学 术 论 文

KDD 中双库协同机制的研究（Ⅰ）

杨炳儒，王建新
（北京科技大学信息工程学院，北京 100083）

[摘要] 针对 KDD（基于数据库的知识发现）主流发展中存在的典型问题提出了用知识库去制约与驱动数据库，并通过数据库改善知识库结构的知识发现的新思想，形成了具有双库协同机制的 KDD 的开放系统 KDD*，从而提高了知识发现的速度、精度和认知自主性，并使知识库在结构上具备了实时维护与自我进化的能力，同时阐述了作为双库协同基础的数据库和知识库在本质上的对应关系。

[关键词] 知识结点；本原知识库；本原数据库；数据子类结构；双库协同机制
[中图分类号] TP180 [文献标识码] A [文章编号] 1009－1742（2002）04－0041－11

1 引论

当前 KDD（基于数据库的知识发现）发展的主流是寻求在各类数据库和应用问题的背景下高性能、高扩展性的发掘算法。笔者另辟蹊径，试图将 KDD 作为一个认知系统和智能体研究其内在的机理。这种内在机理的研究有助于当前发展的主流，有助于解决 KDD 所面临的若干挑战和难题，有助于扩展现有 KDD 的功能。基于此，笔者杨于 1997 年从知识发现、认知科学与智能系统交叉结合的角度，独立地提出了双库协同机制，并构建了将 KDD 与双库协同机制相结合的 KDD* 结构，在结构与功能上形成了相对于 KDD 而言的一个开放的、优化的扩体。此外还相继提出了双基融合机制，信息扩张机制及其扩展性结构模型等，可能形成一个新的内在机理研究的新方向。

当我们做出了若干成果去查新文献的时候发现了一些零散的提法，如 1996 年，S. S. Anand 等，在提出的基于证据理论的数据发掘一般框架 ESD 中，提及了“用户的先验知识与先前发现的知识可以耦合到发现过程中”[1]；1992 年，在 G. Piatesky-Shapiro 等开发的知识发现平台 KDW 中提出过“采用领域知识辅助初始发现的聚焦，限制性的搜索”的思想[2]；1993 年，J. P. Yoon 与 L. Kerschberg 提出一个数据库中知识发现与进化的概念，提出使用正反两个方面的例子来发现新旧知识的协调一致，以及知识与数据库同步进化的思想[3]。然而，他们都没有系统地研究其理论基础与具体实现方法。我们经过三年多的科研实践，从理论上基本解决了双库协同的机理及其技术实现，开发出了相应的软件，并初步用于实际领域中[4~7]。系列型论文（Ⅰ），（Ⅱ）（KDD 中双库协同机制的研究（Ⅰ），（Ⅱ））将给出双库协同机制的内在规律性、对 KDD 主流发展的作用及其若干应用性的结果。

双库协同的含义是什么呢？这里给出其非形式化的描述。在给定真实数据库和基础知识库的前提下，在数据发掘过程中，称具备以下特征的 KDD 中的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发掘数据库的可达范畴与基于属性间关系的

[收稿日期] 2001－12－13；修回日期 2002－02－08
[基金项目] 国家自然科学基金资助项目（69835001）；教育部科技重点资助项目（[2000] 175）
[作者简介] 杨炳儒（1943－），男，天津市人，北京科技大学教授，博士生导师

发掘知识库的推理范畴之间建立等价关系，两个范畴的等价关系为定向发掘和定向搜索（非全局性空间搜索）奠定了理论基础。

2）在KDD聚焦过程中，除依据用户需求确定聚焦外，通过启发协调算法可以形成依发掘知识库中知识短缺而生成的系统自身提供的聚焦方向，进而形成在数据库中的定向发掘（算法和进程）。

3）在获得假设规则到知识评价的过程中产生中断进程，即先不对假设规则进行评价，而是通过中断协调算法到发掘知识库中进行定向搜索（算法和进程），以期发现产生的假设规则与基础知识库（不包含挖掘出的新知识）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。

4）知识库的结构不是单独由人为因素决定的，而是参照数据库中的数据客观地、量化地决定的，并且，随着数据库中数据的积累，知识库的结构也随之动态变化，从而，知识库具有了在内容和结构上自我进化的能力。

笔者旨在完成对1（理论基础）的讨论，论述的核心是建立如图1所示的架构，下面将分述之。而对2、3及4（实现策略等）将在另文（Ⅱ）中讨论。

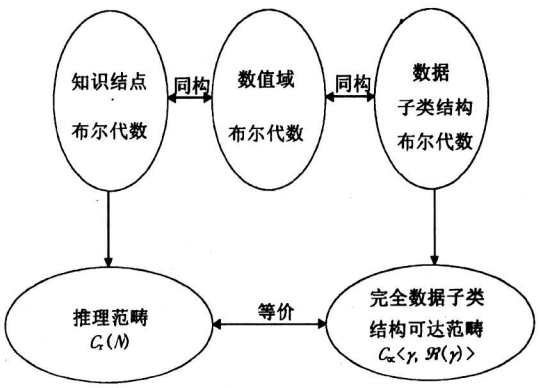

图1 有关论域 X 的5个结构的关系

Fig.1 Corresponding relation between 5 structure of the universe of discourse X

2 三个布尔代数及其关系

2.1 数值域布尔代数

2.1.1 *数值域* 论域 X 的每一个属性都对应一个“数值域”。数值域可以是有限集，也可以是无限集；如果数值域是无限集，则它可以是连续的，也可以是离散的，也可以是连续和离散兼有的。例如“温度”这个属性对应的数值域可以是（15，100］这个实的半开区间。“性别”这个属性对应的数值域是｛男，女｝这个离散的集合。但要求每一个属性对应的数值域是一个序集（全序集）。

属性 X_i 的数值域记为 D_i，$i=1, 2, \cdots, s$。论域 X 的一个采样是指从论域 X 测量、收集的一个 s 维向量，其每一维上的分量属于相应属性的数值域。特别地，论域 X 的一个现象是指论域 X 的一个无误差的或属于允许误差范围内的采样。

把论域 X 的所有属性各自对应的数值域的笛卡儿乘积称为论域 X 对应的数值域，记作 D。即

$$D = D_1 \times D_2 \times \cdots \times D_s。$$

2.1.2 *数值子域* 因为数值域是一个集合，所以可以对它进行划分。依据不同的需要，划分的方式也不一样。划分后，数值域的子集称为数值子域。

定义1 把数值域 D_i 划分为 t_i 个数值子域 D_{i1}，D_{i2}，$\cdots$，D_{it_i}，当这个划分满足如下条件时，称为数值域 D_i 的正则划分：

1）某一数值域 D_i 的所有数值子域的并为数值域 D_i 本身，即：

$$\bigcup_{j=1}^{t_i} D_{ij} = D_i,\ i = 1,2,\cdots,s。$$

2）某一数值域的任意两个数值子域的交为空，也就是对任意 $j_1 \neq j_2$；$1 \leqslant j_1, j_2 \leqslant t_i$，有

$$D_{ij_1} \cap D_{ij_2} = \varnothing。$$

3）数值域 D_i 中元素的序和由其产生的数值子域 D_{i1}，D_{i2}，$\cdots$，D_{it_i} 的序必须满足如下两种关系之一：

a. 数值子域依数值域中元素顺序按照升序排列，即对任意的 v_1，$v_2 \in D_i$，$v_1 \leqslant v_2$，若 $v_1 \in D_{ij_1}$，$v_2 \in D_{ij_2}$，则有 $j_1 \leqslant j_2$。

b. 数值子域依数值域中元素顺序按照降序排列，即对任意的 v_1，$v_2 \in D_i$，$v_1 \leqslant v_2$，若 $v_1 \in D_{ij_1}$，$v_2 \in D_{ij_2}$，则有 $j_1 \geqslant j_2$。

2.1.3 *数值域布尔代数* 对每一个数值域的正则划分可以构成一个拓扑空间。对 $i=1, 2, \cdots, s$，令

$$E_i = \{D_{i1}, D_{i2}, \cdots, D_{it_i}\},$$

则 E_i 是由若干个数值子域构成的集合，称为属性 X_i 的属性子域族。E_i 的任意子集称为子域族。显然，属性子域族是子域族的特殊情形。子域族的每

一个元素是一个数值子域。

属性 X_i 的属性子域族 E_i 的所有子集形成它的幂集 γ_i。幂集 γ_i 包含空集$\varnothing$、属性子域族 E_i 和所有其他由属性 X_i 的若干个数值子域构成的域族。显然，$\langle E_i, \gamma_i\rangle$ 是一个拓扑空间，称为属性 X_i 的数值子域划分拓扑空间。由于不考虑由属性的数值域按照自然的幂集形成的拓扑空间，所以把属性的数值子域划分拓扑空间简称为数值域拓扑空间。对于任意 $y\in\gamma_i$，$y = \{D_{ij} \mid D_{ij}\in E_i, j\in J\subseteq \{1, 2, \cdots, t_i\}\}$，定义 y 的真集为

$$\psi(y) = \bigcup_{j\in J} D_{i,j}。$$

显然，真集 $\psi(y)$ 是数值域 D_i 的子集。

由于拓扑空间的乘积仍然是拓扑空间，故 $\langle E_1, \gamma_1\rangle \times \langle E_2, \gamma_2\rangle \times \cdots \times \langle E_s, \gamma_s\rangle$ 也是一个拓扑空间，记这个乘积拓扑空间为 $\langle E, \gamma\rangle$，称为论域 X 的数值域划分拓扑空间，简称为论域 X 的数值域拓扑空间。

在论域的数值域拓扑空间中，存在着这样的元素：它是 s 个数值子域形成的序组，其中每一个数值子域取自相应的属性子域族。称这样的序组为数值域拓扑空间的基本元，由所有的基本元形成的集合称为基本元集。基本元集如下式所示：

$$\{(d_1, d_2, \cdots, d_s) \mid d_1 \in E_1, d_2 \in E_2, \cdots, d_s \in E_s\}。 \quad (1)$$

把基本元集记为 F。F 中元素的数目为

$$v = t_1 \times t_2 \times \ldots \times t_s。$$

在文中，v 的意义保持不变，并且，把集合 $\{1,2,\cdots,v\}$ 记为 Δ。可以把基本元集排成一列，记为 $F_1, F_2, \cdots, F_v$。

由乘积拓扑空间的性质，γ 是由 F 形成的幂集。因此，γ 中元素的数目为 2^v 个。对任意的 $y\in\gamma$，y 有如下的表示形式：

$$y = \{F_i \mid F_i \in F, i \in I \subseteq \Delta\}。 \quad (2)$$

与在数值子域拓扑空间中定义真集的方式类似，可以定义数值域拓扑空间幂集 γ 中的元素的真集。首先，对任意的基本元 $y = (d_1, d_2, \cdots, d_s)$，定义 y 的真集为

$$\psi(y) = d_1 \times d_2 \times \cdots \times d_s,$$

其中“×”为欧几里德乘积。$\psi(y)$ 不再是一个序组，而是数值域 D 中的一个子集。由此，可以定义 γ 中任意元素的真集。对任意 $y\in\gamma$，$y = \{F_j \mid F_j \in F, j \in J\}$，$y$ 的真集定义为 $\psi(y) = \bigcup_{j\in J} F_j$，其中“∪”号为数值域中通常意义下集合间的并。显然，由论域的数值域拓扑空间定义的任意真集是数值域的子集。

由真集定义可以得出，对任意 $y, y_1, y_2\in\gamma$，

$$\psi(y_1 \cap y_2) = \psi(y_1) \cap \psi(y_2); \quad (3)$$

$$\psi(y_1 \cup y_2) = \psi(y_1) \cup \psi(y_2); \quad (4)$$

$$\psi(\sim y) = \sim \psi(y)。 \quad (5)$$

但是，在式 (3)、式 (4) 和式 (5) 中，等式左边的“∩”号、“∪”号和“～”号与等式右边的符号意义是不一样的。左边的符号表示对普通的集合（数值域中的集合）进行的操作，而右边表示对集族进行的操作。通过真集，可以建立数值域和它的拓扑空间之间的联系。

对任意的 $y_1, y_2\in\gamma$，根据式 (2)，y_1 和 y_2 可以分别表示成 $y_1 = \{F_i \mid F_i \in F, i \in I_1 \subseteq \Delta\}$，$y_2 = \{F_i \mid F_i \in F, i \in I_2 \subseteq \Delta\}$。于是，$y_1 \cap y_2 = \{F_i \mid F_i \in F, i \in I_1 \cap I_2 \subseteq \Delta\}$，即 $y_1 \cap y_2 \in \gamma$，同理可得 $y_1 \cup y_2 \in \gamma$，$\sim y_2 \in \gamma$。所以，在集族 γ 中，元素之间的交运算、并运算以及元素的补运算的结果仍然是 γ 中的元素，即交、并、补运算在 γ 中是封闭的。

定理 1　集族 γ 及其元素间的交运算“∩”和并运算“∪”以及元素的补运算“～”构成一个代数系统（γ，∩，∪，～），并且这个代数系统是一个布尔代数，这个布尔代数的零元是$\varnothing$；幺元是基本元集 F。（证明略）

把上述由论域 X 的数值域拓扑空间的幂集 γ，按照其元素间的交运算、并运算以及元素的补运算形成的布尔代数，称为论域 X 的数值域布尔代数。

2.2　知识结点布尔代数

2.2.1　*属性程度词*　程度词是描述论域中属性状态的词语，即语言值。对温度而言，“很高”、“高”、“中等”、“低”和“很低”等都是程度词。

属性程度词是把某一属性和它的一个程度词放在一起，表示该属性的某种状态的词语。例如，“温度很低”是一个属性程度词。假设论域 X 中的每一个属性的属性程度词有有限个。对 $i = 1, 2, \cdots, s$，属性 X_i 含有 $t_i (t_i \geqslant 2)$ 个程度词。把这 t_i 个属性程度词按照程度的升序或降序排列，分别记为 $A_{i1}, A_{i2}, \cdots, A_{it_i}$

把属性 X_i 的所有属性程度词构成的集合记为 B_i；而把论域 X 的所有属性程度词构成的集合记

为B。显然，$B=\bigcup_{i=1}^{s}B_i$。

两个属性程度词集合在一起，中间添加适当的析取或合取关系，会形成新的意义。例如，假设属性程度词a为“温度高”，b为“压强大”，则$a\wedge b$表示“温度高且压强大”，$a\vee b$表示“温度高或压强大”。

任意一个属性程度词A_{ij}的非表示与属性程度词相反的意义。如果属性程度词A_{ij}表示“温度很高”，则A_{ij}的非表示“温度很高”的非，定义A_{ij}的非为

$$\neg A_{ij}=A_{i1}\vee\cdots\vee A_{i,j-1}\vee A_{i,j+1}\vee\cdots\vee A_{it_i}。\tag{6}$$

容易看出，这个对非的定义与通常二值逻辑意义下对非的定义是不同的。

2.2.2 *知识结点*

定义2 相应于论域X的知识结点，是指如下不含否定联结词的合式公式：

$$\theta_0a_1\theta_1a_2\cdots\theta_{m-1}a_m\theta_m\tag{7}$$

其中：$a_i\in B$，$i=1,2,\cdots,m$；$\theta_i\in J$，$i=0,1,\cdots,m$。这里J是由符号“$\wedge$”、“$\vee$”、“(”、“)”等4个符号及其相应的组合以及NOP（空）而形成的集合；但θ_i在其中取值要使合式公式（7）有意义；只有θ_0和θ_m可以取空。合式公式应记作F。

例如，这里取$m=3$；a_1=温度很高，a_2=温度高，a_3=压强小；θ_0：(，θ_1：$\vee$，θ_2:)$\wedge$结构，θ_3：NOP。于是$\theta_0a_1\theta_1a_2\theta_2a_3\theta_3$是一个知识结点，它表示（温度很高$\vee$温度高）$\wedge$压强小，即同时发生或同时存在“温度很高或温度高”和“压强小”这件事情或这种状态。

当$m=0$时，式（7）为空，记为$\varnothing$，它表示空知识结点。属性X_i的所有属性程度词的析取是一个很重要的知识结点，称为X_i的析取总项，它表示该属性不作任何限制的状态，记这个知识结点为U_i，$i=1,2,\cdots,s$。即$U_i=A_{i1}\vee A_{i2}\vee\cdots\vee A_{it_i}$。与空知识结点相反，由论域$X$的所有属性程度词的析取构成的知识结点称为全知识结点，记为Ω，即

$$\Omega=\vee_{a\in B}a。$$

论域X的所有知识结点构成的集合称为论域X的知识结点集，记为N。

2.2.3 *知识结点集的结构* 在知识结点集N上，可以定义知识结点的非以及知识结点间的合取、析取。对任意的n_1，$n_2\in N$，设$n_1=F_1$，$n_2=F_2$，自然地定义知识结点n_1与n_2的合取、析取分别为：$n_1\wedge n_2=F_1\wedge F_2$；$n_1\vee n_2=F_1\vee F_2$。

定义知识结点$n=\theta_0a_1\theta_1a_2\cdots\theta_{m-1}a_m\theta_m$的非为$\neg n=\neg(\theta_0a_1\theta_1a_2\cdots\theta_{m-1}a_m\theta_m)=\theta_0'a_1'\theta_1'a_2'\cdots\theta_{m-1}'a_m'\theta_m'$。其中$\theta_i'\in J$，$i=0,1,\cdots,m$。当$a_i'$是肯定型的属性程度词时，令其保持不变；当$a_i'$是否定型属性程度词时，把它代换成式（6）右端的形式。经过这样的代换之后，$\neg n$仍然是一个不含否定联结词的合式公式。

从知识结点的三个运算的定义可以得出，两个知识结点的合取和析取仍然是知识结点，一个知识结点的非仍然是知识结点。从而，合取、析取和非运算在知识结点集中是封闭的。

范式定理 任一知识结点$n=\theta_0a_1\theta_1a_2\cdots\theta_{m-1}a_m\theta_m$均可唯一等价地表示为主析取范式的形式：

$$m=L_1\vee L_2\vee\cdots\vee L_y。\tag{8}$$

其中对$i=1,2,\cdots,y$，$L_i=u_{i1}\wedge u_{i2}\wedge\cdots\wedge u_{iw_i}$，均是单个属性程度词合取的形式，其中的项称为成员项。成员项$u_{ij}\in\{a_1,a_2,\cdots,a_m\}$，$j=1,2,\cdots,w_i$。事实上，$u_{i1}$，$u_{i2}$，$\cdots$，$u_{iw_i}$中的来自于同一属性的任意两个成员项$u_{ij_1}$，$u_{ij_2}$有且仅有两种关系之一：

1）如果$u_{ij_1}\neq u_{ij_2}$，则由于$u_{ij_1}\wedge u_{ij_2}=\varnothing$，合取式$L_i$为$\varnothing$，项$L_i$就可以去掉。

2）如果$u_{ij_1}=u_{ij_2}$，则由于$u_{ij_1}\wedge u_{ij_2}=u_{ij_1}$，合取式$L_i$经过这样的合并后，不存在相同的属性词，$L_i$经合并后仍记为$L_i$。于是，合并整理后的$L_i$成为简单合取式；并且知识结点$n$中简单合取式的个数不大于$y$，不失一般性仍记为$y$；每一个简单合取式中属性程度词的数目不大于$s$个，把合并整理后的每一个合取式中的属性程度词按照所归属的属性的顺序排列。若L_i中缺少某一个属性X_j的成员项，那么把该属性的析取总项$A_{j1}\vee A_{j2}\vee\cdots\vee A_{jt_j}$人为地添加到相应的位置（知识结点的意义不变）。应当注意的是，替代以后的L_i不是简单合取式，但可以利用分配律把它变换成若干简单合取式的析取。最后得到知识结点n是若干个含有s个命题变元的简单合取式的析取：

$$n=\bigvee_{i=1}^{y}\left(\bigwedge_{j=1}^{s}u_{ij}\right),\tag{9}$$

其中$u_{ij}\in B_j$。

定义 3　称知识结点的式（9）所示的表示形式为知识结点的标准形，相应的结点称为标准知识结点。

由式（9）可以看出，标准知识结点是由一些简单合取式析取而得，每一个简单合取式有 s 个成员项，每个成员项依次取自相应的属性的程度词集。这样的简单合取式称为论域 X 的基本简单合取式。所有的基本简单合取式的全体构成的集合称为基本简单合取式集，记为 H。容易看出，H 中元素的数目为前面定义的 v。

把 H 中所有的元素依次记为 H_1，H_2，…，H_v。于是对任意 $n\in N$，有

$$n = \vee_{j\in J} H_j。 \tag{10}$$

其中 J 是集合 Δ 的子集。

引理 1　论域 X 中任意两个不相同的基本简单合取式的析取为空知识结点。（证明略）

推论　若两个知识结点的标准型中没有相同的基本简单合取式，则这两个知识结点的合取为空知识结点。

定理 2　论域 X 的任意知识结点均可以表示成式（9）所示的标准形；并且标准形是唯一的。（证明略）

表示成标准形后，知识结点间的析取、合取运算以及知识结点的取非运算都有简单的形式：设 n，n_1，$n_2\in N$，$n = \vee_{i\in I} H_i$，$n_1 = \vee_{i\in I_1} H_i$，$n_2 = \vee_{i\in I_2} H_i$，其中 I，I_1，I_2 均是包含于 Δ 的指标集，容易证明：

$$n_1 \wedge n_2 = \vee_{i\in I_1\cap I_2} H_i; \tag{11}$$

$$n_1 \vee n_2 = \vee_{i\in I_1\cup I_2} H_i; \tag{12}$$

$$\neg n = \vee_{i\in C\setminus I} H_i。 \tag{13}$$

其中 $H_i\in H$。

在得出标准知识结点的合取、析取和取非运算分别如式（11）、式（12）和式（13）所示后，可以更清楚地看出知识结点集的结构及其三种运算的封闭性。容易证明，$(N, \wedge, \vee, \neg)$ 是一个代数系统。

定理 3　$(N, \wedge, \vee, \neg)$ 是一个布尔代数，它的零元是空知识结点 Φ；幺元是全知识结点 Ω。（证明略）

把这个由论域的知识结点集按照其元素间的合取运算、析取运算以及元素的非运算构成的布尔代数称为知识结点布尔代数。

2.3　数据子类结构布尔代数

2.3.1　数据子类结构

定义 4　对给定的论域 X，建立如下的关系数据库模式：

$$\mathcal{R}(\mathbf{N}, X_1, X_2, \cdots, X_s)。 \tag{14}$$

其中：$\mathbf{N}$ 是主码，它取自自然数集，能唯一地识别一个元组；X_i 是论域 X 的属性，$i=1, 2, \cdots, s$。在模式（14）上建立的关系数据库称为论域 X 的数据库，记为 $\mathcal{R}(X)$。数据库 $\mathcal{R}(X)$ 中的任意一个元组 u 是一个 $s+1$ 维向量 $(num, x_1, x_2, \cdots, x_s)$ 的形式，把采样 $(x_1, x_2, \cdots, x_s)$ 记为 $a(u)$。$\mathcal{R}(X)$ 的所有元组的全体记为 U，即 $U = \{u_j \mid j\in I\}$，I 是总指标集。

对拓扑空间 $\langle F, \gamma\rangle$ 中的任意开集 $y\in\gamma$，可以定义一个元组集 $\{u \mid u\in\mathcal{R}(X), a(u)\in\psi(y)\}$，并记这个元组集为 $\langle y, \mathcal{R}(y)\rangle$。称这样的元组集为论域 X 的数据子类结构。y 称为数据子类结构 $\langle y, \mathcal{R}(y)\rangle$ 的数据部分，$\mathcal{R}(y)$ 称为数据子类结构 $\langle y, \mathcal{R}(y)\rangle$ 的元组部分。论域 X 的所有数据子类结构形成数据子类结构集，记为 $\langle\gamma, \mathcal{R}(\gamma)\rangle$。

2.3.2　数据子类结构集的结构

在数据子类结构集中，可以定义两个数据子类结构的相等关系为数据部分和元组部分都分别对应相等。即：$\langle y_1, \mathcal{R}(y_1)\rangle = \langle y_2, \mathcal{R}(y_2)\rangle$ 当且仅当 $y_1 = y_2$ 和 $\mathcal{R}(y_1) = \mathcal{R}(y_2)$ 分别成立。据此，不同的数据子类结构可以具有相同的元组集。

下面可以建立数据子类结构集的元素运算关系。对数据子类结构集 $\langle\gamma, \mathcal{R}(\gamma)\rangle$ 中的任意 3 个元素（可以是相等的元素）$\langle y_1, \mathcal{R}(y_1)\rangle$，$\langle y_2, \mathcal{R}(y_2)\rangle$ 和 $\langle y, \mathcal{R}(y)\rangle$，定义

$$\langle y_1, \mathcal{R}(y_1)\rangle \cap \langle y_2, \mathcal{R}(y_2)\rangle = \langle y_1 \cap y_2, \mathcal{R}(y_1) \cap \mathcal{R}(y_2)\rangle, \tag{15}$$

$$\langle y_1, \mathcal{R}(y_1)\rangle \cup \langle y_2, \mathcal{R}(y_2)\rangle = \langle y_1 \cup y_2, \mathcal{R}(y_1) \cup \mathcal{R}(y_2)\rangle, \tag{16}$$

$$\sim \langle y, \mathcal{R}(y)\rangle = \langle \sim y, \sim \mathcal{R}(y)\rangle。 \tag{17}$$

应当注意的是，式(15) 中的三个“$\cap$”号的意义两两不相同，对式(16) 和式(17) 中的“$\cup$”号和“$\sim$”号也是如此。从定义式(15)、式(16) 和式(17) 可以看出，数据子类结构的“并”、“交”和“补”运算在数据子类结构集中是封闭的。容易证明 $(\langle\gamma, \mathcal{R}(\gamma)\rangle, \cap, \cup, \sim)$ 构成一个代数系统。

定理 4　$(\langle\gamma, \mathcal{R}(\gamma)\rangle, \cap, \cup, \sim)$ 构成一个布尔代数，其中零元为 $\langle\varnothing, \mathcal{R}(\varnothing)\rangle$，幺元为 $\langle\Omega,$

$\mathscr{R}(\Omega)\rangle$。(证明略)

把这个由数据子类结构集及其元素间的交运算和并运算以及元素的补运算形成的布尔代数称为数据子类结构布尔代数。

2.4 三个布尔代数的关系

2.4.1 数值域布尔代数和知识结点布尔代数的关系 在2.1.2节中给出了数值域正则划分的定义。从属性程度词的定义可以看出，任意一组属性程度词可以唯一确定数值域的一个正则划分。事实上，对论域 X 的任意一个属性，如果令每一个属性程度词对应一个数值的范围，并规定这些范围互不相交，这些范围的并为整个属性的数值取值范围，并且这些范围按升序或降序排列，则该属性的所有属性程度词确定了数值子域的一个正则划分。

例如，假设属性“温度”的数值域为[0,100]这个闭区间，如果定义属性程度词“温度很低”对应区间[0,10]，“温度低”对应(10,35]，“温度中等”对应(35,65]，“温度高”对应(65,90]，“温度很高”对应(90,100]，则属性“温度”的5个属性程度词决定了数值域[0,100]的一个正则划分。

通过属性程度词和数值域正则划分的关系，就可以建立知识结点和数值域拓扑空间之间的关系，有如下定理。

定理5 论域 X 的数值域布尔代数(γ, $\cap$, $\cup$, $\sim$)和知识结点布尔代数(N, $\wedge$, $\vee$, $\neg$)之间存在着同构关系。

证明 首先，对式(1)所示的基本元 $y=(d_1, d_2, \cdots, d_s)$ 和式(9)中所含的基本简单合取式 $z = a_1 \wedge a_2 \wedge \cdots \wedge a_s$，建立如下的映射关系：

$$f: F \to H,$$

使得

$$f((d_1, d_2, \cdots, d_s)) = a_1 \wedge a_2 \wedge \cdots \wedge a_s。$$

其中 d_j 对应的属性程度词为 a_j。易证这个映射关系是一一对应。为了叙述方便，改变基本元或基本简单合取式的序号，从而可以使得

$$f(F_j) = H_j,$$

其中 $j=1, 2, \cdots, v$。

然后，就可以把 f 扩张到整个的幂集 γ 上：

$$f: \gamma \to N,$$

使得对任意的 $y\in\gamma$，$y = \{F_j \mid j\in J\subseteq\Delta\}$，有

$$f(y) = \vee_{j\in J} H_j,$$

易证这个扩张之后的映射关系仍然是一一对应。然后证明这两个布尔代数之间存在着保运算的关系。事实上，对任意的 $y_1, y_2\in\gamma$，$y_1 = \{F_j \mid j\in J_1\}$，$y_2 = \{F_j \mid j\in J_2\}$，有

$$f(y_1 \cap y_2) = f(\{F_j \mid j \in J_1\} \cap \{F_j \mid j \in J_2\}) =$$

$$f(\{F_j \mid j \in J_1 \cap J_2\}) = \vee_{j\in J_1\cap J_2} H_j =$$

$$(\vee_{j\in(J_1\setminus(J_1\cap J_2))\cup(J_1\cap J_2)} H_j) \wedge$$

$$(\vee_{j\in(J_2\setminus(J_1\cap J_2))\cup(J_1\cap J_2)} H_j) =$$

$$(\vee_{j\in J_1} H_j) \wedge (\vee_{j\in J_2} H_j) = f(y_1) \wedge f(y_2)。$$

上面的推导过程中用到了引理1及其推论。同理可以证明

$$f(y_1 \cup y_2) = f(y_1) \vee f(y_2)$$

和 $f(\sim y_1) = \neg f(y_1)$。

于是 f 这个一一映射使两个布尔代数之间保运算，从而这两个布尔代数同构，并且 γ 的零元 Φ 对应着 N 的零结点 Φ；γ 的幺元 F 对应着 N 的全结点 Ω。同时可以得到映射 f 的逆映射 f^{-1} 存在。

2.4.2 三个布尔代数的关系 在2.3.1节中，通过数值域拓扑空间中的元素定义了元组集。根据这个定义，有：

定理6 数值域布尔代数和数据子类结构布尔代数之间存在着同构关系。

证明 首先，建立数值域布尔代数和数据子类结构布尔代数之间的映射关系：

$$g: \gamma \to N$$

使得对于任意 $y\in\gamma$，

$$g(y) = \langle y, \mathscr{R}(y)\rangle。$$

g 显然是一个一一对应。于是，对于任意 y, y_1, $y_2\in\gamma$，根据式(15)、式(16)和式(17)，有

$$g(y_1 \cap y_2) = \langle y_1 \cap y_2, \mathscr{R}(y_1 \cap y_2)\rangle =$$

$$\langle y_1, \mathscr{R}(y_1)\rangle \cap \langle y_2, \mathscr{R}(y_2) = g(y_1) \cap g(y_2)。$$

同理可得

$$g(y_1 \cup y_2) = g(y_1) \cup g(y_2)$$

$$和\ g(\sim y) = \sim g(y)。$$

因此，数值域布尔代数和数据子类结构布尔代数之间存在着同构关系，并且，数值域布尔代数的零元$\varnothing$对应数据子类结构布尔代数的零元 $\langle\varnothing, \mathscr{R}(\varnothing)\rangle$；数值域布尔代数的幺元 F 对应数据子类结构的幺元 $\langle\Omega, \mathscr{R}(\Omega)\rangle$。同时可以得到映射 g 的逆映射 g^{-1} 存在。

由定理5和定理6，自然可以得出定理7。

定理7 论域 X 的数值域布尔代数、知识结点布尔代数和数据子类结构布尔代数两两同构。

（证明略）

定理 7 说明了数值域拓扑空间幂集 γ、知识结点集 N、数据子类结构集 $\langle \gamma, \mathscr{R}(\gamma)\rangle$ 具有相同的代数结构。这对于知识结点的产生、数据子类结构的繁衍（见文（Ⅱ））等具有重要意义。

3　两个范畴及其关系

3.1　推理范畴

3.1.1　*推理范畴*　给定论域 X，则它有一个知识结点集 N。例如，假设在论域 X 中有两个知识结点 n_1 = “温度高”，n_2 = “压强大”。若论域 X 中存在着固有的规律是“如果温度高则压强大”，那么知识结点 n_1 到 n_2 有推理关系：$n_1 \rightarrow n_2$，或记为 $r(n_1, n_2) = r$，它表示：如果温度在“温度高”对应的子域上取值，则压强必然在“压强大”对应的子域上取值；若存在着固有的规律是：“‘如果温度高则压强大’不成立”，即 $n_1 {}^{\times}\!\!\rightarrow n_2$，或记为 $r(n_1, n_2) = \phi$，它表示：至少存在并可能出现着一个现象，使它的温度分量在“温度高”对应的数值子域上取值时压强分量不在“压强大”对应的数值子域上取值。

定义 5　称上述的知识结点 n_1 到 n_2 的推理关系 $n_1 \rightarrow n_2$ 为一条正规则，$n_1 {}^{\times}\!\!\rightarrow n_2$ 为一条反规则。n_1 称为规则（正规则或反规则）的始知识结点，n_2 称为规则的终知识结点。论域 X 上所有的正规则和反规则共同构成论域 X 的规则集。

定理 8　论域 X 的知识结点集 N 连同其元素间的推理关系 r 构成一个范畴。

证明　令对象类为 N，态射集 $\hom[n_i, n_j]$ 的元素为推理关系 r 或 $\varnothing$，所以，$\hom[n_i, n_j]$ 或者为空集，或者只有一个元素 r。如果 $s \in \hom[n_h, n_i]$，$t \in \hom[n_i, n_j]$，$u \in \hom[n_j, n_k]$，则有

1）$t \cdot s \in \hom[n_h, n_j]$，这是因为由 $n_h \rightarrow n_i$ 和 $n_i \rightarrow n_j$ 可以推出 $n_h \rightarrow n_j$；

2）$u \cdot (t \cdot s) = (u \cdot t) \cdot s$；

3）$s \cdot l_{ph} = s$，并且 $l_{pi} \cdot s = s$。

这里 $l_{ph} \in \hom[n_h, n_h]$，$l_{pi} \in \hom[n_i, n_i]$。

把 N 连同其元素间的推理关系 r 构成的范畴称为论域 X 的推理范畴，记为 $C_r(N)$。

3.1.2　*本原知识库*

定义 6　论域 X 的知识结点集 N 连同知识结点之间固有的规则集称为论域 X 的本原知识库，记为 $K^p(X)$。

本原知识库真实地反映了论域 X 中知识结点之间的推理关系存在与否：在本原知识库 $K^p(X)$ 中，$r(n_1, n_2) = r$（推理关系存在）当且仅当论域 X 中 $n_1 \rightarrow n_2$；在 $K^p(X)$ 中，$r(n_1, n_2) = \phi$（推理关系不存在）当且仅当论域 X 中 $n_1 {}^{\times}\!\!\rightarrow n_2$。由此，给定论域 X，其本原知识库是唯一确定的。而且，由于反规则集完全可以由正规则集确定，所以如果推理范畴确定，本原知识库也就确定了。

本原知识库 $K^p(X)$ 的规模庞大，规则总数为 2^{2v}。因此，在文（Ⅱ）中提出知识库简约的概念和实现方法。

3.2　完全数据子类结构可达范畴

3.2.1　*数据库和数据子类结构的类型*

定义 7　假设 y 是数值域布尔代数 $\langle \gamma, \mathscr{R}(\gamma)\rangle$ 中的一个元素，$\psi(y)$ 是 y 的真集，$n = f^{-1}(y)$ 是 y 对应的知识结点。称 2.3.1 节定义的数据库 $\mathscr{R}(X)$ 中的元组 u 满足 y（或满足 n），如果

$a(u) \in \psi(y)$，记为 u/y，或者 u/n。

若论域 X 对应的数据库 $\mathscr{R}(X)$ 的任意一个元组 u 对应的向量 $a(u)$ 是论域 X 的一个现象（无误差的采样），则称 $\mathscr{R}(X)$ 为论域 X 的本原数据库，记为 $\mathscr{R}^p(X)$。特别地，在本原数据库中，若对论域 X 的本原知识库 $K^p(X)$ 中的任意一条反规则 $n_1 {}^{\times}\!\!\rightarrow n_2$，$\mathscr{R}^p(X)$ 中均存在元组 u，使得 $a(u)/n_1$ 和 $a(u)/(\neg n_2)$ 都成立，则称 $\mathscr{R}^p(X)$ 为论域 X 的完全数据库，记为 $\mathscr{R}^c(X)$。

由本原数据库产生的数据子类结构称为本原数据子类结构；由完全数据库形成的数据子类结构称为完全数据子类结构。

3.2.2　*可达范畴*

定义 8　在论域 X 的数据子类结构集 $\langle \gamma, \mathscr{R}(\gamma)\rangle$ 上建立元素之间的可达关系“∞”：$\langle y_1, \mathscr{R}(y_1)\rangle \infty \langle y_2, \mathscr{R}(y_2)\rangle$ 当且仅当 $\mathscr{R}(y_1) \subseteq \mathscr{R}(y_2)$。若 $\langle \gamma, \mathscr{R}(\gamma)\rangle$ 中元素 $\langle y_1, \mathscr{R}(y_1)\rangle$ 到元素 $\langle y_2, \mathscr{R}(y_2)\rangle$ 无可达关系，则称 $\langle y_1, \mathscr{R}(y_1)\rangle$ 到 $\langle y_2, \mathscr{R}(y_2)\rangle$ 有非可达关系。所有的可达关系构成可达关系集；所有的非可达关系构成非可达关系集。

由论域 X 的关系数据库 $\mathscr{R}(X)$ 生成的数据子类结构集 $\langle \gamma, \mathscr{R}(\gamma)\rangle$ 连同其元素间的可达关系集和非可达关系集构成论域 X 的数据子类结构库；与本原数据库和完全数据库相对应，可以类似地定义论域 X 的本原数据子类结构库和完全数据子类结构库。

显然，可达关系“$\propto$”是一个偏序。因为任一具有偏序关系的集合均构成一个范畴，所以有：

定理9 论域 X 的数据子类结构集$\langle\gamma,\mathscr{R}(\gamma)\rangle$连同其元素间的可达关系“$\propto$”构成一个范畴。(证明略) 把数据子类结构集$\langle\gamma,\mathscr{R}(\gamma)\rangle$连同其元素间的可达关系构成的范畴称为论域 X 的数据子类结构可达范畴，记为 $C_{\propto}\langle\gamma,\mathscr{R}(\gamma)\rangle$。相应地有：本原数据子类结构可达范畴，记为 $C_{\propto}\langle\gamma,\mathscr{R}^{P}(\gamma)\rangle$；完全数据子类结构可达范畴，记为 $C_{\propto}\langle\gamma,\mathscr{R}^{C}(\gamma)\rangle$。显然，这3个范畴分别可由相应的数据子类结构库唯一确定。

3.3 两个范畴之间的关系——双库结构对应关系定理

引理2 论域 X 的推理范畴 $C_r(N)$ 到数据子类结构可达范畴 $C_{\propto}\langle\gamma,\mathscr{R}(\gamma)\rangle$(包括 $C_{\propto}\langle\gamma,\mathscr{R}\_^{D}(\gamma)\rangle$ 与 $C_{\propto}\langle\gamma,\mathscr{R}^{C}(\gamma)\rangle$ 之间存在函子。

证明 首先，建立论域 X 的知识结点集 N 到数据子类结构集$\langle\gamma,\mathscr{R}(\gamma)\rangle$之间的自然的一一映射 $F_O = g f^{-1}: N \to \langle\gamma,\mathscr{R}(\gamma)\rangle$。其中，$f$ 和 g 的意义如2.4.2节所阐述。当把数据子类结构集换成本原数据子类结构集$\langle\gamma,\mathscr{R}^{P}(\gamma)\rangle$或完全数据子类结构集 $C_{\propto}\langle\gamma,\mathscr{R}^{c}(\gamma)\rangle$ 时，F_O 的意义不变。

对任意$(n\to k)\in \mathrm{Hom}\ C_r(N)$，在元组集$\mathscr{R}(n)$中任取 u，必有 $a(u)\in\psi(f^{-1}(n))$，即 u/n。但由于 u 是本原数据库中的元组，故它必须满足论域 X 本身所固有的属性间的相关规则。由规则的定义，可得 u/k，从而 $a(u)\in f^{-1}(\psi(k))$，即 $u\in\mathscr{R}(f^{-1}(k))$。于是，

$$\mathscr{R}(F_O(n))\subseteq((F_O(k)),$$

从而

$$\langle f^{-1}(n),\mathscr{R}(f^{-1}(n))\rangle\propto\langle f^{-1}(k),\mathscr{R}(f^{-1}(k))\rangle。\tag{18}$$

所以，若有 $n\to k$，就有式(18)成立。

由这个关系得到了一个从正规则集到可达关系集的映射 F_H：

$$F_H(n\to k)=(\langle f^{-1}(n),\mathscr{R}(f^{-1}(n))\rangle\propto\langle f^{-1}(k),\mathscr{R}(f^{-1}(k))\rangle)。$$

证明映射对(F_O,F_H)是一个函子。设任意 $\eta,\zeta\in\mathrm{Hom}\ C_r(N)$，$\eta=(m\to n)$，$\zeta=(n\to k)$。由 F_O 的定义，$F_O(m)=\langle f^{-1}(m),\mathscr{R}(f^{-1}(m))\rangle$，$F_O(n)=\langle f^{-1}(n),\mathscr{R}(f^{-1}(n))\rangle$，$F_O(k)=\langle f^{-1}(k),\mathscr{R}(f^{-1}(k))\rangle$，来验证$(F_O,F_H)$满足函子的4个条件：

1) $F_O(\mathrm{dom}(\eta))=\mathrm{dom}(F_H(\eta))$，由 F_H 的定义，显然成立；

2) $F_O(\mathrm{cod}(\eta))=\mathrm{cod}(F_H(\eta))$，由 F_H 的定义，显然成立；

3) $\mathrm{comp}(\eta,\zeta)\in\mathrm{Hom}\ C_r(N)$，所以 $\mathrm{comp}(F_H(\eta),F_H(\zeta))\in\mathrm{Hom}\ C_{\propto}\langle\gamma,\mathscr{R}^{P}(\gamma)\rangle$，于是，

$$F_H(\mathrm{comp}(\eta,\zeta))=$$
$$F_H(\mathrm{comp}(m\to n,n\to k))=$$
$$F_H(m\to k)=(F_O(m)\propto F_O(k))=$$
$$\mathrm{comp}(F_O(m)\propto F_O(n),F_O(n)\propto F_O(k))=$$
$$\mathrm{comp}(F_H(\eta),F_H(\zeta));$$

4) 对知识结点 n，必有 $n\to n$，因此有 $F_O(n)\propto F_O(n)$，即 $F_H(l(n))=l(F_O(n))$。故(F_O,F_H)是 $C_r(N)$ 到 $C_{\propto}\langle\gamma,\mathscr{R}^{P}(\gamma)\rangle$ 的一个函子。

从引理2可见，若 $C_r(N)$ 中 m 到 n 的推理关系存在，则在 $C_{\propto}\langle\gamma,\mathscr{R}^{P}(\gamma)\rangle$ 中 $F_O(m)$ 到 $F_O(n)$ 的可达关系就存在。下面进一步给出最重要的结构对应定理。

定理10(结构对应定理) 论域 X 的推理范畴 $C_r(N)$ 与完全数据子类结构可达范畴 $C_{\propto}\langle\gamma,\mathscr{R}^{c}(\gamma)\rangle$ 等价。

证明 假设函子(F_O,F_H)的意义如引理2所述。由引理2的证明知，F_O是一个一一映射，故 F_O^{-1} 存在。

证明 F_H 也是一个一一映射。取 $C_{\propto}\langle\gamma,\mathscr{R}^{c}(X)\rangle$ 中的任意一个态射$(F_O(m)\propto F_O(n))$，需证明 $m\to n$。用反证法：若不然，则 $m\overset{\times}{\to}n$。由完全数据库 $\mathscr{R}^{c}(X)$ 的定义，至少存在一个元组 u，使得 u/m 且 $u/\neg n$，即 $u\in\mathscr{R}(f^{-1}(m))$ 但 $u\notin\mathscr{R}(f^{-1}(n))$，也即关系 $\mathscr{R}(f^{-1}(m))\subseteq\mathscr{R}(f^{-1}(n))$ 不成立，从而 $F_O(m)\propto F_O(n)$ 不成立。这与假设$(F_O(m)\propto F_O(n))$是态射矛盾。因此，$m\to n$，所以 F_H^{-1} 存在。

易证，(F_O^{-1},F_H^{-1}) 是 $C_{\propto}\langle\gamma,\mathscr{R}^{c}(\gamma)\rangle$ 到 $C_r(N)$ 的一个函子。所以 $C_r(N)$ 与 $C_{\propto}\langle\gamma,\mathscr{R}^{c}(\gamma)\rangle$ 等价。

至此，已阐述了由论域 X 引出的5个代数结构及其之间的关系(见图1)。其中，3个布尔代数是由论域的属性和对数值域的划分决定的，是形式的；两个范畴是由论域中各个属性之间固有的相关规律决定的，是内容的。

4　进一步的讨论

4.1　基本误差概率

定理 10 的结论是：在确定完全数据子类结构可达范畴时，本原知识库就随之确定。但实际上，只能由论域的数据库得到数据子类结构库。因而，前面的讨论意味着，对本原知识库的实现要经过如下几个步骤：

1）数据子类结构库→数据子类结构可达范畴；

2）数据子类结构可达范畴→完全数据子类结构可达范畴；

3）完全数据子类结构可达范畴→推理范畴；

4）推理范畴→本原知识库。

其中，第 1、3、4 步是确切的，第 2 步是近似的；主要原因是数据库中元组数量不足和采样时存在误差。为了解决这个问题，首先定义几个符号：

定义 9　对论域 X 的数值域拓扑空间 $\langle F, \gamma\rangle$ 中的任意两个开集 d，e，它们对应的知识结点分别为 n，k，定义：

1）$P_0(d) = P_0(n)$ 为按照论域 X 中固有的概率分布某一现象属于 $\psi(d)$（即满足知识结点 n）的概率；$P_0(e \mid d) = P_0(k \mid n)$ 为按照论域 X 中固有的概率分布和属性相关关系，一个现象在属于 $\psi(d)$ 的条件下，它同时又属于 $\psi(e)$ 的条件概率；

2）$P(d) = P(n)$ 为采样时某一采样对应的现象（而不是采样本身）属于 $\psi(d)$ 的概率；$P(e \mid d) = P(k \mid n)$ 为采样时，一个现象在属于 $\psi(d)$ 的条件下，它同时又属于 $\psi(e)$ 的条件概率；

3）$P_m(d) = P_m(n)$ 为一个采样属于 d（即满足 n）的概率；$P_m(e \mid d) = P_m(k \mid n)$ 为当现象属于 $\psi(d)$（即满足 n）时，它对应的采样属于 $\psi(e)$（即满足 k）的概率。

这里，以引理的形式给出相关于定义 9 的一些结论：

引理 3　对论域 X 的数值域拓扑空间 $\langle F, \gamma\rangle$ 中的任意两个开集 d，e，假设它们对应的知识结点分别为 n，k，则：

1）如果 $P_0(d) = 0$，则 $P(d) = 0$。即测量、收集采样时，不可能获取到论域中不可能出现的现象对应的采样。于是，如果记 $I_0 = \{i \mid P_0(F_i) = 0, i = 1,2,\cdots,v\}$，记 $I_1 = \{1,2,\cdots,v\}/I_0$，则对 $i \in I_0$，有 $P(F_i) = 0$。

2）如果 n 到 k 有推理关系 $n \to k$，则 $P_0(k \mid n) = 1$，即 $P_0(e \mid d) = 1$，且反之亦然；

3）$P(e \mid d) (P_0(e \mid d)$，即 $P(k \mid n) \equiv P_0(k \mid n)$，即每一个采样对应的现象都要满足论域 X 固有的属性间的相关规律。

4）$P_m(d) = P_m(d \mid d)P(d) + P_m(d \mid \sim d)P(\sim d)$。（证明略）

定义 10　对数值域拓扑空间 $\langle F, \gamma\rangle$ 中的任意开集 d，$P_m(\sim d \mid d)$ 的意义是：属于 $\psi(d)$ 的现象通过采样误差而使其对应的样本不属于 $\psi(d)$ 的概率，称为 d 的误差概率。对基本集 F_i，设 $P_m(\sim F_i \mid F_i) = \alpha_i$，$i = 1,2,\cdots,v$。根据误差理论，$\alpha_i$ 是一个接近于 0 的正实数，记 $\alpha = \max(\alpha_1, \alpha_2, \cdots \alpha_v)$，则 α 是一个很小的正实数，称为论域 X 的基本误差概率。

引理 4　如果 d 是数值域拓扑空间 $\langle F, \gamma\rangle$ 中的任意开集，$d \neq \varnothing$，则 d 的误差概率不大于论域 X 的基本误差概率。

证明　假设 $d = \bigcup_{i \in I} E_i$，由引理条件，$d \neq \varnothing$，于是根据全概率公式，

$$P_m(d \mid d) = P_m(d \mid \bigcup_{i \in I} F_i) = \sum_{i \in I} P_m(d \mid F_i)P(F_i) \geqslant \sum_{i \in I} P_m(F_i \mid F_i)P(F_i) \geqslant$$

$(1 - \alpha)\sum_{i \in I} P(F_i) = (1 - \alpha)P(d) \geqslant (1 - \alpha)$。

故 $P_m(\sim d \mid d) \leqslant \alpha$。

定义 9 的 3 中定义的“一个采样属于开集 d 的概率 $P_m(d)$”是一个很重要的量，根据引理 3 的 4 项和引理 4，可以给出对这个量的估计：

引理 5　如果 d 和 e 是拓扑空间 $\langle F, \gamma\rangle$ 中的任意两个开集，α 是所定义的论域 X 的基本误差概率，则

1）$(1 - \alpha)P(d) \leqslant P_m(d) \leqslant (1 - \alpha)P(d) + \alpha$。特别地，当 $P(d) = 0$ 时，$P_m(d) \leqslant (1 - \alpha)$；当 $P(d) = 1$ 时，$P_m(d) \geqslant (1 - \alpha)$。

2）当 $P(d) \neq 0$ 时，$P_m(e)/P_m(d)$ 有意义，它有如下估计（证明略）：

$$\frac{(1 - \alpha)P(e)}{(1 - \alpha)P(d) + \alpha} \leqslant \frac{P_m(e)}{P_m(d)} \leqslant \frac{(1 - \alpha)P(e) + \alpha}{(1 - \alpha)P(d)}$$

引理 6　设 I_1 的意义如引理 3 的 1 所述。记 $p = \min\{P(F_i) \mid i \in I_1\}$。设 d 和 e 分别是拓扑空间 $\langle D, \gamma\rangle$ 中的任意两个开集，$e \subset d$，且至少存在 $k \in I_1$，使得 $F_k \subseteq d$，但 $F_k \not\subset e$。如果 $p > 2\alpha + \alpha^2/(1 + \alpha)$，则有：

1) $P_m(d) > \alpha$;

2) $P_m(e)/P_m(d) < 1-\alpha$。

证明　由引理条件，$P(d) \geqslant p > 2\alpha$，根据引理5的1项，$P_m(d) \geqslant (1-\alpha)P(d) \geqslant (1-\alpha)p > 2\alpha(1-\alpha) = 2\alpha - 2\alpha^2$。由于基本误差概率 α 是一个接近于0的正数，所以 $\alpha > 2\alpha^2$。故本引理的1项得证。

由 $p > 2\alpha + \alpha^2/(1+\alpha)$ 可得，$p > \alpha + \alpha/(1-\alpha)$。从而，$1 - p + \alpha/(1-\alpha) < (1-\alpha)$。由此可得

$$(1-\alpha) - (p(1-\alpha) - \alpha) < (1-\alpha)^2。$$

由于 $P(d) \leqslant 1$，$(p(1-\alpha) - \alpha) \geqslant 0$，故把上式的 $(p(1-\alpha) - \alpha)$ 代换成 $(p(1-\alpha) - \alpha)/P(d)$，则不等式仍然成立。即

$$(1-\alpha) - \frac{(1-\alpha)p - \alpha}{P(d)} < (1-\alpha)^2。$$

变形得

$$\frac{(1-\alpha)(P(d) - p) + \alpha}{P(d)} < (1-\alpha)^2。$$

根据引理条件中对基本集上的概率分布的规定，至少存在一个基本集 F_k，$k \subseteq I_1$，使得 $F_k \not\subset d$，但 $F_k \notin e$。因此，$P(e) \leqslant P(d) - p$。故用 $P(e)$ 代替上式中的 $P(d) - p$，不等式仍然成立，即

$$\frac{(1-\alpha)P(e) + \alpha}{P(d)} < (1-\alpha)^2。$$

由引理5的2项可得 $P_m(e)/P_m(d) < 1-\alpha$。

4.2 可信度

定义11　设 U 是关系数据库 $\mathscr{R}(X)$ 的元组集 R 的任意一个子集，定义 U 的规模为 $S(U) = \text{Size}(U) = |U|$，即 U 中元组的数目。如果 $S(\mathscr{R}(d)) \neq 0$，则把比值 $S(\mathscr{R}(d) \cap \mathscr{R}(e))/S(\mathscr{R}(d))$ 称为 e 对 d 的可信度，记为 $H(e/d)$；若 d 和 e 分别对应知识结点 n 和 k，则 $H(e/d)$ 也记为 $H(k/n)$，称为知识结点 k 对 n 的可信度。

对论域 X 中任意两个知识结点 n 和 k，$n \neq \varnothing$，设它们分别对应拓扑空间 $\langle F, \gamma\rangle$ 中的两个开集 d 和 e。对于论域 X 的一个固定的数据库 $\mathscr{R}(X)$，数据子类结构集 $\langle \gamma, \mathscr{R}(\gamma)\rangle$ 也是固定的，$S(R)$、$S(\mathscr{R}(d))$ 和 $H(k/n)$ 均是一个相应的固定的数（但对 $H(k/n)$ 要 $S(\mathscr{R}(d)) \neq 0$），其中 $S(R)$ 是数据库所有元组的总数；但当数据库 $\mathscr{R}(X)$ 变化时，$S(\mathscr{R}(d))$、$S(\mathscr{R}(d))/S(R)$ 和 $H(k/n)$ 均是一个随机变量。

引理7　两个随机变量 $S(\mathscr{R}(\mathbf{d}))/S(R)$ 和 $H(k/n)$ 的数学期望分别为：$P_m(d)$ 和 $P_m(e \cap d)/(d)$。（证明略）

在误差存在的情况下，为了使论域 X 的数据子类结构可达范畴 $C_\infty\langle \gamma, \mathscr{R}(\gamma)\rangle$ 仍可以用来决定推理范畴 $C_r(N)$，把定义9中的可达关系“∞”作适当修正。

4.3 可达关系的概率估计

定义12　设参数 β 和 B 都是一个正实数，$0 \leqslant \beta < B \leqslant 1$。对于论域 X 的给定的数据库 $\mathscr{R}(X)$，称数据子类结构 $\langle d, \mathscr{R}(d)\rangle$ 到 $\langle e, \mathscr{R}(e)\rangle$ 有可达关系，如果 $S(\mathscr{R}(d)) \leqslant \beta |R|$；或者如果 $S(\mathscr{R}(d)) > S(R)$，$S(\mathscr{R}(d) \cap \mathscr{R}(e)) > S(R)$，并且 $H(e/d) \geqslant B$. 仍把 $\langle d, \mathscr{R}(d)\rangle$ 到 $\langle e, \mathscr{R}(e)\rangle$ 有可达关系记为 $\langle d, \mathscr{R}(d)\rangle \infty < e, \mathscr{R}(e)\rangle$，称数据子类结构 $\langle d, \mathscr{R}(d)\rangle$ 到 $\langle e, \mathscr{R}(e)\rangle$ 有非可达关系，如果 $S(\mathscr{R}(d)) > S(R)$，$S(\mathscr{R}(d) \cap \mathscr{R}(e)) \leqslant S(R)$；或者如果 $S(\mathscr{R}(d)) > S(R)$，$S(\mathscr{R}(d) \cap \mathscr{R}(e)) > S(R)$，并且 $H(e/d) < B$。

数据子类结构库 $\langle D, \mathscr{R}(X)\rangle$ 中的每一个关系（定义12所述的可达关系或非可达关系）都对应着本原知识库 $K^p(X)$ 中一条规则，理想的对应是：每一个可达关系对应的规则均为正规则；每一个非可达关系对应的规则均为反规则。但由于误差的存在和数据量的不足，使理想的对应关系不一定成立。但是有

定理11　I_1 的意义如引理3所述。设 $p > 2\alpha + \alpha^2/(1-\alpha)$；对定义12中定义的参数 β 和 B，令 $\alpha < \beta < (1-\alpha)p$，令 $(1 - p + p\alpha)/(1-\alpha) < B < 1 - \alpha$。则随着论域 X 的数据库 $\mathscr{R}(X)$ 中元组数目 $S(R)$ 的增加，本原知识库中每一条正规则对应的数据子类结构库中的关系为一个可达关系的概率均趋于1；每一条反规则对应的关系为非可达关系的概率均趋于1。

证明　在论域 X 中任意取两个知识结点 n 和 k，设它们分别对应数值域拓扑空间 $\langle F, \gamma\rangle$ 中的两个开集 d 和 e。

首先，当 $P(n) = 0$ 时，由定理条件中对基本集上概率分布的约定可得 $P_0(n) = 0$，因而恒有 $n \to k$。但由引理7，数学期望 $E(S(\mathscr{R}(d))/S(R)) = P_m(d)$，把随机变量 $S(\mathscr{R}(d))/S(R)$ 简记为 s，把 R 的规模 $S(R)$ 简记为 R。由辛钦大数定律，对任

意 $\varepsilon>0$，

$$P(|s/R - P_m(d)|<\varepsilon)\to 1,$$

即
$$P(P_m(d)R - \varepsilon R(s(P_m(d)R + \varepsilon R)\to 1,$$

$$P(s < P_m(d)R + \varepsilon R)\to 1。$$

由引理5，$P_m(d)\leqslant\alpha$，故 $(s<\alpha R+\varepsilon R)\to 1$。而定理条件中约定 $\beta>\alpha$，又因 ε 是任意的，所以 $P(s<\beta R)\to 1$。

由定义12，正规则 $n\to k$ 对应的$\langle d, \mathcal{R}(d)\rangle$到$\langle e, \mathcal{R}(e)\rangle$ 的关系为可达关系的概率随数据库 $\mathcal{R}(X)$ 的元组数目的增加而趋于1。

当 $P(n)\neq 0$时，根据定理条件，有 $P_0(d)\neq 0$，这种情况下，当 $P(e)=0$时，必有 $P_0(e)=0$，从而 $n^{\times}\to k$，用与上面类似的方法可以证明，它对应着$\langle d,\mathcal{R}(d)\rangle$到$\langle e, \mathcal{R}(e)\rangle$ 的关系为非可达关系的概率趋于1。当 $P(e\cap d)\neq 0$时又分两种情况：n 到k 为正规则($n\to k$) 的情况和 n 到k 为反规则($n^{\times}\to k$) 的情况。因为第一种情况的证明与当 $P(n)=0$ 时的证明类似，所以这里只讨论第二种情况。

当 n 到k 为反规则时，$P_0(d)\neq 0$，且至少存在 $k\in I_1$，使得 $E_k\subseteq d$，但 $Ek\not\subset e$。与前面的结论类似，首先有

$$P(S(\mathcal{R}(d))>\beta|R|)\to 1, \qquad (19)$$

和
$$P(S(\langle e\cap d, \mathcal{R}(e\cap d)\rangle)>\beta|R|)\to 1。 \qquad (20)$$

把随机变量 $H(e/d)$ 简记为 h，则由引理7，$E(h)=P_m(e\cap d)/P_m(d)$，把 $P_m(e\cap d)/P_m(d)$ 简记为 r。由辛钦大数定律，对任意 $\varepsilon>0$，随着元组数目的增加，

$$P(|h-r|<\varepsilon)\to 1,$$

因而
$$P(h<r+\varepsilon)\to 1。$$

由引理6的2，$r<1-\alpha$. 故 $P(h<1-\alpha+\varepsilon)\to 1$，但 ε 是任意的，$B>(1-\alpha)$，所以

$$P(h<B)\to 1。 \qquad (21)$$

由式(19)、式(20)和式(21)以及定义12，反规则($n^{\times}\to k$)对应$\langle d, \mathcal{R}(d)\rangle$到$\langle e, \mathcal{R}(e)\rangle$的关系为非可达关系的概率随元组数目的增加而趋于1。

基本误差概率 α 可以通过经验给出，或利用领域的基础知识和数据库计算得出。当 α 为0时，定理11的结论显然成立。此时的数据子类结构库退化为本原数据子类结构库。可以证明：如果对任意 $i\in I_1$，$P(F_i)>0$，则随着元组数目的增加，本原数据库成为完全数据库的概率趋于1。

5　结论

阐述了双库协同机制的涵义，双库协同机制下的知识库及其结构、数据库及其结构，以及两库之间在本质上的对应关系。只有揭示出这种对应关系，才能完成定向搜索和定向发掘，为实现文(Ⅱ)中的两个协调算法及其构件奠定了理论基础，并为构造新的知识发现的结构模型 KDD* 作了铺垫。

参考文献

[1] Anand S S, Bell D A, Hughs J G. EDM: A general framework for data mining based on evidence theory [J]. Data & Knowledge Eng, 1996, 18: 189～223

[2] Piatetsky-Shapiro G, Matheus C J. Knowledge discovery work-bench for exploring business databases [J]. International Journal of Intelligent Systems, 1992, 7: 675～686

[3] Yoon J P, Kerschberg L. A frame work for knowledge discovery and evolution in databases [J]. IEEE Transactions on Knowledge and Data Eng, 1993, 5: 973～979

[4] Yang Bingru . KD(D&K) and double-bases cooperating mechanism [J]. Journal of System Engineering and Electronics, 1999, 10(2): 48～55

[5] Yang Bingru. Double-base cooperating mechanism in KDD [J], International Symposium on Computer, 1998, 149～152

[6] Yang Bingru. FIM and CASE for evaluation of hazard level on fuzzy language field [J]. Fuzzy Sets and System, North Holland, 1997, 95(2): 83～89

[7] 杨炳儒. 关于 KDD 的一类开放系统 KDD* 的研究 [J], 计算机科学, 2000, 27(2): 83～87

(下转第 57 页)

参考文献

[1] 王正中. 复杂系统仿真方法及应用 [J]，计算机仿真，2001，18(1)：3～6

[2] http://www.swarm.org

[3] http://repast.sourceforge.net

[4] http://ww.brook.edu/es/dynamics/models/ascape

[5] http://www.media.mit.edu/starlogo/

[6] http://www.cs.sandia.gov/tech-reports/rjpryor/Aspen.html

[7] http://www.santafe.edu

[8] 李宏亮，金士尧，王俊伟，等. 复杂自适应系统分布仿真平台 JCass 的研究与应用 [R]. 长沙：国防科技大学计算机学院，2001

[9] 史忠植. 智能主体及其应用 [M]. 北京：科学出版社，2000

[10] The high level architecture, defense modeling and simulation organization [DB/OL]. http://www.dmso.gov/hla.

[11] 金士尧，党　岗，凌云翔，等. 银河高性能分布仿真系统的设计与实现 [J]. 计算机研究与发展，2001，(4)：458～466

Design and Research of Computer Simulation for Complex System

Jin Shiyao, Li Hongliang, Dang gang, Wang Zhaofu, Liu Xiaojian

(*School of Computer Science, National University of Defense Technology, Changsha* 410073, *China*)

[**Abstract**] The complex systems and complexity is the kernel scientific problem of the 21st century. Due to the complexity and indetermination of complex systems, it is difficult to study the complex systems with the traditional reductive theory. The agent-based fuzzy computer simulation is approved in the paper, and a distributed simulation platform based on the agents is designed.

[**Key words**] complex systems; computer simulation; agent; HLA/RTI

(cont. from p.51)

A Study on Double Bases Cooperating Mechanism in KDD (Ⅰ)

Yang Bingru, Wang Jianxin

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

[**Abstract**] This paper and paper (Ⅱ), aiming at the problems in the mainstream development of KDD, put forward a new academic thought that, in the knowledge discovery, the database can be restricted and deriven by the knowledge base whose structure can be in turn improved by the database. An open KDD system, KDD*, with double-base cooperating mechanism can thus be created. At the same time, the velocity, precision, and the auto-cognition of knowledge discovery can be imroved and the knowledge base is structurally equipped with the capacity of real time maintenance and to evolve by itself. This paper mainly deals with the corresponding relation between the database and knowledge base, which are the basis of double-base cooperating mechanism.

[**Key words**] knowledge node; primitive knowledge base; primitive database; data sub-class structure; double-base cooperating mechanism