一种专家系统知识获取时的属性约简算法

任永昌,朱 萍,李仲秋
(渤海大学 信息科学与技术学院,辽宁 锦州 121013)

摘 要:知识获取是构造专家系统的“瓶颈”,提供准确的推理知识是进行决策规划的关键。文中运用粗糙集理论,通过粗糙集的约简消除冗余的条件属性,实现对知识库的精简。首先研究知识获取,在阐明知识的层次结构基础上,给出了概念化、形式化、知识库求精三个知识获取过程;然后研究属性约简算法,在研究集合差异度和属性的重要性、约简算法推导过程的基础上,给出了属性约简算法的六个步骤。最后根据属性约简算法及其步骤,对功能点分析法构建软件成本估算专家系统时,组成技术复杂因子的 14 个因素进行了约简。

关键词:专家系统;知识获取;属性约简算法;粗糙集理论
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X (2012)09-0050-03

An Attributes Reduction Algorithm of Expert System Knowledge Acquisition

REN Yong-chang, ZHU Ping, LI Zhong-qiu
(College of Information Science and Technology, Bohai University, Jinzhou 121013, China)

Abstract: Knowledge acquisition is the "bottleneck" of construction expert system, to provide an accurate inference of knowledge is the key decision-making plan. It uses the rough sets theory, eliminate redundant condition attribute through the rough sets reduction to achieve the streamlining of the knowledge library. First study the knowledge acquisition, in exposition knowledge hierarchical structure foundation, has given three knowledge acquisition of the conceptualization, formal, the knowledge library refinement and so on. And then study attributes reduction algorithms, on the basis of researching sets difference and the attribute importance, the reduction algorithms inferential reasoning process, has given the attribute reduction algorithms six steps. Finally, according to the attributes reduction algorithms and the steps, 14 factors of the composition technology complexity factor are reduced when software cost estimation expert system is constructed by function analysis method.

Key words: expert system; knowledge acquisition; attributes reduction algorithms; rough sets theory

0 引 言

随着信息技术的发展,人类积累的数据和知识迅速增长,如何从大量的数据和知识中获取最有价值的信息,是当前科学研究和决策面临的新课题[1]。属性约简是粗糙集理论研究的基本问题之一,通过属性约简,可以除去属性集中的冗余属性或不重要属性,有利于减少决策知识的维数,降低决策的复杂性,提高决策的准确度。

1 知识获取

知识获取就是把用于求解专门领域问题的知识从众多的海量数据中抽取出来,并转换为特定的计算机表示的过程。计算机表示有产生式表示、面向对象表示、框架表示和自定义表示等。知识获取是构建专家系统花费时间最长、最困难的部分,是开发专家系统的“瓶颈”。

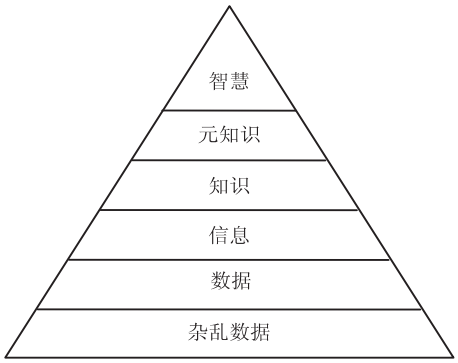
1.1 知识的层次结构

知识反映了客观世界中事物之间的联系,不同事物之间或者相同事物间的不同关系形成了不同的知识。知识可以分成不同的层次,如图 1 所示[2]。这里所说的专家系统知识获取时的知识是图 1 中层次结构中的一部分。

在知识的层次结构中,最底层(第 6 层)是杂乱数据,由几乎没有意义的或含糊难解的数据组成;第 5 层是数据,是一些有潜在意义数据项;第 4 层是信息,是经过加工处理后具有意义的数据;第 3 层是知识,代表专门化的信息;第 2 层是元知识,是“在…之上”的意思,是教人们如何运用知识的知识,元知识的有效运用

收稿日期:2012-02-05;**修回日期**:2012-05-12
基金项目:国家自然科学基金项目(70871067);2011 年辽宁省东欧及独联体国家重点引智项目;2011 辽宁省科学事业公益研究基金
作者简介:任永昌(1969-),男,教授,博士,从事数据挖掘、粗糙集理论研究。

可以提高专家系统的性能;最顶层(第 1 层)是智慧,在哲学意义上,智慧是所有知识的顶峰,基于人工智能的智慧工程正在持续发展。



智慧:有效地使用知识;元知识:知识的规则;
知识:使用信息的规则;信息:对知识潜在有用;
数据:潜在有用的信息;杂乱数据:无明显信息。

图 1　知识的层次结构

1.2　知识获取过程

具体获取可分为三个过程。一是概念化,确定专家系统有关的概念、信息、数据等,哪些是已知的,哪些是推理得到的;二是形式化,将概念、信息和数据转换成专家系统所要求的知识表示形式,建立专家系统求解模型框架,并确立推理规则和控制策略等;三是知识库求精,解决知识库中存在的矛盾、错误和冗余,对知识库进行测试、检查,发现问题并进行修改,直到得到满意的结果为止。知识获取过程图如图 2 所示[3]。在知识的抽取转换过程中,属性约简是其中的重要环节。

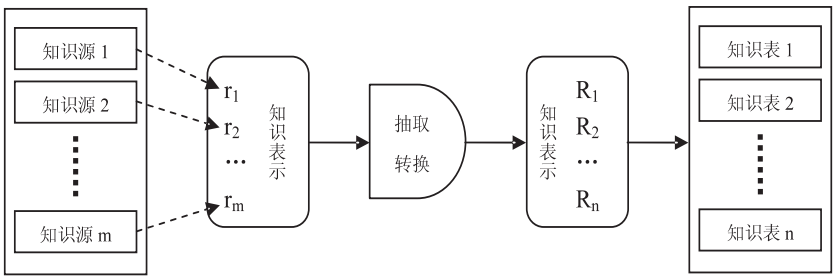


图 2　知识获取示意图

2　属性约简算法

通常知识库中存在大量的数据,提高决策精度的重要过程之一就是对知识库中的数据进行约简[4]。所谓知识约简,是指在不影响知识表达能力的前提下,通过消除冗余知识,从而获得知识库简洁表达的方法[5]。

2.1　属性的重要性

设有集合 A 和集合 B,集合差异度可以定义为[6]:

$$D(A,B)=1-\frac{|A\cap B|}{|A\cup B|}=1-H(A,B) \tag{1}$$

设 $S=(U,A)$ 是一个信息系统,$P,Q\subseteq A$,P 和 Q 所对应的知识分别为:

$$U/SIM(P)=\{S_P(x_1),S_P(x_2),\cdots,S_P(x_{|U|})\} \tag{2}$$

$$U/SIM(Q)=\{S_Q(x_1),S_Q(x_2),\cdots,S_Q(x_{|U|})\} \tag{3}$$

则知识 $U/SIM(P)$ 和知识 $U/SIM(Q)$ 的知识距离为:

$$D(P,Q)=\frac{1}{|U|}\sum_{i=1}^{|U|}\left(1-\frac{|S_P(x_i)\cap S_Q(x_i)|}{|S_P(x_i)\cup S_Q(x_i)|}\right) \tag{4}$$

设 $S=(U,A)$ 是一个完备信息系统,当 $P,Q,R\subseteq A$,且 $P\prec Q\prec R$ 时,有 $D(P,R)\geqslant D(P,Q)$ 和 $D(P,R)\geqslant D(Q,R)$ 成立。

利用知识之间的距离可以分析信息系统中每一个属性的重要性[7]。

设 $S=(U,A)$ 是一个信息系统,$a\in B\subseteq A$ 是一个属性,则知识 B 中属性 a 的重要性定义为:

$$sig_B(a)=D(B,B-\{a\}) \tag{5}$$

可以得到以下性质:

性质 1,$0\leqslant sig_B(a)\leqslant 1$;

性质 2,属性 a 在 B 中是必要的,当且仅当 $sig_B(a)>0$;

性质 3,$core(B)=\{a\in B\mid sig_B(a)>0\}$。

设 $S=(U,A)$ 是一个信息系统,$a\in B\subseteq A$ 是一个属性,$D(B,A)=0$,$sig_B(a)=0$,则知识 B 中属性 a 的不重要性定义为:

$$unsig_B(a)=D(B,\{a\}) \tag{6}$$

2.2　约简算法

设 (U,A) 是决策表,$A=C\cup D$,$C\cap D=\Phi$,其中 C 是条件属性集,D 是决策属性集。令 $(\Phi\subset X\subseteq C,\Phi\subset Y\subseteq D,U/Y\neq U/\delta=\{U\}$,可以找到 X 的一个极小子集 X_0 使得 $S_{X_0}(Y)=S_X(Y)$[8]。

也就是说,X_0 是 X 的一个子集,使 $X\leftrightarrow\cdots\leftrightarrow X_0\rightarrow\cdots\rightarrow\Phi(Y)$,其中 $X\supset\cdots\supset X_0\supset\cdots\supset\Phi$,$S_X(Y)=\cdots=S_{X_0}(Y)\supset\cdots\supset S_\Phi(Y)=S_\delta(Y)$,$S_\delta(Y)=\cup_{W\in U/Y}(\cup_{V=U,V\subseteq W}V)=\Phi$,其中 $U/Y\neq U/\delta=\{U\}$。称 X 的这种子集 X_0 为 X 的约简(对于 Y 而言)。约简可能不是惟一的,故引入下述约简的概念。

设 (U,A) 是决策表,$A=C\cup D$,$C\cap D=\Phi$,其中 C 是条件属性集,D 是决策属性集。令 $(\Phi\subset X\subseteq C,\Phi\subset Y\subseteq D,U/Y\neq U/\delta=\{U\}$。如果 $X_0\in X$ 满足:

(1) $S_{X_0}(Y)=S_X(Y)$,即 $X_0\leftrightarrow X(Y)$;

(2)如果 $X'\subset X_0$,则 $S_X(Y)\supset S_{X'}\subset(Y)$,即如果 $X'\subset X_0$,则 $X'\leftrightarrow X(Y)$ 不成立。

则称 X_0 是 X 的一个约简(对于 Y 而言)。

空子集 Φ 的约简为 Φ(对于 Y 而言)。

也就是说,X_0 是 X 的极小恒等依赖子集(对于 Y

而言),即对于 $X \supseteq \cdots \supseteq X_0 \supset X' \supseteq \cdots$,有 $X \leftrightarrow \cdots \leftrightarrow X_0 \rightarrow X' \rightarrow \cdots (Y)$。

从这个定义可知,所有约简的时间复杂性是指数的。对于知识发现,X 的一个约简 X_0(对于 Y 而言)意味着:

(1)规则"X_0 蕴涵 Y"和规则"X 蕴涵 Y"同样强;

(2)如果 $X' \subset X_0$,则规则"X' 蕴涵 Y"严格弱于规则"X 蕴涵 Y"。

2.3 约简算法步骤

粗糙集是用来处理不确定、不完备数据的重要工具之一[9]。信息系统分为不完备信息系统和完备信息系统。对于完备信息系统的属性约简,通常是从计算核属性集开始;对于不完备信息系统的属性约简,直接从空集开始逐步加入当前属性重要性最大的属性[10~12]。

输入:一个信息系统 $S = (U, A, V, f)$,其中 $A = C \cup \{d\}$,C 是条件属性集合,d 是决策属性,条件属性个数为 n。

输出:S 的属性约简集 red。

Step 1,初始化 $red = \Phi$,$P = C$;

Step 2,利用公式(1)计算条件属性 C 相对于决策属性集 D 的差异度,利用公式(5)计算条件属性 $C_i \in C$ 相对于决策属性集 D 的属性重要性;

Step 3,将属性重要性值归一化为属性 C_i 的权重 ω_i,利用权重联系度关系计算条件属性 C 相对于决策属性集 D 的近似分类质量 $r(C)$;

Step 4,若 $red = \Phi$,需要选取权重最大的属性 C_k,$red = C_k$,C_k 的权重为 1,就要计算 $r(red)$,若 $r(red) \geqslant r(C)$,转 Step 6,否则继续下一步;

Step 5,对每个 $C_i \in P - red$,将属于 $red \cup \{C_i\}$ 中的属性的权重做归一化处理,然后利用权重联系度关系计算 $r(red \cup \{C_i\})$,若属性 p 使 $r(red \cup \{p\})$ 值最大,$red = red \cup \{p\}$,若 $r(red \cup \{p\}) < r(C)$,$P = P - \{p\}$,然后转 Step 5;否则约简结束,转到 Step 6;

Step 6,属性约简的输出结果是 red。

3 属性约简实例

运用功能点分析法(Function Point Analysis)构建软件成本估算专家系统时,最初得到的技术复杂因子(TCF, Technical Complexity Factor)由 14 个因素组成,见表 1。

表 1 技术复杂因子

序号	名称	序号	名称
1	数据通信	8	在线更新
2	分布式处理	9	复杂处理
3	性能	10	重用性
4	配置项负载	11	安装难易程度
5	事务率	12	操作难易程度
6	在线数据项	13	多个地点
7	用户使用效率	14	修改难易程度

每个因子按照其对系统的重要程度分为六个级别,取值见表 2。

表 2 权重表

0	1	2	3	4	5
没有影响	偶有影响	轻微影响	一般影响	较大影响	严重影响

进行初步成本估算时,得到的估算结果是成本高、成本中等、成本低三种形式之一。通过调查问卷得到的原始成本估算数据如表 3 所示。

运用上述方法,约简掉 5 个属性,序号分别为{1,3,5,8,12},TCF 名称分别是{数据通信、性能、事务率、在线更新、操作难易程度}。最小属性约简集的序号为{2,4,6,7,9,10,11,13,14},TCF 名称分别是{分布式处理、配置项负载、在线数据项、用户使用效率、复杂处理、重用性、安装难易程度、多个地点、修改难易程度}。约简率为 35.7%。

4 结束语

知识获取和表示就是把解决特定问题所用的专门知识从某些知识来源中提炼出来,并表示成计算机能接受和使用的方式。在专家系统中,提供准确的推理

表 3 软件成本估算原始数据

序号	技术复杂因子取值														估算结果
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
1	1	3	0	5	2	4	3	0	5	4	3	5	4	5	高
2	2	3	0	4	3	3	5	0	2	1	2	3	2	3	中等
3	0	2	2	1	3	4	2	0	3	5	2	2	3	1	低
4	0	4	1	2	3	4	2	1	2	4	1	4	4	2	中等
5	2	5	2	3	4	5	4	0	2	4	5	2	5	3	高
6	0	2	0	1	2	5	5	2	0	2	3	3	5	1	中等
7	3	2	0	1	5	2	2	0	1	0	3	4	2	2	低
8	1	0	4	3	2	4	2	4	1	5	4	2	1	3	中等

(下转第 58 页)

化随机应变,对于每一步都有十分清晰的计划,从需求调研开始,整个开发团队就都进入了忙碌的阶段:需求、设计、代码、测试。而瀑布模型则是一步一步的走下去的。敏捷开发注重“民主”,每个开发中作者都是自己的领导者。近年来,随着敏捷开发思想的提出,以及 UP(Unified Process,统一流程)、敏捷 UP、Scrum 和 XP(极限编程实践)等一系列的实践方法得到应用,迭代、增量的开发模式得到了更多的赞誉声音,目前,最为热门的是以 Scrum 和 XP 进行组合的敏捷开发方式,已经被腾讯、华为、上海贝尔等一些大公司所采用。

敏捷开发方法是一个过程,是一个持续的应用原则、模式以及实践来改进软件的结构和可读性的过程。它致力于保持系统设计在任何时间都尽可能得简单、干净和富有表现力[12]。

并不是选用某种开发模式,或者更前进的开发技术,就一定能保证将项目做成功。而是参考业界的一些成功的最佳实践活动经验,不断地对所用的开发模式进行检讨,合理地调整开发过程,在现有的开发资源的情况下不断地将项目的开发过程发挥得更高效。

参考文献:

[1] Livermore J A. Factors that impact implementing an agile software development methodology [C]//Proceedings of 2007 IEEE Southeast Conference. [s. l.]:[s. n.],2007:82-86.

[2] 赵剑冬,林 建. 敏捷方法在软件项目开发中的实践[J]. 计算机工程与设计,2007(6):2772-2774.

[3] 夏显鄂,梁洪峻. 敏捷软件开发与计划驱动开发的概述比较[J]. 计算机工程与设计,2007(8):4035-4037.

[4] 何 伟,杨宗德,张 兵. 基于 Symbian OS 的手机开发与应用[M]. 北京:人民邮电出版社,2006.

[5] 罗运模. 软件能力成熟度模型集成(CMMI)培训教程[M]. 北京:清华大学出版社,2003:106-130.

[6] 施瓦伯. Scrum 敏捷项目管理[M]. 北京:清华大学出版社,2007:100-115.

[7] 马 丁. 敏捷软件开发:原则、模式与实践[M]. 北京:人民邮电出版社,2008.

[8] Starting an iPhone Application Business For Dummies[M]. [s. l.]:Wiley,2010.

[9] Sutherland J,Schoonheim G,Rustenburg E,et a1. Fully Distributed Serum: The Secret Sauce for Hyperproductive Offshored Development Teams[C]//Agile 2008 Conference. [s. l.]:[s. n.],2008.

[10] Moore R,Reff K,Graham J,et al. Serum at a Fortune 500 Manufacturing Company[J]. AGILE,2007(8):175-180.

[11] 成奋华,金 敏. 基于敏捷过程的 IT 项目范围管理的研究与应用[J]. 计算机技术与发展,2010,20(10):233-236.

[12] 沈 雷,沈备军. 敏捷方法的研究与实践[J]. 计算机工程,2005(4):219-222.

(上接第 52 页)

知识是进行决策规划的关键。知识获取是构造专家系统的“瓶颈”问题,专家知识的好坏直接影响整个系统的性能,因此知识获取方法得到了广泛的研究和应用。属性约简是知识获取的关键步骤,可以从一定程度上消除决策系统中的条件属性[13,14]。文中运用粗糙集理论,通过粗糙集的约简消除了冗余的条件属性,实现了对知识库的精简。进行属性值的约简相对简单,属性约简则困难很多。结果表明,该方法可以简化复杂系统的结构,并能有效维护知识库的结构和性能。

参考文献:

[1] 裴小兵. 粗糙集的知识约简研究[D]. 武汉:华中科技大学,2006.

[2] Giarratano J C,Riley G D. 专家系统原理与编程[M]. 北京:机械工业出版社,2006.

[3] 任永昌. 软件成本估算及其专家系统研究[D]. 阜新:辽宁工程技术大学,2008.

[4] 关 欣,衣 晓,何 友. 一种新的粗糙集属性约简方法及其应用[J]. 控制与决策,2009,24(3):464-467.

[5] 王 彪,段禅伦,吴 昊,等. 粗糙集与模糊集的研究及应用[M]. 北京:电子工业出版社,2008.

[6] Dai Jianhua,Li Yuanxiang,Liu Qun. A hybrid genetic algorithm for reduct of attributes in decision system based on rough set theory[J]. Wuhan University Journal of Natural Sciences,2002,7(3):285-289.

[7] 史 琨. 粗糙集的知识获取方法研究[D]. 太原:山西大学,2009.

[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.

[9] 代广珍,徐 超. 基于 RS 理论的快速属性约简求核方法[J]. 计算机技术与发展,2011,21(4):133-135.

[10] 颜 艳. 基于粗糙集的属性约简算法及其应用研究[D]. 无锡:江南大学,2008.

[11] Hedar Abdel-Rahman,Wang Jue,Fukushima M. Tabu search for attribute reduction in rough set theory[J]. Soft Computing,2008,12(9):909-918.

[12] Deng Tingquan,Ma Minghua,Wang Xinxia,et al. An improved approach to attribute reduction with ant colony optimization [J]. Fuzzy Information and Engineering,2010,2(2):145-155.

[13] 沈 玮,赵佳宝. 一种新的启发式粗集决策表属性约简算法[J]. 计算机技术与发展,2010,20(10):16-20.

[14] 汪 凌,胡 培. 基于粗糙集的决策系统知识获取算法及实证分析[J]. 情报杂志,2009,28(3):144-147.

一种专家系统知识获取时的属性约简算法

作者： 任永昌， 朱萍， 李仲秋
作者单位： 渤海大学 信息科学与技术学院,辽宁 锦州 121013
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2012(9)