

一个不确定性数据库模型及其语义^{*})

An Uncertain Database Model and Its Semantics

蒋运承 张师超

(广西师范大学数学与计算机系 桂林 541004)

Abstract In this paper, based on probabilistic and cognitive uncertainties, we set up an uncertain database model and presented the uncertain relational algebra, and proved that the uncertain normal forms are closed under operations of the uncertain relational algebra. At last, we studied the semantics of the uncertain relation.

Keywords Uncertain database, Quantitative fuzzy data, Semantics of uncertain relation

1 引言

一般地,不确定性可粗分为概率的和认知的不确定性,概率的不确定性体现在事物的随机性,而认知的不确定性体现在事物的模糊性。本文综合考虑了这两方面的不确定性,建立了一个不确定性数据库模型:用认知的不确定性(即模糊性)来描述属性,用概率的不确定性(即随机性)来描述元组。从而,我们所建立的不确定性数据库模型能有效地体现现实世界。另外,我们提出了“ α -相等”的概念,用它构造了一个不确定性关系代数,并证明了不确定性范式在该关系代数操作下是封闭的。最后,本文对该不确定性数据库模型的语义进行了研究。

2 不确定性数据库模型 URM 与概念

我们所建立的不确定性关系模型是 Codd 所提出的关系模型的扩展:属性值用模糊数表示,对于每一个元组 t , 增加一个属于 $p(t)$, 称之为概率因子,用来描述元组 t 在该关系中发生的概率。下面详细介绍该不确定性数据库模型。

2.1 属性的有关概念

由于属性值是用模糊数据表示的,而模糊数据有很多种表示方法,如隶属函数表示、模糊区间数表示、模糊中心数表示等等^[1],本文采用隶属函数表示。隶属函数是由 Zadeh 提出的一种表示模糊数据的方法,它把模糊数据 X 定义为论域 D 上的一个隶

属函数:

$$f: D \rightarrow [0, 1]$$

它把 D 上的每个元素映射到 $[0, 1]$ 区间上的一个值,表示该元素属于 X 的程度,值越大表示隶属的程度越高。但是,模糊数据又可以分为定量模糊数据和定性模糊数据,为了在数据库中有效地计算和度量它们,我们有必要分别研究它们的具体表示方法和度量方法。

2.1.1 定量模糊数据的有关概念。定量模糊数据是基于客观数字的模糊数据,如年龄在“30 岁左右”,身高为“大约 1.73 米”等等。对于这些定量模糊数据,我们用梯形函数来表示它们的隶属函数,即:论域 D 上的隶属函数定义如下:

$$T(x; a, b, c, d) = \begin{cases} 0 & \text{当 } x > d \text{ 或 } x < a \text{ 时} \\ (x-a)/(b-a) & \text{当 } a \leq x < b \text{ 时} \\ 1 & \text{当 } b \leq x \leq c \text{ 时} \\ (d-x)/(d-c) & \text{当 } c < x \leq d \text{ 时} \end{cases}$$

其中参数 a, b, c, d 由专家给定,且 $l_0 \leq x, a, b, c, d \leq u_p$, l_0 和 u_p 分别是论域 D 的下界和上界。

显然,上述隶属函数 $T(x; a, b, c, d)$ 在论域 D 上是连续的,这不适合数据库的应用。我们有必要对 $T(x; a, b, c, d)$ 离散化,在 $T(x; a, b, c, d)$ 上均匀地取有限个点组成一个子集,这个子集就可以用来描述定量模糊数据的隶属度,且满足数据库的要求。

下面给出判断两定量模糊数据近似相等的方法。

^{*}) 本文由国家 863 计划项目和广西教委基金项目资助

定义1 设 U 和 U' 是论域 D 上的两定量模糊数据,其隶属函数分别为: $T(x;a,b,c,d)$ 和 $T(x;a',b',c',d')$,则 U 和 U' 的近似程度定义为:

$$P_x(U,U') = \nabla A / (\Delta A + \nabla A)$$

其中 $\Delta A = \int_{\min(a,a')}^{\max(d,d')} |T(x;a',b',c',d') - T(x;a,b,c,d)| dx$

$$\nabla A = \int_a^d T(x;a,b,c,d) dx + \int_{a'}^d T(x;a',b',c',d') dx - \Delta A$$

定义2 设 U 和 U' 是论域 D 上的两定量模糊数据,如果它们的近似程度 $P_x(U,U') > \alpha$,则称 $U \alpha$ -等于 U' ,记为: $U =_\alpha U'$,其中 α 是专家给定的充分靠近1且小于或等于1的正数。

定义3 设 U 和 U' 是论域 D 上的两定量模糊数据,其隶属函数分别为 $T(x;a,b,c,d)$ 和 $T(x;a',b',c',d')$,如果 $U =_\alpha U'$,则:(1) U 合并 U' (即 $U \oplus U'$) 的隶属函数定义为: $T(x; \min(a,a'), \min(b,b'), \max(c,c'), \max(d,d'))$; (2) U 合取 U' (即 $U \odot U'$) 的隶属函数定义为: $T(x; \max(a,a'), \max(b,b'), \min(c,c'), \min(d,d'))$ 。

2.1.2 定性模糊数据的有关概念。定性模糊数据是基于主观的非数字的模糊数,如:健康状况“比较好”,学习动机“较一般”等等。由于这些模糊数据是基于主观的,它们的论域由人们指定且一般都已离散化,如:健康状况的论域可指定为{差,一般,良好,健康}。因此,这些定性模糊数据可直接用它们的可能性分布来表示,如上述的“比较好”可表示为:

$$\{0/\text{差}, 0.8/\text{一般}, 0.9/\text{良好}, 0.6/\text{健康}\}$$

下面给出判断两定性模糊数据近似相等的方法。

定义4 设论域 $D = \{d_1, d_2, \dots, d_n\}$, S_m 是定义在 D 上的一个相似关系, $S_m: D \times D \rightarrow [0,1]$ 。 X 和 Y 是 D 上的两定性模糊数据, Π_x 和 Π_y 分别是 X 和 Y 的可能性分布, π_x 和 π_y 分别是 X 和 Y 的可能性分布函数,则 X 包含 Y 的程度 $\eta(X,Y)$ 定义为:

$$\eta(X,Y) = \left(\sum_{i=1}^n \sup_{1 \leq j \leq n} (\min(\pi_x(d_i), \pi_y(d_j)) \times S_m(d_i, d_j)) \right) / N$$

其中 $N = \sum_{i=1}^n \min(\pi_x(d_i), \pi_y(d_i))$, 且 K 满足:

$$\min(\pi_x(d_i), \pi_y(d_i)) = (\sup_{1 \leq j \leq n} (\min(\pi_x(d_i), \pi_y(d_j)) \times S_m(d_i, d_j))) / S_m(d_i, d_i)$$

定义5 设论域 $D = \{d_1, d_2, \dots, d_n\}$, X 和 Y 是 D 上的两定性模糊数据,则 X 和 Y 的近似程度 P_x

(X,Y) 定义为:

$$P_x(X,Y) = \min(\eta(X,Y), \eta(Y,X))$$

定义6 设 X 和 Y 是论域 D 上的两定性模糊数据,如果它们的近似程度 $P_x(X,Y) > \alpha$,则称 $X \alpha$ -等于 Y ,记为: $X =_\alpha Y$ 。其中 α 是由专家给定的充分靠近1且小于或等于1的正数。

定义7 设论域 $D = \{d_1, d_2, \dots, d_n\}$, X 和 Y 是 D 上的两定性模糊数据, Π_x 和 Π_y 分别是 X 和 Y 的可能性分布, π_x 和 π_y 分别是 X 和 Y 的可能性分布函数,如果 $X =_\alpha Y$,则:(1) X 合并 Y (即 $X \oplus Y$) 定义为: $\{\max(\pi_x(d_i), \pi_y(d_i)) / d_i : d_i \in D\}$; (2) X 合取 Y (即 $X \odot Y$) 定义为: $\{\min(\pi_x(d_i), \pi_y(d_i)) / d_i : d_i \in D\}$,其中 $i=1,2,\dots,n$ 。

2.2 元组的有关概念

定义8 设 r 是不确定性关系模式 $R = \{A_1, A_2, \dots, A_n, P\}$ 上的一个不确定性关系, s_i 和 s_j 是 r 上的两元组,如果 $\inf_{1 \leq k \leq n} P_k(s_i[A_k], s_j[A_k]) > \alpha$ 且 $|s_i[P] - s_j[P]| \leq 1 - \alpha$,则称 $s_i \alpha$ -等于 s_j ,记为: $s_i =_\alpha s_j$,其中 α 是由专家给定的充分靠近1且小于或等于1的正数。

定义9 设 s_i 和 s_j 是不确定性关系 r 上的两个元组,如果 $s_i =_\alpha s_j$,则:(1) s_i 合并 s_j (即 $s_i \oplus s_j$) 定义为:对各个属性(除概率因子 P 外)对应地按定义3(1)或定义7(1)分别进行合并,至于概率因子 P 有不同取法,如: $\max\{P_i, P_j\}$ 或 $P_i + P_j - P_i * P_j$,选用什么要根据具体应用而定。(2) s_i 合取 s_j (即 $s_i \odot s_j$) 定义为:对各个属性(除概率因子 P 外)对应地按定义3(2)或定义7(2)分别进行合取,至于概率因子 P 有不同取法,如 $\min\{P_i, P_j\}$ 或 $P_i * P_j$,选用什么要根据具体应用而定。

定义10 设 $R = \{R', P\}$ 是一个不确定性关系模式,设 K 是 R 上的关键字集,且 $K \subseteq R'$, r 是 R 上的一个不确定性关系,如果 r 满足:(1)对于 r 中任意元组 τ 及 $A \in R$, $\tau(A)$ 是唯一的,即 $\tau(A)$ 中只有一个值;(2)对于 $\forall \tau_1, \tau_2 \in r, \forall A \in K (\tau_1(A) =_\alpha \tau_2(A))$,当且仅当 $\tau_1 = \tau_2$,其中 α 是专家给定的充分靠近1且小于或等于1的正数。则称: r 是一个不确定性范式。

3 不确定性关系代数

本节将采用“ $=_\alpha$ ”构造一个不确定性关系代数,它能较好地刻画数据的不确定性。

(1) 扩展并 ($\cup^*$)

设 r_1, r_2 分别为不确定性关系模式 R_1, R_2 上的

不确定性关系。若进行 r_1 和 r_2 的并运算,要求 $R_1 = R_2$ 及 r_1 和 r_2 有相同的关键字集。设 r_1 和 r_2 的并为 r , r 是 $R = R_1 = R_2$ 上的一个关系, K 是 R 的关键字集且 $K \subseteq R$, $A = R - K$ 是 R 的非关键字集,则:

$r = r_1 \cup r_2 = \{ \tau | (\exists \tau_1 \in r_1, \exists \tau_2 \in r_2 (k' \in K(\tau_1(k') = \tau_2(k')))) \wedge \tau = (\tau_1 \oplus \tau_2)) \vee (\tau \in r_1 \wedge (\forall \tau' \in r_2 (\exists k' \in K(\tau(k') \neq \tau'(k')))) \vee (\tau \in r_2 \wedge (\forall \tau' \in r_1 (\exists k' \in K(\tau(k') \neq \tau'(k'))))) \}$ 。其中 $\oplus$ 运算按定义 9(1) 进行。

(2) 扩展交 ($\cap^*$)

扩展交的要求和假设与扩展并相同,那么:

$r = r_1 \cap r_2 = \{ \tau | (\exists \tau_1 \in r_1, \exists \tau_2 \in r_2 (\forall k' \in K(\tau_1(k') = \tau_2(k')) \wedge \tau = (\tau_1 \odot \tau_2))) \}$, 其中 $\odot$ 运算按定义 9(2) 进行。

(3) 扩展差 ($-^*$)

扩展差的要求和假设与扩展并相同,那么:

$r = r_1 - r_2 = \{ \tau | \tau \in r_1 \wedge (\forall \tau' \in r_2 (\exists k' \in K(\tau(k') \neq \tau'(k')))) \}$

(4) 扩展笛卡尔积 ($\times^*$)

设 r_1 和 r_2 分别是不确定性关系模式 $R = \{A_1, A_2, \dots, A_n, P_1\}$ 和 $S = \{B_1, B_2, \dots, B_m, P_2\}$ 上的不确定性关系。设 $R' = R - \{P_1\}$, $S' = S - \{P_2\}$, r_1 和 r_2 的笛卡尔积为 r , r 是关系模式 $R' S' \cup \{P\} = \{A_1, A_2, \dots, A_n, B_1, B_2, \dots, B_m, P\}$ 上的一个不确定性关系,并且 r 的元组 $\tau = \{a_1, a_2, \dots, a_n, b_{n+1}, b_{n+2}, \dots, b_{n+m}, p\}$ 。设 $\mathcal{F}_1$ 和 $\mathcal{F}_2$ 分别是 r_1 和 r_2 上的元组,并假设 $\mathcal{F}_1 = (\tau_1, p_1)$, $\mathcal{F}_2 = (\tau_2, p_2)$, $\tau_1 \in R'$, $\tau_2 \in S'$, 则 r 上的元组简记为 $\tau = (\tau_1, \tau_2, p)$ 。设 $K_1 \subseteq R'$, $K_2 \subseteq S'$ 且 K_1 和 K_2 分别是 R 和 S 的关键字集,则 $K_1 K_2$ 是 $R' S' \cup \{P\}$ 的关键字集,那么:

$r = r_1 \times r_2 = \{ \tau | \tau = (\tau_1, \tau_2, p), \tau_1 \in R', \tau_2 \in S' \}$

其中 $P = P_1 \wedge P_2$, 这里 $\wedge$ 表示 $[0, 1] \times [0, 1] \rightarrow [0, 1]$ 的运算,如 $\min\{P_1, P_2\}$, $P_1 * P_2$ (普通乘法) 等等。

(5) 扩展自然连接 ($\bowtie^*$)

设 r_1 和 r_2 分别是不确定性关系模式 $R_1 = \{A_1, A_2, \dots, A_n, P_1\}$ 和 $R_2 = \{B_1, B_2, \dots, B_m, P_2\}$ 上的不确定性关系,设 $R'_1 = R_1 - P_1$, $R'_2 = R_2 - P_2$, 要对 r_1 和 r_2 做自然连接操作,必须有 $R'_1 \cap R'_2 \neq \emptyset$, 我们设 $X = R'_1 \cap R'_2$, $A = R'_1 - X$, $B = R'_2 - X$, 那么 $R = (A, X, B, P)$ 就是连接结果关系 r 的不确定性关系模式, r 满足:

$r = r_1 \bowtie r_2 = \{ \tau | \exists \tau_1 \in r_1, \exists \tau_2 \in r_2 (\forall k' \in X(\tau_1(k') = \tau_2(k')) \wedge \tau = (\tau_1(A), \tau_1(X) \odot \tau_2(X), \tau_2(B), P)) \}$, 其中 $\odot$ 运算按定义 9(2) 进行; $P = P_1 \wedge P_2$, 这

里 $\wedge$ 表示 $[0, 1] \times [0, 1] \rightarrow [0, 1]$ 的运算,如 $\min\{P_1, P_2\}$, $P_1 * P_2$ (普通乘法) 等等。

(6) 扩展投影 (π^*)

扩展投影是使用扩展并 $\cup^*$ 操作将传统投影合成一个关系。设 r_1 是不确定性关系模式 R_1 上的一个不确定性关系, $Y \subseteq R_1$, 则在 Y 上的扩展投影记为:

$r = \pi_Y^*(r_1) = \cup^*(\pi_Y(r_1))$

(7) 扩展选择 (σ^*)

设 r_1 是不确定性关系模式 R 上的一个不确定性关系, Q 是一个不确定性谓词公式。对于 Q , 我们定义 r_1 上的选择为:

$r = \sigma_Q^*(r_1)$

其中 Q 与传统的谓词公式的区别是“=”被置换成“ $=_*$ ”。

上面, 我们已给出了不确定性关系代数操作的定义, 从定义看, 显然有如下定理:

定理 1 在扩展并, 交, 差, 笛卡尔积, 自然连接, 投影和选择操作下, 不确定性范式是封闭的。

4 不确定性数据库模型的语义

设 R 是一个不确定性关系模式, r 是 R 上的一个不确定性关系, $\tau = (t, p(t))$ 是 r 上的一个元组, 其中 $p(t)$ 表示 t 在 r 中发生的概率, 则显然有 t 在 r 中不发生的概率为 $1 - p(t)$ 。

设 $r = \{\tau_1, \tau_2, \dots, \tau_m\} = \{(t_1, p(t_1)), (t_2, p(t_2)), \dots, (t_m, p(t_m))\}$ 是不确定性关系模式 R 上的一个不确定性关系。对于 $t_1, t_2, \dots, t_m$ 来说, 它们至多有 2^m 种不同的组合 (即: $t_1, t_2, \dots, t_m$ 的幂集), 对于每一种组合, 可以附加一个概率因子 P , 表示该组合中的元组发生且关系中其它元组不发生时的概率。我们称由这 2^m 个集合 (每个集合附加一个概率因子) 组成的集合为 R 的一个不确定性世界 $UW(R)$ 。下面给出严格的定义。

定义 11 设 $r = \{(t_1, p(t_1)), (t_2, p(t_2)), \dots, (t_m, p(t_m))\}$ 是不确定性关系模式 R 的一个不确定性关系, 则 R 的一个不确定性世界 $UW(R)$ 定义为:

$UW(R) = \{ (T, P) | T = t_{i_1}, t_{i_2}, \dots, t_{i_j}; P = P(t_{i_1}) \times P(t_{i_2}) \times \dots \times P(t_{i_j}) \times (1 - P(t_{i_{j+1}})) \times (1 - P(t_{i_{j+2}})) \times \dots \times (1 - P(t_m)) > 0 \}$

其中 $\{t_{i_1}, t_{i_2}, \dots, t_{i_j}\} \subseteq \{t_1, t_2, \dots, t_m\}$, $0 \leq j \leq m$ 。

显然, R 的一个不确定性世界有效且完备地描述了关系 r 中各元组随机联合发生的各种不确定性组合。现在我们利用不确定性世界的概念来描述不

确定性关系模型的语义。

定理 2 对于不确定性模式 R 的每一个不确定性世界 $UW(R)$, 设 $UW(R) = \langle T, P \rangle = \{(T_1, P_1), (T_2, P_2), \dots, (T_k, P_k)\}$, 则有: $P_1 + P_2 + \dots + P_k = 1$, 其中 $k=2^n$ 。

该定理可用数学归纳法证明, 这里省略。

下面举例说明。设有一不确定性关系如下:

姓 名	年 龄	身 高	P	
李 平	大约 30	1.65 左右	0.9	t_1
王 敏	大约 20	1.59 左右	0.2	t_2
张 明	大约 50	1.73 左右	0.5	t_3
胡 军	大约 28	1.51 左右	0.8	t_4

则它的不确定性世界为:

4 元组世界
$\{t_1, t_2, t_3, t_4\}$ 0.072
3 元组世界
$\{t_1, t_2, t_3\}$ 0.018 $\{t_1, t_2, t_4\}$ 0.072
$\{t_1, t_3, t_4\}$ 0.288 $\{t_2, t_3, t_4\}$ 0.008
2 元组世界
$\{t_1, t_2\}$ 0.018 $\{t_1, t_3\}$ 0.072 $\{t_1, t_4\}$ 0.288
$\{t_2, t_3\}$ 0.002 $\{t_2, t_4\}$ 0.008 $\{t_3, t_4\}$ 0.032
1 元组世界
$\{t_1\}$ 0.072 $\{t_2\}$ 0.002 $\{t_3\}$ 0.008 $\{t_4\}$ 0.032
0 元组世界
$\emptyset$ 0.008

总结 本文建立了一个不确定性数据库模型, 讨论了该模型的有关概念, 并构造了一个不确定性关系代数, 证明了不确定性范式在该关系代数操作下是封闭的。该模型的特点是: (1)它是传统关系模

型的扩展, 即对属性和元组进行不确定性化; (2)它既体现了认知的不确定性, 又体现了概率的不确定性, 能有效地反映客观现实; (3)它有明确的定义; (4)它还可以进行推广, 如对概率因子 P 可以象其它属性一样进行不确定性化等等。

不确定性数据库容易与其它数据库结合起来形成新的数据库, 如: 不确定性数据库与时态数据库结合可形成不确定性时态数据库, 与演绎数据库结合可形成不确定性演绎数据库等等。有关这方面的工作, 我们将另文研究。另外, 不确定性数据库的应用前景很广阔, 笔者希望借本文抛砖引玉。

参 考 文 献

- 1 何新贵. 模糊关系数据库的数据模型. 计算机学报, 1989(2): 120~126
- 2 张师超, 等. 一个模糊时态关系代数. 软件学报, 1994(9): 59~64
- 3 Gupta M M. Intelligence, uncertainty and information, Analysis and Management of Uncertainty: Theory and Applications, 1992: 3~11
- 4 Yu Hairong, Ramer A. A Combined Approach to Uncertain Data Analysis. In: Liu X, Cohen P, Berthold M, eds. Advances in Intelligent Data Analysis (IDA-97). LNCS 1280, 1997. 123~134
- 5 Pittarelli M. An Algebra for Probabilistic Databases. IEEE Transactions on Knowledge and Data Engineering, 1994, 6(2): 293~302
- 6 Buckles B P, Pntry F E. A Fuzzy Representation of Data For Relational Database. Fuzzy Sets and Syst, 1982(7): 213~226

(上接第 89 页)

因为 A^T 满秩, 其左逆或右逆存在, 所以:

$$x' = (A^T)^{-1} \times (M - M_0)$$

其中, $x' = (x, x_a)^T$

x 是原 Petri 网的点火次数向量, x' 是变换后的正则 Petri 网的点火次数向量, x_a 是添加的点火次数向量。

$$x = [x(t_1), x(t_2), \dots, x(t_n)]^T$$

$$x_a = [x(t_{n+1}), \dots, x(t_{n+r})]^T$$

到此为止, 我们就已算出点火次数向量。

参 考 文 献

- 1 Reisig W. Petri Nets: An Introduction. Springer-Verlag, Berlin, 1985

- 2 Peterson J L. Petri Net Theory and the Modeling of Systems. Prentice-Hall, Inc., England Cliffs, N. J. 07632
- 3 Murata T. State equation, controllability, and maximal matching of Petri nets. IEEE Transaction on Automatic Control, 1977, AC-22(3): 412~416
- 4 Murata T. Petri nets: Properties, analysis and applications. Procs. of the IEEE, 1989, 77(4): 541~580
- 5 Lee D, et al. Handles and reachability analysis of free choice nets. In: Proc. of the 16th Int. Conf. on Application and Theory of Petri Nets. LNCS 935. 1995. 298~315
- 6 Desel J. A structural property of free-choice nets. Petri Net Newsletters, 1986(25): 16~20
- 7 陈大新. 矩阵理论. 上海交通大学出版社