计算机科学 2006Vol. 33№. 5

面向 KDD 的粒度计算建模研究*)

蒙祖强[1]　蔡自兴[2]

(广西大学计算机与电子信息学院　南宁 530004)[1]　(中南大学信息科学与工程学院　长沙 410083)[2]

摘　要　给出了全粒度空间的拓扑结构模型，进一步介绍了面向粒度计算的产生式决策逻辑语言 GDL-language，然后定义了面向粒度描述的正基语言，阐明该语言公式的语义解释，给出了一种"全粒度空间＋正基语言"的粒度计算模型，并找到了正基语言系统、粒度空间和基本概念空间的关系定理。最后，把 KDD 任务归结为基于该模型的粒度计算问题，这样就可以把各种 KDD 任务统一到一个理论框架下，也便于比较和研究。这些工作无疑对今后知识发现的研究起着重要的作用。

关键词　粒度计算，KDD，决策逻辑语言，建模

Research on Granular Computing Modeling for KDD

MENG Zu-Qiang[1]　CAI Zi-Xing[2]

(College of Computer and Information Engineering, Guangxi University, Nanning 530004)[1]

(College of Information Science & Engineering, Central South University, Changsha 410083)[2]

Abstract　Topological structural model of AllGS is given, and GDL-language is further introduced for GrC(granular computing). Then L_{+base}, a language for describing granules, is defined and its semantic explanation and a kind of granular computing model are also presented. Some theorems between L_{+base}, granular space and BCS(Basic Concept Space) are found. At last, the problems of KDD come down to ones of granular computing, and then KDD tasks become problems in a theoretical framework, which leads to convenience for research on KDD. These work will be beneficial to further study on KDD.

Keywords　GrC, KDD, GDL-language, Modeling

1　引言

知识发现(KDD, Knowledge Discovery in Databases)是AI研究的热点之一，目前的研究主要是以知识发现的任务描述、知识评价与知识表示为主线，有效的知识发现算法为中心[1]。相比之下，在有关 KDD 的形式化和数学建模方面所做的工作就很少了[2]。诚然，研究不同的知识发现算法是很重要的，但独立于具体的算法研究 KDD 的形式化工作同样很重要。如果能够对 KDD 建立一个通用理论框架，使得各种方法都可以在此框架下得以研究和比较，这无疑是对 KDD 研究的重要贡献。为此，人们探讨了多种理论和方法，试图建立 KDD 的通用模型，其中以粒度计算最为成功和有效。

粒度计算(GrC, Granular Computing)是逐步形成并有待进一步完善的一种计算方法，到目前为止，还难以给它下一个精确的定义。但普遍认为，粒度计算是一把"大伞"，它是覆盖了所有有关粒度的理论、方法论、技术和工具的研究[5]，其核心内容归结为两个方面：粒度(granule)的创建和运用粒度进行计算。前者是处理粒度的形成、表示和解释，后者则涉及运用粒度解决问题。理论上，论域的任何一个子集都可以看作是一个粒度，而实际上在问题处理过程中通常是根据不可分辨性、相似性或近似性等把有关的对象(点)抽取在一起而形成的集合。粒度计算的基本原理和方法符合人类对问题解决的从简思路，并且这种方法是高效和可行的，这也是它得以迅速发展的主要原因。文[2]基于粒度计算探讨了 KDD 的建模问题，提出了一种粒度计算模型。该模型实际上是有限对象的集合，每一个对象由有限个属性-值对来描述。该模型的贡献之处在于，建立了决策逻辑语言和论域之间的联系。但是这种模型比较粗糙，一方面，没有对知识空间的结构特性进行分析，没有揭示知识的粒度和层次特征；另一方面，由于决策逻辑语言的广泛性，缺少对知识粒度的针对性描述，从而失去对知识形成过程的有效性分析，导致建立的模型缺乏通用性而无上述的理论意义。我们在文[4]对知识空间的结构进行了初步的探讨，建立了知识空间的拓扑结构模型。本文在此基础上，建立该模型与决策逻辑语言子集的对应关系，形成了"全粒度空间＋正基语言"的一种计算模型，并讨论它的若干性质。从而得到了一个可接受的模型，可为许多基本概念提供一个公用的解释，同时也是许多问题解决、讨论和比较的理论框架。

2　信息系统及全粒度空间

在粒度计算理论中，数据一般是利用信息系统(Information System)来表示。具体讲，信息系统表示为四元组：

$$IS=\langle U,A,\{V_a\},f_a\rangle_{a\in A}$$

其中，U 是论域，表示有限对象(样本)的集合，A 是有限属性的集合；V_a 是对象在属性 $a\in A$ 上所有可能取值的集合(即 a 的值域)；$f_a: U\to V_a$ 是论域 U 到值 V_a 上的映射，称为信息

*)本文得到国家自然科学基金项目(60234030, 60404021)及广西大学科研基金项目资助。蒙祖强　博士，从事机器学习与数据挖掘，多 agent 等方面研究；蔡自兴　教授，联合国专家，纽约科学院院士，博士生导师，从事人工智能，智能机器人，机器学习和智能控制等方面研究。

函数(information function)。

一个信息系统可以用一个二维关系表来表示,称为信息表,表中每一行表示一个对象,每一列表示一个属性列。当没有重复的对象时,信息表表示一个关系数据库。在不引起混淆的情况下,信息系统$\langle U,A,\{V_a\},f_a\rangle_{a\in A}$简写为$\langle U,A\rangle$,即$IS=\langle U,A\rangle$。

在信息系统$\langle U,A\rangle$中,利用信息函数f_a在U上构造一个关于属性集$B\subseteq A$的关系R_B,定义如下:

$R_B=\{(s_1,s_2)\mid f_a(s_1)=f_a(s_2)$,对所有$a\in B,s_1,s_2\in U\}$

显然,R_B是一个等价关系。为简单起见,在不与属性子集混淆或者不需特别强调的情况下,直接把R_B写成B,称为等价关系B。

易知,U上的任一等价关系都可以形成U的相应的等价划分。把等价关系B形成的等价划分记为P_B,或者记为记商集的形式U/B。

定义1 由论域U上的等价关系B确定的划分U/B称为信息系统IS关于B的粒度空间(Granular Space),记为GS_B,当B已知时,分别简记为粒度空间和GS,GS_B中的元素称为关于B的粒度(granule),通常表示为G;当$B=A$时,GS_B中的粒度称为信息系统IS的基本元,对于基本元G,令$p=|G|$,则G称为关于B的p元基本元;规定$GS_\varnothing=\{U\}$,令$AllGS=\{G\in GS_B\mid B\in P(A)\}$,$AllGS$称为$IS$的全粒度空间,其中$P(A)$表示集合$A$的幂集(下同)。

可见对$\forall B\subseteq A$,关于B的粒度空间GS_B是全粒度空间$AllGS$的一个子集。当把GS_B作为一个整体来看待时,GS_B可被认为是一种特殊的粒度,由此可以形成更高一层的粒度空间。

定义2 令$P(A)$为信息系统IS属性集A的幂集,集合$\{GS_B\mid B\in P(A)\}$称为信息系统IS的超粒度空间(Super-Granular Space),记为S_GS,GS_B称为超粒度。

3 全粒度空间的超树结构模型

定义3 对于$\forall GS_{B1},GS_{B2}\in S_GS$,如果$B_1\supseteq B_2$,称$GS_{B1}$内涵包含于$GS_{B2}$,记为$GS_{B1}\subseteq^* GS_{B2}$。

定理1 信息系统$\langle U,A\rangle$的超粒度空间S_GS及其元间的关系$\subseteq^*$构成一个格,用二元组$\langle S_GS;\subseteq^*\rangle$表示[4]。

定义4[4] 在格$\langle S_GS;\subseteq^*\rangle$中从0元$GS_A$到1元$GS_\varnothing$的每一条路径都对应着一棵树,称为完全树。从0元$GS_A$到1元$GS_\varnothing$存在多条路径,每一条路径都对应着一棵完全树,所有完全树的集合构成了"森林",但是这种"森林"很特别,其每一棵树的树根和树叶子都重合在一起,而且有些树枝也重叠在一起,所以不宜用通常所说的森林来描述这种"森林"。我们认为,它更像一棵树,只不过树枝"叉开"了。因此,用超树来命名这种特殊的"森林",记为$Stree$。严格说,超树是把一个信息系统的格中所有0元到1元的路径对应的完全树在同一坐标系中画出而得到的数据结构。直观上讲,超树是由多棵树组成,但这些树的根节点和叶子节点都相同,并且有部分(不是全部)分枝也是重合在一起。在超树中,假设G为非叶子节点,对于G下一层的所有节点中,凡是与G有树枝相连的节点都称为G的子节点,而G则称为这些节点的父节点(节点的父节点不一定是唯一的)。有时为了阐述之便,把超树看作是若干棵完全树的集合,即$Stree=\{tree_1,tree_2,\cdots,tree_m\}$,$m$为完全树的总数,$tree_i((i=1,2,\cdots,m)$表示完全树。

在超树中,节点的层次从叶节点开始定义起,叶节点所在的层为第一层,若某节点在第h层,则其父节点在第$h+1$层(如果父节点存在的话)。这样,根节点就在最高层。

定理2 令S为超树$Stree$所有节点的集合(包括根节点、叶子节点和内节点),则$AllGS=S$,即$Stree$所有节点的集合即为全粒度空间。

该定理的证明是显然的,因为$AllGS=\{G\in GS_B\mid BP(A)\}$,表明全粒度空间$AllGS$是由所有超粒度所包含的粒度的集合;而超树的构造过程是从超粒度$GS_\varnothing(=GS_{Bn})$开始,然后是超粒度$GS_{Bn-1},\cdots$,最后到超粒度GS_{B1},逐个对超粒度所包含的粒度展开并对其进行"连线"(构建树枝)。可见,超树所包含的节点也是由所有超粒度所包含的粒度的集合,所以$AllGS=S$。

超树是全粒度空间的一种结构模型,对超树的研究有助于对全粒度空间性质的探讨,可以有效地指导知识的发现过程和KDD算法的设计。

例 对于表1所示的信息表,在其对应的格中,考虑序列$Seq1:GS_{\{a,b,c\}}\subseteq^* GS_{\{a,b\}}\subseteq^* GS_{\{a\}}\subseteq^* GS_\varnothing$,其对应的完全树如图1所示。

表1 信息表

U	a	b	c
s_1	1	3	2
s_2	2	1	1
s_3	1	3	3
s_4	2	3	2
s_5	2	1	1

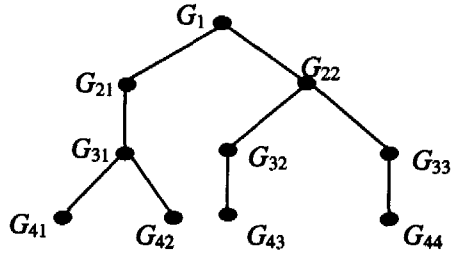

图1 $Seq1$的完全树

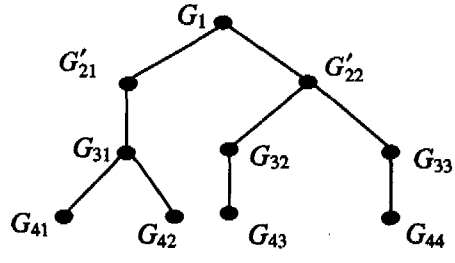

图2 $Seq2$的完全树

其中,粒度$G_1=U$,$G_{21}=\{s_1,s_3\}$,$G_{22}=\{s_2,s_4,s_5\}$,$G_{31}=G_{21}=\{s_1,s_3\}$,$G_{32}=\{s_2,s_5\}$,$G_{33}=\{s_4\}$,$G_{41}=\{s_1\}$,$G_{42}=\{s_3\}$,$G_{43}=G_{32}=\{s_2,s_5\}$,$G_{44}=G_{33}=\{s_4\}$。可以看到$G_{42}\subseteq G_{31}\subseteq G_{21}\subseteq G_1$等,子节点的并等于父节点,如$G_{41}\cup G_{42}=G_{31}$,同一层粒度构成$U$的一个划分等。注意到,由有序序列$Seq2:GS_{\{a,b,c\}}\subseteq^* GS_{\{a,b\}}\subseteq^* GS_{\{b\}}\subseteq^* GS_\varnothing$亦可以构造一棵完全树,如图2。其中只有$G'_{21}(=\{s_1,s_3,s_4\})$和$G'_{22}(=\{s_2,s_5\})$分别与对应的$G_{21}$和$G_{21}$不同以外,其它的对应节点都相同。

如果把它们在同一坐标中画出，可得到如图 3 所示的树（超树的一部分）。

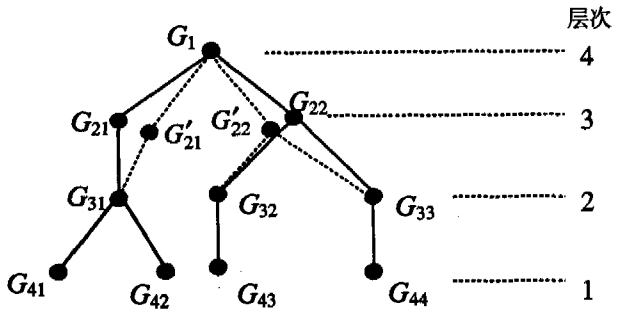

图 3　*Seq*1 和 *Seq*2 构成超树的一部分

从图 3 可看出，由根节点出发从两种不同的途径也可以达到叶子节点。如果把每一条从 0 元 GS_A 到 1 元 $GS_\varnothing$ 的各条路径对应的完全树都在同一空间中画出，将得到一棵包含所有粒度的超树。

4　KDD 的粒度计算模型及其性质

4.1　粒度计算模型

信息系统是通过属性-值对来刻画对象的，也就是说每一个对象的存在是由其属性值之间的差异而得以体现的。所以，通过利用属性-值来建立面向粒度计算的知识描述的语言系统是很自然的事情。

定义 5　对于信息系统 $IS=\langle U,A,\{V_a\},f_a\rangle_{a\in A}$，语言 L 定义如下[3]：

1）(a,v) 称为原子公式，其中 $a\in A,v\in V_a$；

2）原子公式是 L 的一条公式；

3）如果 ϕ 是 L 的公式，则 $\sim\phi$ 也是 L 的一条公式；

4）若 ϕ 和 φ 都是 L 的公式，则 $\phi\vee\varphi,\phi\wedge\varphi,\phi\rightarrow\varphi,\phi\equiv\varphi$ 都是 L 的公式；

5）只有有限次运用上述规则 1）～4）求得公式才是 L 的公式。

定义 6[7]　对 $\forall s\in U,s$ 与 L 中公式的关系定义如下：

1）$s\models(a,v)$ iff $f_a(s)=v$；

2）$s\models\sim\phi$ iff *not* $s\models\phi$；

3）$s\models\phi\wedge\varphi$ iff $s\models\phi$ and $s\models\varphi$

4）$s\models\phi\vee\varphi$ iff $s\models\phi$ or $s\models\varphi$

5）$s\models\phi\rightarrow\varphi$ iff $s\models\sim\phi\vee\varphi$

6）$s\models\phi\equiv\varphi$ iff $s\models\phi\rightarrow\varphi$ and $s\models\phi\rightarrow\varphi$

实际上，$\rightarrow,\equiv$ 可以由 $\wedge,\vee$ 和 $\sim$ 来表示，所以语言系统 L 可表示为代数系统 $\langle L,\wedge,\vee,\sim\rangle$。对于 L 中的任一公式 ϕ，其语义解释为 $\{s\in U|s\models\phi\}$，记为 $m(\phi)$，即 $m(\phi)$ 表示公式 ϕ 的在信息系统 IS 中的含义（meaning）。反过来讲，如果 $X=m(\phi)$，则用户通过 ϕ 就可以了解 X 的基本特征，从而建立了关于 X 的概念。可见，ϕ 是 X 的基于决策逻辑语言的公式描述。

定义 7　对于子集 $X\in P(U)$，如果存在 $\phi\in L$，使得 $m(\phi)=X$，则称 X 在信息系统 IS 中是可描述的，ϕ 称为 X 的一个描述公式，简称为 X 的描述，有时把 ϕ 写成 $DES(X)$。

定理 3　在 L 中，存在下列性质[2,3]，对 $\forall\phi,\varphi\in L$：

$(1)m(a,v)=\{s\in U|f_a(s)=v\}$

$(2)m(\sim\phi)=-m(\phi)$

$(3)m(\phi\wedge\varphi)=m(\phi)\cap m(\varphi)$

$(4)m(\phi\vee\varphi)=m(\phi)\cup m(\varphi)$

$(5)m(\phi\rightarrow\varphi)=-m(\phi)\cup m(\varphi)$

$(6)m(\phi\equiv\varphi)=(m(\phi)\cap m(\varphi))\cup(-m(\phi)\cap-m(\varphi))\cup\cap$

定义 5 定义的语言 L，因过于“广泛”而缺乏针对性。特别是对于全粒度空间 AllGS，可以说是“绰绰有余”，这样反而不利于问题的研究。为此，进一步引入相关的概念。

定义 8　令 $L_{base}=\{\varphi\wedge|\varphi(a,v)$ 或 $\sim(a,v),v\in V_a,a\in A\}$，即 L_{base} 表示语言系统 L 中所有由原子公式或原子公式的非的合取而组成的公式的集合，则称 L_{base} 为 L 中关于 A 的基语言（在不引起混淆的情况下，简称基语言，下同），对 $\phi\in L_{base},\phi$ 称为 L 的基公式；如果 L_{base} 中的任意基公式都不含原子公式的非，则称 L_{base} 为 L 的正基语言，记为 L_{+base}，相应的基公式称为正基公式。对任意 $\phi\in L_{base},\phi$ 为一合取式，其包含的项（原子公式或原子公式的非）的个数称为 ϕ 的长度，记为 Length(ϕ)。

定义 9　对于子集 $X\in P(U)$，如果存在 $\phi\in L_{+base}$，使得 $X=m(\phi)$，则称 X 在信息系统 IS 中是可正基描述的，ϕ 称为 X 的正基描述，有时把 ϕ 写成 $DES_+(X)$。

对任意 $X\in P(U)$，X 都是可描述的，但并不一定是可正基描述的。我们有下列重要的定理：

定理 4　在信息系统 $\langle U,A\rangle$ 中，对于全粒度空间 $AllGS$，$\forall G\in AllGS,G$ 是可正基描述的。

证明：因为 $G\in AllGS$，所以可假设 $\exists B\in P(A)$，使得 $G\in GS_B$，并假设 $B=\{b_1,b_2,\cdots,b_m\}$。由于 GS_B 是 U 的一个等价划分，因此 G 是一个等价类，故 G 是由 B 中每一属性的取值来确定的。假设这些取值分布情况为：$b_1=v_{b1},b_2=v_{b2},\cdots,b_m=v_{bm}$，则 G 的描述为 $(b_1,v_{b1})\wedge(b_2,v_{b2})\wedge\cdots\wedge(b_m,v_{bm})\in L$，且该公式中没有包含任何原子公式的非，所以它是 G 的正基描述。证毕。

由定理 3 和定理 4，超树中的任一节点 G 都是可正基描述的，即 $DES_+(G)$ 存在。

定理 5　在超树 *Stree* 中，假设对任意一条从树根节点到叶子节点的路径上的节点依次是 $G_1,G_2,\cdots,G_h$，则 $|G_1|\geqslant|G_2|\geqslant\cdots\geqslant|G_h|$，且 Length($DES_+(G1)$)$\leqslant$Length($DES_+(G_2)$)$\leqslant\cdots\leqslant$Length($DES_+(G_h)$)，即路径沿着树叶到树根方向上的节点，粒度越来越大，而描述长度越来越短（描述越来越简单）。

证明：由于 $G_1\supseteq G_2\supseteq\cdots\supseteq G_h$，因此 $|G_1|\geqslant|G_2|\geqslant\cdots\geqslant|G_h|$。

对 $G_1,G_2,\cdots,G_h$ 中的任意两个粒度 G_i 和 G_j，其中 $i,j=1,2,\cdots,h$ 且 $i<j$，则 $G_j\subseteq G_i$。分两种情况来讨论：

（1）G_i 为 G_j 的父节点。这时 G_i 和 G_j 必同时为某一棵完全树的节点。完全树的构造是在格的基础上进行的，任何一棵完全树都是由格的一条边在“展开”后得到的。假设包含 G_i 和 G_j 的完全树由 $GS_{B1},GS_{B2},\cdots,GS_{Bn-1},GS_{Bn}$ 所在的边“展开”后得到的（GS_{B1} 为 0 元，GS_{Bn} 为 1 元），则由格的性质有 $GS_{B1}\subseteq^*GS_{B2}\subseteq^*\cdots\subseteq^*GS_{Bn-1}\subseteq^*GS_{Bn}$，进而根据定义 3 得知 $B_1\supseteq B_2\supseteq\cdots\supseteq B_n$。假设 $G_i\in GS_{Bi'}G_j\in GS_{Bj'}$，因为 $G_j\subseteq G_i,GS_{Bj'}\subseteq^*GS_{Bi'}$（如果 $GS_{Bi'}\subseteq^*GS_{Bj'}$，则 $GS_{Bi'}\subseteq^+GS_{Bj'}$[4]，从而由定义 3 知 $G_j\subseteq G_i$ 不成立，故 $B_{j'}\supseteq B_{i'}$。同时可看出 $DES_+(G_i)$ 和 $DES_+(G_j)$ 是由 $B_{i'}$ 和 $B_{j'}$ 中的属性构成的，由于 $|B_{j'}|\geqslant|B_{i'}|$，因此由描述长度的定义可知 Length($DES_+(G_j)$)$\geqslant$Length($DES_+(G_i)$)。

(2)G_i 为 G_j 的祖先节点。由于 G_i 和 G_j 在同一条路径上，故必存在 $G_{t1},G_{t2},\cdots,G_{tm}\in\{G_1,G_2,\cdots,G_h\}$，使得 G_{tk} 为 G_{tk+1} 的父节点($k=1,2,\cdots,m-1$)，其中 $G_{t1}=G_i,G_{tm}=G_j$。这样，由(1)的证明可知，$\text{Length}(DES_+(G_{tm}))\geqslant\cdots\geqslant\text{Length}(DES_+(G_{t2}))\geqslant\text{Length}(DES_+(G_{t1}))$，从而 $\text{Length}(DES_+(G_j))=\text{Length}(DES_+(G_{tm}))\geqslant\text{Length}(DES_+(G_{t1})=\text{Length}(DES_+(G_i))$。而 G_i 不可能为 G_j 的子孙节点，所以 G_i 和 G_j 仅有上面两种关系，而且均有 $\text{Length}(DES_+(G_j))\geqslant\text{Length}(DES_+(G_i))$，从而进一步推导出 $\text{Length}(DES_+(G_1))\leqslant\text{Length}(DES_+(G_2))\leqslant\cdots\leqslant\text{Length}(DES_+(G_h))$成立。证毕。

该定理说明了，超树中在同一路径上的节点(粒度)，如果层次越高，则其包含的对象就越多(粒度越大)，而描述长度就越短(知识越简洁)；如果层次越低，则其包含的对象就越少(粒度越小)，而描述长度就越长(知识越复杂)。这两个定理可为寻求面向各种目的(如可理解性、准确性、兴趣性等)的最佳粒度组合提供了有效的启发信息。

定理 6 对 $\forall X\in P(U)$，X 是可正基描述的，当且仅当 $X\in AllGS$，即 X 为粒度。

证明：(“⇒”)：如果 X 是可正基描述的，则存在 $\phi\in L_{+\text{base}}$，使得 $X=m(\phi)$。由于 ϕ 是正基公式，故 ϕ 不含由原子公式的非，即仅由形如(a,v)的原子公式的合取而构成。不妨假设 $\phi=(a_1,v_{a1})\wedge(a_2,v_{a2})\wedge\cdots\wedge(a_m,v_{am})$，可以看出 X 是商集 $U/\{a_1,a_2,\cdots,a_m\}$ 中的一个等价类。当令 $B=\{a_1,a_2,\cdots,a_m\}$时，由定义 2.6，$X\in GS_B$，所以 X 是一个粒度。

(“⇐”)：其证明见由定理 4。证毕。

全粒度空间和正基语言是密不可分的，是一个“有机的结合体”。这样就形成了“全粒度空间＋正基语言”的粒度计算模型，该模型可以表示为以下的二元组：

$\langle AllGS,L_{+\text{base}}\rangle$

其中，$AllGS$ 是全粒度空间，$L_{+\text{base}}$ 是正基语言。易知，$AllGS$ 是粒度的集合表示，超树是粒度的结构化描述，因此可以进一步把该模型表示为：

$\langle Stree,U,L_{+\text{base}}\rangle$

其中，$Stree$ 是定义 4 所定义的超树，U 是论域(超树的根节点)，$L_{+\text{base}}$同上。从下文的进一步论证可以看到，知识发现问题可以归结为基于该模型的计算问题。

4.2 模型的性质分析及 KDD 任务的模型解释

粒度和描述实际上是密不可分的。没有粒度的描述，那是空的；没有描述的粒度，那是无意义的。两者的有机结合才能够形成可认知的对象。这种结合就形成了所谓的“概念”。一般地，一个概念用二元组(ϕ,X)来表示，其中 $\phi=DES(X)$。ϕ 和 X 分别称为概念(ϕ,X)的内涵(intension)与外延(extension)。

当然，在一个信息系统中所有的子集都是可描述的，但并不是所有的子集 X 都是可正基描述的。由定理 4 可知，对任意 $G\in AllGS$，X 是可正基描述的，当且仅当 $X\in AllGS$。

定义 10 在信息系统 IS 中，令 $CS=\{(\phi,X)\mid X\in P(U),X$ 是可描述的且 $X=m(\phi),\phi\in L\}$，则称 CS 为 IS 的概念空间(Concept Space)；令 $BCS=\{(\phi,X)\mid X\in P(U),X$ 是可正基描述的且 $X=m(\phi),\phi\in L_{+\text{base}}\}$，则称 BCS 为 IS 的基本概念空间(Basic Concept Space)，BCS 中的概念称为基本概念。

定义 11 对 $\forall(\phi_1,X_1),(\phi_2,X_2)\in CS,(\phi_1,X_1)=(\phi_2,X_2)$当且仅当 $\phi_1=\phi_2$ 且 $X_1=X_2$。

这就是说，两个概念相等当且仅当它们的内涵与外延分别相等。

知识发现的过程实际上就是逐步形成概念的过程。但对 U 的任一子集或多个子集，要建立它们有效的描述并形成相应的概念及其它们之间关联的描述，这是 KDD 问题的关键，是 KDD 算法设计的核心任务之一。由上可知，在信息系统中，容易建立粒度的正基描述，从而形成相应的基本概念。实际上，任一概念都可以归结为若干个基本概念的“并”，这对探讨信息系统中知识的形成机理有着重要的作用，以下将看到其意义所在，这也是引入基本概念空间的原因之一。

定理 7 正基语言系统 $L_{+\text{base}}$ 到基本概念空间 BCS 的一个对应关系 θ 定义如下：

$\theta:\phi\to(\phi,X)$

其中，$\phi\in L_{+\text{base}},X=m(\phi)$。那么，关系是 $L_{+\text{base}}$ 到基本概念空间 BCS 的一一映射。

证明：因为 $X=m(\phi)$，X 由 ϕ 唯一确定，所以(ϕ,X)是唯一的，θ 为映射；对 $\forall(\phi,X)\in BCS$，由基本概念的定义可知，其原像为 $\phi=DES_+(X)\in L_{+\text{base}}$，因此 θ 是满的；而且，对 $\forall\phi_1,\phi_2\in L_{+\text{base}}$，如果 $\phi_1\neq\phi_2$，由定义 11 可知，$(\phi_1,X_1)\neq(\phi_2,X_2)$，其中 $X_1=m(\phi_1),X_2=m(\phi_2)$，所以 θ 为单射。

综上可见，θ 是 $L_{+\text{base}}$ 到概念空间 BCS 的一一映射。证毕。

定理 8 令 ρ 为基本概念空间 BCS 到全粒度空间 $AllGS$ 的一个对应关系，定义如下：

$\rho:(DES_+(G),G)\to G$

其中，$(DES_+(G),G)\in BCS,G=m(\phi)$。那么，$\rho$ 是概念空间 BCS 到全粒度空间 $AllGS$ 的满射。

证明：显然，$(DES_+(G),G)$在 $AllGS$ 中的像 G 是唯一的，ρ 为映射；对 $\forall G\in AllGS$，由定理 6，G 是可正基描述的，故 $DES_+(G)$有意义，并令 $\phi=DES_+(G)$，则 $\phi\in L_{+\text{base}}$，所以 $(\phi,G)\in BCS$，即(ϕ,G)就是 G 在 BCS 中的原像，这说明对 $\forall G\in AllGS,G$ 都存在原像。因此，ρ 是概念空间 BCS 到全粒度空间 $AllGS$ 的满射。证毕。

定理 9 令 $\sigma=\rho\circ\theta$，即 σ 为 θ 和 ρ 的复合函数，则 σ 是正基语言系统 $L_{+\text{base}}$ 到全粒度空间 $AllGS$ 一个满射。

由定理 7 和定理 8 可知，该定理是成立的。由定理 9 知道，对于同一个粒度 G，其正基描述有多个，这是导致个性化知识发现出现的根本原因[6]。

令 $\xi(\phi)=\{a\mid(a,v)$为 ϕ 中的原子公式$\}$，$\phi\in L_{+\text{base}}$，即 $\xi(\phi)$表示正基公式 ϕ 所包含的原子公式的集合。这样，信息系统$\langle U,A\rangle$中知识发现的关联模型就可以表述为：

对于 $\forall G_1,G_2\in AllGS$，令 $\phi=DES(G_1)$，$\varphi=DES(G_2)$，显然 $\phi,\varphi\in L_{+\text{base}}$，如果 $G_1\subseteq G_2$ 且 $\xi(\phi)\cap\xi(\varphi)=\varnothing$，则 $\phi\to\varphi$ 为一条关联规则，其支持度为 $\dfrac{|m(\phi\wedge\varphi)|}{|m(\phi)|}$。

分类模型可以表述为：

设 $\varphi_1,\varphi_2 2,\cdots,\varphi_m\in L_{+\text{base}}$，$\xi(\varphi_1)=\xi(\varphi_2)=\cdots=\xi(\varphi_m)$且 $m(\varphi_1)\cup m(\varphi_2)\cup\cdots\cup m(\varphi_m)=U$，显然 $m(\varphi_i)\in AllGS$，为粒度($i=1,2,\cdots,m$)，对于任意 φ_i，如果存在 $G\in AllGS$，使得 $G\subseteq m(\varphi_i)$，则 $\phi\to\varphi_i$ 为一条分类规则，其中 $\phi=DES_+(G)$。

类似地，聚类模型等也可以从$\langle AllGS,L_{+\text{base}}\rangle$模型中得到表述和解释。

由此可见，知识发现问题可以归结为$\langle AllGS,L_{+\text{base}}\rangle$($\langle Stree,U,L_{+\text{base}}\rangle$)中的粒度计算问题，从而可进行有效的比

(下转第 194 页)

用 IRIS 数据集作为实验数据。IRIS 数据集是已建立的用于演示分类算法性能的数据集。IRIS 数据集含有 150 个样本，包括三类对象的 4 个特征。这三类对象分别是 Setosa，Versilcolor 和 Viginica，其样本点在图 1 中分别用“+”，“o”和“*”表示；4 个特征分别是 Sepal length，Sepal width，Petal length 和 Petal width（单位是 cm）。可以根据这 4 个特征将 IRIS 数据集分类。出于可视化的考虑，取 IRIS 数据集的两个特征 Petal length 和 Petal width 进行分类。由于 ACM 是二类分类器，为便于观察分类效果，分类时将 Versilcolor 作为一类，将 Setosa 和 Viginica 作为另一类。略去了 IRIS 数据集的两个特征后，IRIS 数据集的 150 样本点出现了部分重复，分类前去掉这些重复的样本。经实验确定，核函数取指数核函数 $K(x,y)=\exp\left(-\frac{\|x-y\|}{2\sigma^2}\right)$，其中 $\sigma=1.3$。图 1 给出了 IRIS 数据集的数据分布和用 ACM 分类的结果，其中虚线是分类决策线。

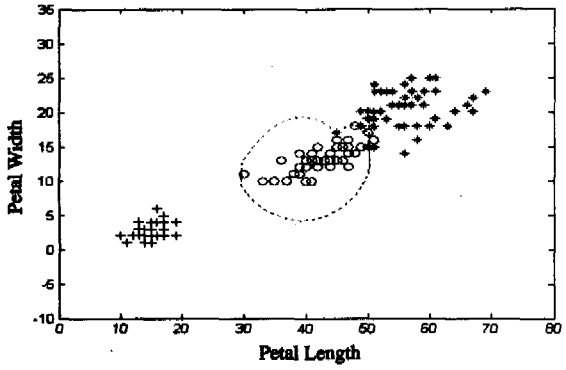

图 1　ACM 对 IRIS 数据的分类结果

我们还分别用二阶多项式核函数、高斯核函数和双曲正切核函数进行分类，结果表明部分核函数的确可以使 ACM 发散。在使 ACM 收敛的核函数中，分类效果也不一样。限于篇幅，未将这些结果一一列出。

实验中显示，ACM 发散有两种情形：一是使 ACM 算法第 21 行中的矩阵 Z 成为奇异阵，从而使牛顿迭代不收敛，而我们在证明 ACM 收敛时假定牛顿迭代是收敛的；二是经过 ACM 算法中从第 19 行到第 35 行的 while 循环，$k_j^T\alpha_i$ 的值没有变化，这不满足上述定理的第二个条件。实验中还显示，ACM 收敛性与样本点集的分布、核函数及其参数的选择有关，对于给定的样本点集，如何选择合适的核函数及其参数，使得 ACM 收敛并有理想的分类效果，还有待于进一步研究。

结束语　T. B. Trafalis 等人提出了解析中心机 ACM 的方法和相应的求解算法。实验结果表明，ACM 机的性能要优于支持向量机。而 S. S. Keerthi 等人的泛化的 SMO 算法及其收敛性结论不能用于 ACM 模型。本文对 ACM 算法的收敛性进行了研究，证明了在一定的条件下 ACM 算法是收敛的，实例分析也支持了这一结论。此外，除了解析中心机外，是否还存在其它形式的中心机模型及其有效算法？另一个值得研究的问题是，当样本非常大时如何发展 ACM 优化问题的有意义的近似算法。作者目前正在研究这方面的问题。

参 考 文 献

1　Trafalis T B, Malyscheff A M. An Analytic Center Machine. Machine Learning [J], 2002, 46: 203～223

2　曾凡仔，岳建海，裘正定. DRC-ACM：一种精确的基于解析中心的分类器. 计算机研究与发展[J]，2004，41(5)：802～806

3　Keerthi S S. Convergence of a Generalized SMO Algorithm for SVM Classifier Design. Machine Learning [J], 2002, 46: 351～360

4　James Tin-Yau Kwok. The Evidence Framework Applied to Support Vector Machines. IEEE Transactions on Neural Networks [J], 2000, 11(5)

5　James Tin-Yau Kwok. Moderating the Outputs of Support Vector Machine Classifiers. IEEE Transactions on Neural Networks [J], 1999, 10(5)

6　边肇祺，张学工，等. 模式识别（第二版）[M]. 北京：清华大学出版社，2002

7　关治，陆金甫. 数值分析基础[M]. 北京：高等教育出版社，2002

8　钱颂迪，等. 运筹学（修订版）[M]. 北京：清华大学出版社，1998

（上接第 181 页）

较和分析，使得格、超树等的有关性质在该模型中得到完美的结合，并且由超树 Stree 的几何特性和逻辑语言的代数性，$\langle Stree, U, L_{+base}\rangle$ 能把几何的直观性和代数的推理性较好结合在一起，为 KDD 的研究提供了一个统一而通用的理论框架。

结束语　不管是对 KDD 算法还是理论研究，探讨 KDD 的建模问题无疑都是很重要的。目前，从已有的文献看，在这方面的研究所取得的成果还是非常有限的。本文以粒度计算理论为工具，研究了这个问题，给出了一种“全粒度空间＋正基语言”的粒度计算模型，并找出该模型中的一些关系定理以及导出了它的若干性质，最后把 KDD 任务归结为基于该模型的粒度计算问题。今后，我们将进一步对其完善和补充，并用以指导算法设计，真正推向实用化。

参 考 文 献

1　杨炳儒，江亚东，申江涛. 基于双库协同机制的 KDD＊及其软件实现. 系统工程与电子技术，2000，22(6)：69～72

2　Yao Y Y. On Modeling Data Mining with Granular Computing. In: 25th Annual International Computer Software and Applications Conf. (COMPSAC'01), October 08 - 12, Chicago, Illinois, 2001. 638～643

3　Pawlak Z. Rough Sets—Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, 1991

4　蒙祖强. 基于分类模型的知识发现机理和方法研究：[博士学位论文]. 长沙：中南大学信息科学与工程学院，2004

5　张铃，张钹. 模糊商空间理论（模糊粒度计算方法）. 软件学报，2003，14(4)：770～776

6　蒙祖强，蔡自兴. 一种面向个性化知识发现的属性约简算法. 小型微型计算机系统，2005，26(2)：209～213

7　Yao Y Y, Liau C J. A generalized decision logic language for granular computing. In: FUZZ-IEEE'02 in The 2002 IEEE World Congress on Computational Intelligence, 2002. 1092～1097