doi:10.3772/j.issn.1002-0470.2010.05.009

专家系统中基于认知的知识自动获取机制①

杨炳儒②* 唐志刚③*** 杨 珺*

(*北京科技大学信息工程学院知识工程研究所 北京 100083)
(**南华大学数理学院 衡阳 421001)

摘 要 针对专家系统中知识自动获取的瓶颈问题,从专家系统自身的潜在规律(机理)出发来改变知识发现的固有流程、形成新的知识发现过程模型和构建基于认知的知识自动获取机制,并利用 T 型协调器,根据基础知识库中的"知识短缺"自动地启发定向挖掘知识的途径,有效地克服领域专家的自身局限,而且做到只对那些有可能成为新知识的假设进行评价,最大限度地减少评价工作量,由此形成知识自动获取机制。这一机制在很大程度上解决了智能系统中的知识自动获取的瓶颈问题。

关键词 双库协同机制,双基融合机制,知识发现,知识自动获取,专家系统

0 引 言

20 世纪 70 年代以来,数据获取与数据存储技术得到了快速发展,数据库、数据集市和数据仓库相继出现于各个行业,如何从中提取出有用的知识,已成为信息领域亟待解决的问题之一。数据库中的知识发现(knowledge discovery in database,KDD)的出现,为人们提供了一条解决这个问题的有效途径。目前,KDD 的发展主流是寻求在各类数据库和应用背景下的高性能、高扩展性的挖掘算法,然而所进行的研究对较高层次的框架乃至理论基础没有进行深入探讨[1,2],因而无法从根本上明显提高现有知识发现过程的性能。这些研究包括:文献[3]提出的基于证据理论的通用数据挖掘框架,这一框架具有统一的知识表示形式,支持并行计算,并有其独特的运算符号;文献[4]提出的基于验证(justification)的完全自治的知识发现框架,该框架的核心是一个推理组件,它根据感兴趣度和执行任务的验证强度来计算每一个挖掘任务的可能性,再根据这些可能性对挖掘任务进行排序,从而实现自治知识发现,同时,该框架还具有启发功能,用于提出新的挖掘任务;文献[5]提出的使用微观经济学和归纳数据库相结合的方法进行数据挖掘理论研究的想法;文献[6]从统计的角度提出的"信息范例",以及从理论上对其在关联规则和分类问题上的应用的探讨。本文把知识发现系统视为认知系统,从专家系统自身的潜在规律(机理)出发,重在研究知识发现的认知自主性,然后再用创新的知识发现机理以及模型作为专家系统的新的知识获取构件,从而可丰富和提升经典专家系统的知识库结构,形成全新的知识自动获取机制。新的知识自动获取机制的最重要的一点是为传统的专家系统增加了新的知识获取渠道,即我们提出的基于数据库与知识库的双库协同机制[7,8]下的知识发现过程模型。

1 理论基础

双库协同机制,即挖掘数据库与挖掘知识库在基于数据库的知识发现(KDD)进程中的协同机制。该机制基本上解决了数据挖掘过程中对领域固有的基础知识库的实时维护,同时在一定程度上解决了认知自主性的问题,实现了计算机自动发现"知识短缺",系统自身根据知识短缺产生创建意向,形成定向挖掘;对挖掘出来的知识通过中断型协调器对知识库进行实时管理与维护。

在给出双库协同机制之前,我们给出了 3 个布尔代数——数值域布尔代数、知识结点布尔代数及数据子类结构布尔代数和两个范畴——推理范畴及完全数据结构可达范畴。这 3 个布尔代数和两个范

① 国家自然科学基金(69835001,60675030,60875029)和教育部科技重点([2000]175)资助项目。
② 男,1943 年生,教授,博士生导师,研究方向:知识发现与智能系统,柔性建模与集成技术,E-mail: bryang-kd@yahoo.com.cn
③ 通讯作者,E-mail:damangshe22@163.com
(收稿日期:2009-05-05)

畴的具体概念以及它们之间的关系见文献[7,8]。

在给定真实数据库和基础知识库的前提下,在数据挖掘过程中具备以下特征的 KDD 运行机制为双库协同机制:(1)在真实数据库上,按数据子类结构形式所构成的挖掘数据库的可达范畴与基于属性间关系的挖掘知识库的推理范畴之间构建范畴间的等价关系;两个范畴的等价关系为定向挖掘和定向搜索奠定理论基础。(2)在 KDD 聚焦过程中,除依据用户需求确定聚焦外,通过 T 型协调算法可以形成依据挖掘知识库中知识短缺而生成的机器自身提供的聚焦方向,进而形成在数据库中的定向挖掘(算法和进程)。(3)在获得假设规则到知识评价的过程中产生的中断进程,先不对假设规则进行评价,而是通过协调算法到挖掘知识库中进行定向搜索,以期发现产生的假设规则与知识库中原有的知识是否重复、冗余和矛盾,并作相应处理,即对知识库进行实时维护。

在双库协同机制的研究中,给出了一系列定义,并演绎出重要的结构对应定理,还提出并实现了 T 型协调算法与中断协调算法。以下给出结构对应定理。

定理(结构对应定理)[7,8]:对于论域 X,在相应的知识子库与数据子库中,关于知识结点的拓扑空间 $< E,\rho(E) >$ 与关于数据子类(结构)的拓扑空间 $< F,\rho(E) >$ 是同一泛同伦型的空间。

基于双库协同机制,提出了 KDD* 新过程模型,简单地说,KDD* = KDD + 双库协同机制(其中 + 表示融合)[9]。

双基融合机制[7]是指构建基础数据库与知识库的内在联系的"通道",通过此通道,可用数据库与 KDD 去制约与驱动基于知识库的知识发现(knowledge discovery in knowledge base, KDK)的挖掘过程,改变 KDK 固有的运行机制,在结构与功能上形成相对于 KDK 而言的一个开放的优化的扩体。

基于双基融合机制,提出了 KDK* 新过程模型,简单地说 KDK* = KDK + 双基融合机制[7]。

以下 3 问题是上述两个机制共同诱导新过程模型形成的生长点:(1)突破基于数据库的知识发现的封闭系统,而与知识库协同起来,由基础知识库制约与驱动 KDD,从而发现新知识;(2)目前许多研究具体挖掘技术,应提升到宏观背景下多个抽象级和不同知识层面上的知识发现系统的一般性框架的研究;3)在"综合基"(数据库和知识库并存)上发现新知识,即将 KDD 与 KDK 有机融合,统一在知识发现的全部运行过程中。

针对上述 3 点以及认知与逻辑发展的必然,我们构造了包容 KDD* 与 KDK* 的具有特色的新系统,即具有双库协同机制与双基融合机制的综合型知识发现系统,简单地讲,就是 KD(D&K) = KDD* + KDK[*7]所表示的系统。

2 T 型协调器和 T 型协调算法

在经典 KDD 进程中,系统的聚焦通常是由用户提供感兴趣方向,KDD 依此为据进行挖掘,这种情况下大量数据中的潜在有用的信息往往被用户忽略。为帮助 KDD 尽可能多地搜索到对用户有用的信息,以弥补用户或领域专家自身的局限性,提高机器的认知自主性,本文构造了 T 型协调器,使得知识发现系统在原有用户聚焦的基础上,增加了系统自身提供聚焦方向的功能,从而有效提高了数据中隐含信息的发现概率。

2.1 发现知识短缺

T 型协调器通过搜索知识库中知识结点的不关联态来发现"知识短缺"。这里对短缺知识做一些限定:(1)短缺知识只考虑单个后件的规则;(2)同一属性的属性程度词不同时出现在同一规则的前件和后件中;(3)根据具体问题确定短缺知识最多的前件个数,前件个数过多势必造成规则难于理解;(4)对某条规则 $e_1 \hat{} e_2 \hat{} \cdots \hat{} e_n \rightarrow h$,其规则长度为 $m+1$。

对于知识库中单前件和单后件的知识,通过把规则的前件和后件看作图的顶点,利用图论中求解可达关系的方法发现"知识短缺";由于知识库中的规则大多具有多个条件,本文用有向超图[10,11]解决这个问题。

定义 1 一个超图是一个二元组 $< V,E >$,其中 $V = \{p_1,p_2,\cdots,p_n\}$ 是一个非空集合,它的元素称为超图的顶点;$E = \{e_1,e_2,\cdots,e_m\}$ 为定义在顶点集 V 上的子集簇,即 $\forall e_j \in V$, $j = 1,2,\cdots,m$,并且满足:(1) $e_j \neq \Phi(j = 1,2,\cdots,m)$;(2) $\bigcup_{j=1}^{M} e_j = V$。

定义 2 一个有向超图是一个二元组 $< V,E >$,其中 $V = \{p_1,p_2,\cdots,p_n\}$ 是素知识结点的集合来作为图的顶点, $E = \{e_1,e_2,\cdots,e_m\}$ 是知识库中规则所对应的有向边。如规则 $r_i = p_1 \wedge p_2 \wedge \cdots \wedge p_k \rightarrow p_j$,则有向边 $e_i = < (p_1,p_2\cdots,p_k),p_j >$ 是一个序偶,其中第一个元素是 V 的一个子集,与规则的前件相对应,第二个元素是 V 的一个元素,与规则的

后件相对应。

定义 3 与同一条超边关联的顶点称为互相邻接,若两条超边有一公共顶点,则称这两条有向超边邻接。

根据计算普通有向图邻接矩阵的 Warshall 算法[12],本文提出如下计算有向超图邻接矩阵的P(H)算法。

Function calculate-reach-matrix

(1) 知识库中所有的知识素结点的 ID 号 1,2,…,n 组成一个矩阵 $p_{n\times n}$, 用一个二维数组来表示,其元素均为 0,即 $p(i,j)=0$, 其中 $i,j=1,2,\cdots,n$;

(2) $e:=1$;

(3) 读取知识库中第 e 条长度为 2 的规则 $r_e:p_i\rightarrow p_j$;

(4) 矩阵 P(H)的元素 $p(i,j)=1$;

(5) Calculate-matrix1(i,j); //调用过程 Calculate-matrix1,见后面

(6) 知识库中长度为 2 的规则是否读完? 若没读完,则 $e:=e+1$, 转步骤(3);否则转(7);

(7) $e:=1$;

(8) 读取知识库中的第 e 条长度大于 2 的规则 $r_e:p_{f1}\wedge p_{f2}\wedge\cdots\wedge p_{fj}\rightarrow p_i$;

(9) Calculate-matrix2($(f_1,f_2,\cdots,f_j),i$); //调用过程 Calculate-matrix2,见后面

(10) 知识库中长度大于 2 的规则是否读完? 若没读完,则 $e:=e+1$, 转步骤(8);否则结束。

Procedure Calculate-matrix1(i,j)

(1) for $k:=1$ to n

(2) $P(k,j):=P(k,j)\vee P(k,i)$;

(3) for $m:=1$ to n

(4) $P(i,m):=P(i,m)\vee P(j,m)$。

Procedure Calculate-matrix2 ($(f_1,f_2,\cdots,f_j),i$)//($j>1$)

(1) 若虚结点 $p_{f1}\wedge p_{f2}\wedge\cdots\wedge p_{fj}$ 不存在,则在可达矩阵的后面加一行表示该结点;

(2) $P(p_{f1}\wedge p_{f2}\wedge\cdots\wedge p_{fj},i)=1$;

(3) for $s:=1$ to n;

(4) $P(p_{f1}\wedge p_{f2}\wedge\cdots\wedge p_{fj},s):=P(p_{f1}\wedge p_{f2}\wedge\cdots\wedge p_{fj},m)\vee p(i,s)$。

至此,我们求出了知识的可达矩阵,该矩阵中为 0 的元素对应了短缺的知识,对于长度小于等于 2 的短缺知识均可从这些 0 元素中得到。

2.2 短缺知识排序

在文中 2.1 节实现了找出长度不大于 2 的短缺知识,但是对长度大于 2 的短缺知识由于矩阵中只包含了在知识库中出现的合结点,则不能全部从可达矩阵 P(H)中得到。为此,本文定义了找出长度大于 2 的短缺知识的规则强度,规则强度包括规则客观的支持度(support)和主观的感兴趣度(interestingness)两方面,以下分别予以说明。

这里使用关联规则的支持度[13]概念描述规则强度的客观方面,即规则 A→B 的支持度是数据库事务的集合中同时包含 A 和 B 的百分比。

定义 4 感兴趣度是指对数据库中的各属性或属性程度词的感兴趣程度,也就是用户对知识库中知识素结点的感兴趣程度。在预处理阶段,首先由用户给出每个属性程度词的感兴趣度,即对知识素结点 e_k 的感兴趣程度,记为 Interest(e_k),其值域为[0,1],该值越大,说明用户对该知识素结点越感兴趣。对于知识合结点 $F=e_1\wedge e_2\wedge\cdots\wedge e_m$, 其感兴趣度为各知识素结点感兴趣度的平均值,即

$$\text{Interesting(F)}=\Big(\sum_{i=1}^{m}Interest(e_i)\Big)/m \tag{1}$$

对于规则 r_i: $F\rightarrow h$, 它的感兴趣度为

$$Interest(r_i)=\Big(\sum_{i=1}^{m}Interest(e_i)+Interest(h)\Big)/(Len(r_i)) \tag{2}$$

其中, $Len(r_i)$ 是规则 r_i 的长度; $Interest(h)$, $Interest(e_i)$ 分别为对 h, e_i 的感兴趣度。一般地,一个规则中包含感兴趣度大的知识素结点越多,感兴趣度小的知识素结点越少,则认为用户对该规则越感兴趣。

定义 5 规则强度(intensity)包含对规则的客观支持度和主观感兴趣度两方面。对规则 $r_i:F\rightarrow h$, 其规则强度为

$$Intensity(r_i)=(Interest(r)+\text{support}(r_i))/2 \tag{3}$$

其中, support(r_i) 为规则 r_i 的支持度, $Interest(r)$ 为对 r 的感兴趣度,规则强度同时考虑了主观和客观两方面,一方面,即使支持度较小,只要用户对该规则特别感兴趣,则规则强度就不会太小,从而该知识还可以被聚焦;另一方面,如果用户对某一规则不太感兴趣,只有该规则具有很高的支持度才有可能被聚焦。

由于规则强度中包含了支持度,因此可利用该支持度对短缺知识分层聚焦,即对长度为 2 的短缺

知识 K_2 进行聚焦,然后对长度为 3 的短缺知识 K_3 进行聚焦,直至 $K_1 = \varphi$。K_2 可直接从可达矩阵 $P(H)$ 中产生, K_2 与知识库中已有的知识构成集合 $K_2'(\forall r_j \in K_2', \text{support}(r_j) > \min\_\sup)$(这里 min_sup 是最小支持度阈值), K_3 将利用支持度从 K_2' 中产生。因为 $\forall r_3 \in K_3, r_3$ 的支持度必不小于 r_3 子集的支持度,即 $\text{support}(r_3) \leqslant \sup(r_2)$,其中 r_2 是 r_3 中的任意两个知识素结点组成的规则,而 $\text{support}(r_3) > \min\_\sup$,故 $\text{support}(r_2) > \min\_\sup$,因此 $r_2 \in K_2'$。该性质与文献[13]中介绍的大项集具有相同的性质,因此,本文利用文献[13]中提出的算法产生大项集,从而计算短缺知识。至此,可通过计算得到任意长度的短缺知识。

接下来,启发协调器通过规则的关联强度按由大到小的顺序进行排序,依次自主地形成新聚焦以发现新知识,即产生创见意向。

2.3 T 型协调算法

以上两节已分别实现了系统自主发现短缺知识和自主聚焦,下面给出 T 型协调算法。

Procedure Heuristic-Coordinator(K_2)//产生所有长度为 2 的短缺知识。

(1)把可达矩阵从数据表 ReachMatrix 中读出,把 $\text{support}(p_i) >$ min-sup 的知识素结点与全部知识合结点存入数组 P 中;

(2)$K_2 = \phi$;

(3)for $i := 0$ to n//可达矩阵的列数;

(4)for $j := 0$ to n//可达矩阵的列数;

(5)if ($P(i,j) = 0$ and $\text{attr}(p_i) \neq \text{attr}(p_j)$ and $\text{support}(p_i p_j) > \min\_\sup$)//$\text{attr}(p_i)$为知识素结点 p_i 所对应的属性,相同属性的不同程度词不能出现在同一规则中,对 i, j 对应的数据表 tablei, tablej 进行挖掘计算 $\text{support}(r_i)$;

(6)$K_2 = K_2 \cup \{i \rightarrow j\}$。

Procedure Heuristic_Coordinator(K_{x-1}, K_x)//由长度为 $x-1$ 的短缺知识产生所有长度为 $x(x>2)$ 的短缺知识。

(1)$K_x = \Phi$;

(2)对于 K_{x-1}中任意两规则 $f_{i1} \wedge f_{i2} \wedge \cdots \wedge f_{ix-1} \rightarrow j$ 和 $g_{i1} \wedge g_{i2} \wedge \cdots \wedge g_{ix-1} \rightarrow i$,若 $f_{i1} = g_{i1}, \cdots, f_{ix-1} = g_{ix-1}, f_{ix-1} = g_{ix-1} j \neq i$,则 $K_x = K_x \cup \{f_{i1} \wedge f_{i2} \wedge \cdots \wedge f_{ix-1} \wedge i \rightarrow j, f_{i1} \wedge f_{i2} \wedge \cdots \wedge f_{ix-1} \wedge j \rightarrow i\}$;

(3)对所有 $r_i \in K_x$;

(4)若 $\text{support}(r_i) <= \min\_\sup$ then 对 r_i 对应的数据表 table1, table2, …, tablep, tableq 进行挖掘,计算 $\text{support}(r_i)$;

(5)$K_x = K_x - r_i$;

这样,ESKD 就能自动发现知识库中短缺的知识,但是这些新入库的知识有可能和原有的知识相互重复、矛盾、冗余,我们就需要把这些知识剔除。在这里,首先给出知识重复、矛盾和冗余的定义,然后给出一个算法,剔除这些算法。

定义 6 若在可达矩阵中 $p((f_{i1}, f_{i2}, \cdots, f_{is}), j) = 1$,则称知识 $R: f_{i1} \wedge f_{i2} \wedge \cdots \wedge f_{is} \rightarrow j$ 是重复的。

该定义是说新得到的知识在知识库中已存在,因而是重复的。

定义 7 知识 $R: f_{i1} \wedge f_{i2} \wedge \cdots \wedge f_{is} \rightarrow j$ 是矛盾的当且仅当在知识库中存在一个知识 $T: f_{i1}, f_{i2}, \cdots, f_{is} \rightarrow i$ 且 $\text{attr}(p_i) = \text{attr}(p_s)$。

该定义是说如果有一个知识的前件与知识库中的某条知识的前件是一样的,但后件是同一个属性的不同程度词,则该知识是矛盾的。

定义 8 知识 $R: f_{i1} \wedge f_{i2} \wedge \cdots \wedge f_{is} \rightarrow j$ 是冗余的当且仅当在知识库中存在一个知识 $T: f_{i1}, f_{i2}, \cdots, f_{is} \rightarrow i$ 和知识 $K: i \rightarrow j$。

该定义是说如果有一个知识可由知识库中的知识推导出,则该知识是冗余的。

Procedure Maintenance-Coordinator($R: f_{i1} \wedge f_{i2} \wedge \cdots \wedge f_{is} \rightarrow j$)//$\text{len}(R) = x$

(1)若 R 是重复的,则{$K_x = K_x - R$; return 0;};

(2)若 R 是矛盾的,则{$K_x = K_x - R$; return 0;};

(3)若 R 是冗余的,则{$K_x = K_x - R$; return 0;};

(4)return (1)。

3 领域专家的知识获取

我们把所创建的知识获取知识用一张示意图来表示,其示意图见图 1。

知识自动获取的步骤如下:(1)首先对领域进行定义,然后引导、记录并分析专家口述知识。(2)搜索知识元素,将检测出来的概念与包含它们的记录段一起加以存储。对记录段进行语义分析,对记录段中的所有词汇进行检查,看其是否包括顺序关系(如小于、等于)和倾向(如稳定、增加)等。(3)将知识元素及其相互联系的运算符共同构成命题演算,并与现有知识的匹配导致完整命题的最终实现。(4)进行中间知识表示,记录分析的所有输出都集成到中间知识表示系统。每个命题由一个运算符(表

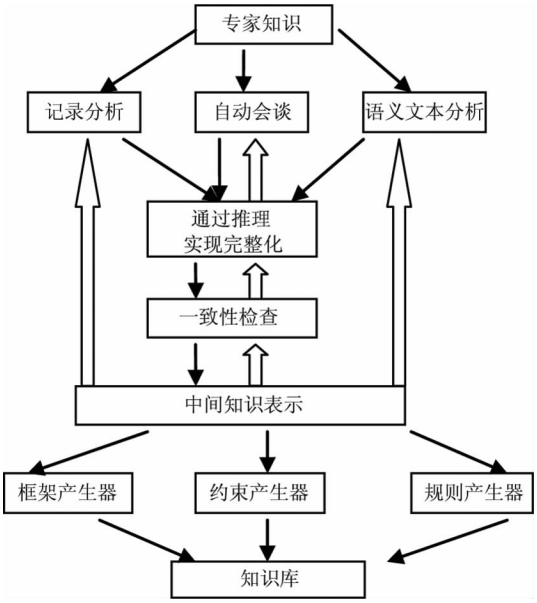

图 1　知识自动获取示意图

示概念之间的关系)、一个段标记(指向提供该命题的记录段的指针)及相关的概念组成。(5)检查结构化对象的网络的完整性,对检查到的不完整性,应重复会谈和文本分析过程。(6)将语义网中的结构化对象翻译成框架形式,并修改操作(通过调用结构编辑器进行),由知识工程师完成规则集合的组织和控制策略的选择。(7)进行约束生成,在发现数据之间有全局性依存关系时,由用户用鼠标选择数据和它们的关系,从而形成约束语言。(8)通过 T 型协调器搜索知识库中"知识结点"的不关联态,计算有向超图的可达矩阵来实现发现"知识短缺",产生"创见意象",从而启发与激活真实数据库中相应的"数据类",以产生"定向挖掘进程",进而用规则强度阈值进行剪枝并由计算机自动完成聚焦,并通过选定的知识挖掘法,从挖掘数据库中提取用户所需要的知识,并用特定的模式表达所提取的知识。

4　实验验证

基于计算机程序的蛋白质二级结构预测研究已经有 30 多年的历史,其主流是各种不同预测方法的研究。大致可分成三类:(1)基于机器学习的方法(如 SVM 方法);(2)使用多序列排列信息的方法(如 BLAST 方法);(3)使用规则和统计结合的方法(如 ILP 方法、Chou-Fasman 方法等)。然而长期以来,蛋白质二级结构预测研究进展缓慢,预测精度一般较低(低于 80%),同时当前所建立的模型与方法,无法完成揭示序列与空间构象的关系,特别是出现了绕过二级,直接由一级预测三级的趋势,但精度均不理想。故蛋白质二级结构预测研究已成为本世纪分子生物学和生物信息学领域中国际性的一大难题。

本文利用基于知识发现的专家系统(ESKD)进行蛋白质二级预测,测试使用文献[14]中的"偏 alpha/beta 型"蛋白质 256B、351C、9PAP、1BP2,组成的测试集(ILP 数据集)是 RS126 测试集以及 CB513 测试集。预测精度均以国际标准 Q3 为指标。

通过 ESKD 系统中知识自动获取机制生成的 alpha、beta 的规则,进行一定程度的约简后得到的精炼的 alpha、beta 规则库,使用改进的 CBA 算法,经过反复实验得到适用于 alpha、beta 的支持度与可信度的阈值,通过实验验证其结果是可靠的,对此我们得到的 alpha、beta 规则库与支持度、可信度的阈值是可以作为知识确定下来,例如:当支持度为 5%,可信度为 30%时,所产生的总规则数为 14 415,后件为中间点结构规则数为 1992,其中支持度小于 10%的规则占规则集的 90.1%,其可信度平均为 56.4%,表现出"支持度低,可信度高"的特点,我们称之为意外规则,并且对于支持度小于 2%的规则由于适用度较低,有可能干扰决策,影响规则的鲁棒性,甚至造成预测模型的鲁棒性降低,我们一般不进行挖掘。

近年来国际上有关此研究的典型文献,从精度的角度进行对比(见表 1)。

表 1　典型文献对比

序号	RS126	CB513	序号	RS126	CB513
文献[15]	78.8%	/	文献[17]	69.8%	69.6%
文献[16]	/	78.44%	文献[18]	71.2%	75.2%

而我们利用 ESKD 系统进行蛋白质二级结构预测研究中,达到如下预测结果:在 ILP 相应的数据库 Q3 精度达 93.88%;在 RS126 数据库 Q3 精度达 84.1%;在 CB513 数据库 Q3 精度达 80.49%,均处国际领先水平。在取得此阶段性成果后,在优化层还有进一步提高预测精度的空间。

5　结　论

为了解决"知识自动获取"这一专家系统中的瓶颈问题,本文在我们提出的基于数据库与知识库的双库协同机制的基础上,提出了专家系统中基于认

知的知识获取机制。该机制的核心思想是把知识发现系统视为认知系统,研究专家系统自身的潜在规律(机理),改变知识发现的固有流程,形成新的知识发现过程模型,然后再用创新的知识发现机理以及模型作为专家系统新的知识获取构件,从而丰富和提升了经典专家系统的知识库结构,形成了全新的动态知识库系统。通过对比实验证明,把基于认知的知识获取机制融入到专家系统中去,在解决"知识自动获取"问题上确实能取得很好的效果。本文认为,基于认知的知识获取机制研究对于专家系统的理论研究具有重要意义,将有可能对新一代专家系统的发展起到重要的推动作用。

参考文献

[1] Han J W, Altman R B, Kumar V, et al. Emerging scientific applications in data mining. *Communications of the ACM*, 2002, 45(8): 54-58

[2] Chen M S, Han J, Yu P S. Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 1996, 8(6): 866-883

[3] Sarabjot A, David B, John H. EDM: A general framework for data mining based on evidence theory. *Data & Knowledge Engineering*, 1996, 18: 189-223

[4] Gray Ray Livingston. A framework for autonomous knowledge discovery from databases. [Ph.D. dissertation], Pittsburgh: University of Pittsburgh, 2001.57-68

[5] Heikki Mannila. Theoretical frameworks for data mining. *SIGKDD Explorations*, 2000, 1 (2): 30-32

[6] Renato C. A theoretical framework for data mining: the "informational paradigm". *Computational Statistics & Data Analysis*, 2002, 38: 501-515

[7] 杨炳儒. 知识工程与知识发现. 北京:冶金工业出版社, 2000,486-501

[8] Yang B R. Knowledge discovery based on inner mechanism: construction, realization and application. USA: Elliott & Fitzpatrick 2004, 101-124

[9] 杨炳儒. 基于内在机理的知识发现理论及其应用. 北京:电子工业出版社, 2004.121-134

[10] 杨炳儒, 宋威, 徐章艳. 基于内在认知机理的知识发现理论及其应用. 自然科学进展, 2005, 15(12): 107-115

[11] 杨炳儒,宋威,徐章艳. 基于知识发现创新技术的专家系统新构造. 中国科学(E 辑),2007,37(6):738-749

[12] Brasil L M, Azevedo F M, Barreto J M. Hybrid expert system for decision support in the medical area. *Medical and Biological Engineering and Computing*, 1999, 37(2):738-739

[13] Yang B R. Fia and case based on fuzzy language field. *Fuzzy Sets and Systems*, 1998, 95(1): 83-89

[14] Muggleton S H, King R, Sternberg M. Protein secondary structure prediction using logic-based machine learning. *Protein Engineering*, 1992, 5(7):647-657

[15] Hu H J, Pan Y, Senior M. Improved protein secondary structure prediction using support vector machine with a new encoding scheme and an advanced tertiary classifier. *IEEE Transactions Nanobioscience*, 2004, 3(4):589-602

[16] Wang L H, Liu J. Predicting protein secondary structure by a support vector machine based on a new coding scheme. *Genome Informatics*, 2004, 15(2):181-190

[17] Liu Y, Carbonel J, Seetharaman J K. Context sensitive vocabulary and its application in protein secondary structure prediction. Vanathi Gopalakrishnan USA Pittsburgh, PA 152132582, USA 2004

[18] Guo J, Chen H, Sun Z R, et al. A novel method for protein secondary structure prediction using dual-Layer SVM and profiles. *PROTEINS: Structure, Function, and Bioinformatics Wiley-Liss*, 2004,54:738-743

A cognition based automatic knowledge acquisition mechanism for expert systems

Yang Bingru[*], Tang Zhigang[*,**], Yang Jun[*]

([*] School of Information Engineering, University of Science and Technology Beijing, Beijing 100083)

([**] School of Math and Physics, Nanhua University, Hengyang 421001)

Abstract

The research was carried out to solve the bottleneck problem of automatic acquisition of knowledge in expert systems. Starting from the potential law (mechanism) of an expert system itself, the study tried to change the inherent knowledge discovery process, form the new knowledge discovery process model, and build a cognition-based automatic knowledge acquisition mechanism. Simultaneously, by using the T-coordinator, it automatically started the directed excavation of knowledge, according to the knowledge shortages of the basic knowledge, to effectively overcome the limitations that could become the new knowledge and minimize the workload of evaluation, thus forming a mechanism for automatic acquisition of knowledge. This mechanism can solve, to a great extent, the bottleneck problem of automatic acquisition of knowledge in intelligent systems.

Key words: double-base cooperating mechanism, double-basis fusion mechanism, knowledge discovery, automatic knowledge acquisition, expert system