林人,硕士研究生,研究方向为人工智能、知识工程、创新设计;赵 克,博士,教授,研究方向为人工智能、知识工程、创新设计。

1 系统信息获取

文中所开发的几何专家系统涉及规则3300多条,功能强大,由许多小模块有机构成。为了简明易懂,不妨以"判断作辅助线的原因模块"为例来说明几何专家系统中的知识获取方法。图1是本模块框图。

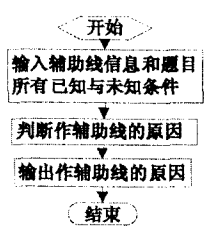

图1 判断作辅助线的原因模块框图

1.1 信息的收集

信息收集是指从各个知识源获取知识。知识源包括领域专家、书本,及系统的运行实践等。收集是对问

题的基本待征、求解问题所采用的方法的记录以及求解结果的收集与整理。由于是基于初中几何领域的问题处理,所以对于判断作辅助线的原因模块的信息搜集就比较容易了,可从各个知识源,比如课本、教师和各类习题资料及辅导书等中获取平面几何辅助线相关的定义,定理等知识及题目。笔者认为,在这个特定的区域范围之内的所有个体组成的就是问题的论域 U,所要做的就是从论域 U 中提取解题及规则判断所需要的关键信息。

1.2 信息的抽象与分析

信息的解释是对收集到的信息的评述和对关键知识的辨识。信息的分析是指在前面的基础上,识别专家系统所需的重要概念,并且决定概念之间的关系以及如何运用这些关系来解决问题。在这一阶段因为隐含知识和耦合知识的存在,会给分析过程造成相当大的困难,这也就是整个知识获取过程的重点难点所在。需要对这个论域 U 进行划分,也就是所说的分类。对于判断作辅助线的原因模块来说需要研究的对象,也就是论域 U,是平面几何领域内的辅助线问题域。对于论域 $U=\{$平面几何领域内的辅助线问题域$\}$,B^1 是条件属性集合$\{B_1,B_2,\cdots,B_m\}$,可以得到论域 U 上的一种分类:$U/B^1=F^1=\{X_1,X_2,\cdots,X_m\}(U=\bigcup_{i=1}^{m}X_i,m>1)$,其中集合簇 F^1 是论域 U 上定义的知识。对于不同的条件属性集合如 $B^2=\{B'_1,B'_2,\cdots,B'_m\}$,则有另一种分类 $U/B^2=F^2=\{Y_1,Y_2,\cdots,Y_m\}(U=\bigcup_{i=1}^{m}Y_i,m>1)$[3]。

由此可见,并不是任意一种分类标准得到的集合簇都能满足要求,关键在于找出恰当的条件属性集合 B。对于每个概念 $X_i(X_i\subseteq U)$,包含于 X_i 中的最大可定义集称为 X_i 的下近似集 $B_-(X_i)$;反之,包含于 X_i 中的最小可定义集称为上近似集 $B^-(X_i)$。可以得到:

$$B_-(X_i)=\bigcup\{Y_i\mid(Y_i\in U\mid \mathrm{IND}(B)\wedge Y_i\subseteq X)\}$$

$$B^-(X_i)=\bigcup\{Y_i\mid(Y_i\in U\mid \mathrm{IND}(B)\wedge Y_i\cap X\neq\varnothing)\}$$

其中:

$$U\mid\mathrm{IND}(B)=\{X\mid(X\subseteq U\wedge\forall x\forall y\forall b(b(x)=b(y)))\}$$

$U\mid\mathrm{IND}(B)$ 表示条件属性集合 B 对 U 的划分,也就是论域 U 的 B 基本集的集合。由此,可以通过定义 $d_B(F)=\sum_{i=1}^{n}\mid B_-(X_i)\mid/\sum_{i=1}^{n}\mid B^-(X_i)\mid$ 来描述条件属性集合 B 对 F 近似分类的准确度。根据条件属性 B_i 在问题域中出现的次数多少即属性频度即可判断属性的重要性,频度越高,条件属性 B_i 就越重要。根据条件属性对论域 F 近似分类的准确度,可以对条件属性集合进行约简,实现知识的简化[3]。

通过对收集到的辅助线相关知识及所有题目的分析后发现,每道题的设计都是跟知识点(例如:两圆公切线,构造弦切角,三角形中位线等等)相关的,即出题者往往都是通过一道题来考察学生对某个知识点的掌握程度。因此,将知识点作为分类标准对问题域进行分类。这样根据条件属性集合 $B=\{$知识点$\}$,对于论域 $U=\{$平面几何领域内的辅助线问题域$\}$ 可得到结果属性集合 $U=\{U_1,U_2,\cdots,U_n\}(n>1)$。图2是知识点分类图。

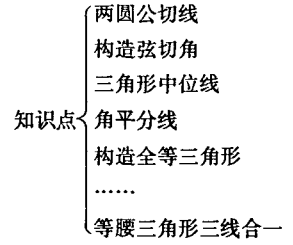

图2　知识点分类图

1.3 设　计

专家系统中系统设计的主要目的就是为系统绘制蓝图,在各种技术和实施方法中权衡利弊,精心设计,合理使用各种资源,最终勾画出新系统的详细设计方案。系统设计的基本任务大体上分为两个步骤:

①把总任务分解为许多基本的、具体的任务。

②为各个具体任务选择适当的技术手段和处理方法,即详细设计。

在具体模块的设计中,遵循可扩充性和简单性两大原则[4]。

对于判断作辅助线的原因模块,将知识点作为分类标准对问题域进行分类,所以让每一个知识点对应一些规则。对于每一条辅助线,让题目的所有已知和未知条件逐个与各个知识点来进行匹配,最后推出作每一条辅助线的依据到底是哪个已知或未知条件。图3为本模块具体流程图。

2 基于粗糙集的属性约简

由于处理问题的复杂性,所开发的几何专家系统规则库十分庞大,规则条件属性很多。这不仅严重影响了系统的效率,而且也使规则库变得冗余,不易理解,从而使得对规则的增删和修改变得困难,容易导致冗余规则的产生,甚至破坏规则的一致性和完整性,引起规则库维护代价的增加[5]。因此,利用粗糙集的可辨识矩阵方法对规则集进行约简,得到最小规则集和最简属性集,从而进一步精炼和简洁规则库。

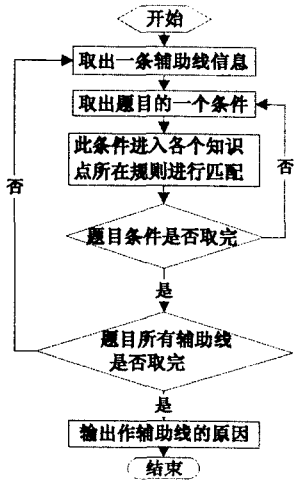

图 3 判断作辅助线的原因模块详细流程图

2.1 一些相关概念

定义 1 信息系统和决策表。

一个信息系统可表示为：$S=<U,A,V,f>$。其中 U 为对象的非空有限集合；A 为属性的非空有限集合；V 为属性的值域集；f 为信息函数，$f:U\times A\rightarrow V$。如果 $A=C\cup D, C\cup D=\varnothing$，$C$ 为条件属性集，D 为决策属性集，则把信息系统 $S=<U,A,V,f>$ 称为决策系统，用 $S=<U,C\varnothing\{d\}>$ 或 $S=<U,C\cup D>$ 来表示，其中 d 为单一的决策属性[6]。

定义 2 属性的约简和核。

对信息系统 $S=(U,C)$，属性 $a\in B\subseteq C$ 是 B 中必要的，当且仅当 $\mathrm{IND}(B)\neq\mathrm{IND}(B-\{a\})$，否则，属性 a 在 B 中是冗余或可省略的。属性集 B 的约简是一个集合 $B'\subseteq B$，当且仅当满足：①B' 是独立的。②$\mathrm{IND}(B')=\mathrm{IND}(B)$。属性集 $B\subseteq C$ 的所有约简族的交集称为属性集 B 的核，记为 $\mathrm{CORE}(B)$，有：$\mathrm{CORE}(B)=\cap\mathrm{RED}(B)$[7]。

定义 3 可辨识矩阵。

可辨识矩阵的定义为：令 $S=<U,A,V,f>$ 是一个信息系统，U 为论域，且 $U=\{x_1,x_2,\cdots,x_m\}$，C 是条件属性集，D 是决策属性集，$a(x)$ 是记录 x 在属性 a 上的值，可辨识矩阵 $\boldsymbol{C}_D$ 表示为：

$$\boldsymbol{C}_{ij}=\begin{cases}\{a\in A:A(x_i)\neq a(x_j)\} & D(x_i)\neq D(x_j)\\ 0 & D(x_i)=D(x_j)\quad i,j=1,2,\cdots,n\\ 1 & D(x_i)\neq D(x_j)\ a(x_i)=a(x_j)\end{cases}$$

由可辨识矩阵的定义可知，矩阵是一个依主对角线对称的矩阵，在分析时，只考虑其上三角或下三角部分即可。

2.2 基于可辨识矩阵的属性约简算法

(1) 计算决策表的可辨识矩阵 $\boldsymbol{C}_D$；

(2) 对于可辨识矩阵中所有取值为非 1 的元素 $\boldsymbol{C}_{ij}(c_{ij}\neq 0, c_{ij}\neq 1)$，建立相应的析取逻辑表达式 L_{ij}：$L_{ij}=\bigcup_{a_i\in C_{ij}}a_i$；

(3) 将所有的析取逻辑表达式 L_{ij} 进行合取运算，得一个合取范式 L，即 $L=\bigcap_{C_{ij}\neq 0,C_{ij}\neq 1}L_{ij}$；

(4) 将合取范式 L 转换为析取范式的形式，得 $L'=\forall_i L_i$；

(5) 输出属性约简结果。析取范式中的每个合取项对应一个属性约简的结果，每个合取项中所包含的属性组成约简后的条件属性集合。

在可辨识矩阵中，如果存在一个元素，其取值为包含单属性元素的集合，则表明该属性是区分这个矩阵元素所对应的两个样本所必须的属性，也是唯一能够区分这两个样本的属性。可辨识矩阵中的这些元素所包含的属性组成的属性集合就是该决策表系统的相对属性核。可以首先将这些属性取出，将可辨识矩阵中包含核属性的元素的值修改为 0，从而得到一个新的矩阵，再在新矩阵的基础上实施算法的第 2,3,4 步，得到一个析取方式逻辑表达式，最后将所有的核属性加入析取范式的每个合取项，得到属性约简的结果[8]。

这里仍然以判断作辅助线的原因模块为例来进行约简。假设一条辅助线所涉及的知识点是构造弦切角，那么作此辅助线所依据的条件可能有多个。将决策表的条件属性定义为：$C1$：线之间的交点，1 表示辅助线与切线有一个交点，2 表示辅助线与圆有两个交点，3 表示切线，圆与辅助线有一个公共交点；$C2$：输入条件的性质，1 表示条件是中间过程推出的，2 表示条件是已知的，3 表示条件是待证的；$C3$：输入条件的具体值，1 表示条件中含有辅助线的名称，2 表示条件中有“切线”标志；$C4$：辅助线类型，1 表示类型为连接，0 表示未告诉类型；D：决策属性，判断条件是否为作辅助线的依据，1 表示是，0 表示不是。

所以判断作辅助线的原因模块关于构造弦切角的决策规则用决策表的形式表示见表 1。

表 1 关于构造弦切角知识点的决策表

U	$C1$	$C2$	$C3$	$C4$	D
1	1	1	1	0	0
2	1	1	1	1	0
3	2	1	1	0	1
4	3	2	1	0	1
5	3	3	2	0	1
6	3	3	2	1	0
7	2	3	2	1	1
8	1	2	1	0	0
9	1	3	2	0	1
10	3	2	2	0	1
…	…	…	…	…	…

利用属性约简算法约简后的规则表见表 2。

表2　约简后的决策表

U	C1	C3	C4	D
1	1	1	0	0
2	1	1	1	0
3	2	1	0	1
4	3	1	0	1
5	3	2	0	1
6	3	2	1	0
7	2	2	1	1
9	1	2	0	1
…	…	…	…	…

其中,以规则7为例,其表示:

if 辅助线与圆有两个交点且输入条件中有"切线"标志,同时辅助线类型为连接

then 此输入条件就是作辅助线的依据条件

所开发的几何专家系统规则库经过约简后,不但约掉了规则中的冗余属性,同时约掉了冗余决策规则,使规则库变得精炼、简洁,易于维护,从而大大提高了系统的运行效率。最终,使初中几何专家系统的规则库的规则数目减少为2700多条,运行效率提高了12%左右。

3　结束语

知识获取在构建整个专家系统的过程中所占的地位举足轻重。探讨了初中几何专家系统领域内的知识获取及实现的一般方法,解决了知识获取中的难点,然后利用基于粗糙集的约简理论来消除和减少规则库的冗余,使得平面几何系统规则库中的规则精炼、简洁,易于维护,同时大大提高了系统的效率。系统运行结果证明,此分析方法是有效的,有利于问题域的求解与实现。

参考文献:

[1] 田盛丰. 人工智能原理与应用——专家系统、机器学习、面向对象的方法[M]. 北京:北京理工大学出版社,1998:12-59.

[2] 蔡自清, Durkin J, 龚　涛. 高级专家系统:原理、设计及应用[M]. 北京:科学出版社,2005:33-66.

[3] 刘　东. 知识管理的基本过程与知识的分类管理模式[J]. 南京政治学院学报,2002,18(6):44-47.

[4] 陈　平,褚　华. 软件设计师教程[M]. 北京:清华大学出版社,2004:222-237.

[5] 亿珍珍, 赵　克,许　威. 基于粗集的知识库冗余性化简研究[J]. 计算机工程与设计,2004,25(10):1731-1733.

[6] 王国胤. Rough集理论与知识获取[M]. 西安:西安交通大学出版社,2001:134-139.

[7] 曾黄麟. 粗集理论及其应用——关于数据推理的新方法[M]. 重庆:重庆大学出版社,1995:55-125.

[8] 张文修,吴伟志,梁吉业,等. 粗糙集理论与方法[M]. 北京:科学出版社,2003:26-40.

(上接第155页)

```
    dec ebx;计数器减一
    jnz L1;循环
    pop esi
NoAction:add esi,0x20;指向下一个程序段
        dec ecx
        jnz Begin
LoadEnd:
```

3　结束语

Bootloader的设计使内核的启动脱离GNU GRUB的束缚,真正让软件开发人员了解和控制程序的运行,虽然笔者是以软盘为例来实现的,但在实际应用中可在虚拟机环境中使用。首先做一个软盘映像文件;其次,使用二进制编辑软件将该软盘映像文件的前512字节内容用BOOT的内容替换;第三,将该软盘映像文件作为一个磁盘挂在文件系统中,或利用磁盘工具,将LOADER、CONFIG.CFG和内核文件复制到映像文件中,再卸载该映像文件,然后在Bochs、QEMU或VMware等虚拟机软件中启动。通过虚拟机可方便地对内核进行调试,不需要每修改一次内核程序,就启动一次机器。通过设计Bootloader可以简化系统的启动过程,使系统能更快地投入运行。当然,Bootloader的设计还不太完善,因为在装入内核代码时,Intel 80x86 CPU工作在实模式下,只能访问1MB的空间,而内核被装到0x10000至0x90000存储空间中,要求内核的长度不能大于512kB。如何打破这一限制将是今后进一步研究解决的问题。

参考文献:

[1] Straumann T. Open Source Real Time Operating System Overview[C]//8th International Conference on Accelerator & Large Experimental Physics Control Systems. San Jose, California:[s.n.],2001.

[2] 陈海军,申卫昌,史　颖. 嵌入式系统引导程序详探[J]. 计算机技术与发展,2006,16(1):123-128.

[3] 徐亚鹏,谢凯年. 用U-Boot构建IXP2350目标系统的引导程序[J]. 计算机技术与发展,2007,17(5):10-14.

[4] 于　渊. 自己动手写操作系统[M]. 北京:电子工业出版社,2006.

[5] 何先波,唐宁九,吕　方,等. ELF文件格式及应用[J]. 计算机应用研究,2001,18(11):144-145.

[6] 倪继利. Linux内核分析及编程[M]. 北京:电子工业出版社,2005.