

数据扩展的相容性判定问题

赵蕾, 程国胜

(南京信息工程大学 数理学院, 江苏 南京 210044)

摘要: 以概念格为工具, 讨论数据扩展引起的相容性问题。将数据作为概念格中的对象, 在给定数据基本集并假设数据特征一定的条件下, 考虑数据扩展相容性问题, 解决了数据扩展的相容性判定问题并给出了相应的判定定理。目的是使在特征一定的情况下, 数据对象达到最大化。

关键词: 概念格; 覆盖; 数据扩展; 相容性

中图分类号: TP182 文献标识码: A 文章编号: 1000-2022(2007)03-0424-04

The Decision for Consistency of Data Enlargement

ZHAO Lei CHENG Guo-sheng

(School of Mathematics and Physics NUIST, Nanjing 210044 China)

Abstract We discuss the problem of consistency arose from data enlargement in this paper. Let data as the objects of concept lattice, we deal with the consistency of data enlargement when the data basic set is given and the data hypothetically have a certain character, and give the decision theorems of data enlargement consistency.

Key words concept lattice, covering, data enlargement, consistency

0 引言

Wille^[1]首次提出概念格理论, 该理论现在已成为数据分析和规则提取的一种有效工具, 并被广泛研究^[2-3]和应用到机器学习^[4]、软件工程^[5-7]和信息获取^[8-10]等领域。在对数据进行分析时, 一方面采集数据并从其中提取数据的特征或属性^[11-14]; 另一方面是数据库的管理与维护, 在确定的数据特征下合理增删修改数据库里的数据, 在实际应用中有着重要意义^[15]。本文尝试以概念格为工具, 在假设数据特征(概念格的属性集合)确定的条件下, 将数据作为概念格中的对象, 讨论数据扩展相容性问题, 给出了数据相容性的判定定理。这为数据库数据在相同属性前提下的对象扩展提供了理论依据, 丰富了数据知识的研究, 对合理有效地管理数据库有着重要意义。

1 基础知识

设论域 U 表示数据库中所有数据对象。 A 表示数据的属性集, $O \subseteq U$ 为对象集, R 是 O 和 A 之间的二元关系, 即 $R \subseteq O \times A$, 称三元组 (O, A, R) 是一个形式背景, oRa 表示 o 与 a 之间存在关系 R 。

设 $P \subseteq O, Q \subseteq A$, 定义运算

$$\begin{aligned} P^* &= \{a \in A \mid oRa, \forall o \in P\}, \\ Q^* &= \{o \in O \mid oRa, \forall a \in Q\}. \end{aligned}$$

此处, P^* 表示 P 对象具有的所有属性的集合, Q^* 表示具有属性 Q 的 O 中所有对象的集合。

定义 1^[2] 二元组 (P, Q) ($P \subseteq O, Q \subseteq A$) 称为形式背景 (O, A, R) 中的形式概念(简称概念), 其中 $P^* = Q, Q^* = P$ 。 P 和 Q 分别称为概念的外延和内涵。

收稿日期: 2006-03-31 改回日期: 2006-09-30

基金项目: 江苏省自然科学基金资助项目(05KJD-110-123); 南京信息工程大学科研启动资金资助项目

作者简介: 赵蕾(1979-), 女, 江苏徐州人, 硕士, 研究方向: 数据处理与计算理论, njlzha@sohu.com.

对于形式背景 (O, A, R) 中任意两个概念 (P_1, Q_1) 和 (P_2, Q_2) , 定义以下偏序关系

$$(P_1, Q_1) \leqslant (P_2, Q_2) \Leftrightarrow P_1 \leqslant P_2 \text{ (或等价的) } Q_2 \leqslant Q_1.$$

则形式背景 (O, A, R) 中的所有概念连同其上定义的偏序关系称为概念格, 记为 $L(O, A, R)$ 。易见 $L(O, A, R)$ 是一个完备格^[2]。

形式背景 (O, A, R) 中的概念所具有的基本性质见文献 [2]。

定义 2 设 $L(O_1, A, R_1)$ 和 $L(O_2, A, R_2)$ 是两个概念格 ($O_1, O_2 \in U$), 如果对于 $\forall (x, y) \in L(O_1, A, R_1)$ 总存在 $(x', y') \in L(O_2, A, R_2)$, 使得 $y = y'$, 则称 $L(O_2, A, R_2)$ 覆盖 $L(O_1, A, R_1)$, 记作 $L(O_2, A, R_2) \geqslant L(O_1, A, R_1)$ 。

若 $L(O_2, A, R_2) \geqslant L(O_1, A, R_1)$ 且 $L(O_1, A, R_1) \geqslant L(O_2, A, R_2)$, 则称两个概念格同构, 记作 $L(O_1, A, R_1) \cong L(O_2, A, R_2)$ 。

本文约定, 在形式背景 (O, A, R) 中, $\forall x \in O, x^* = \{x\}^* \neq f$, $\forall y \in A, y^* = \{y\}^* \neq f$ 。

2 数据扩展与相容性判定定理

在形式背景 (O, A, R) 下, $\forall G \subseteq O$, 记 $R_G = R \cap (G \times A)$, 则 (G, A, R_G) 也是形式背景, 以下以 (G, A, R_G) 为基础讨论。设 $y^* \in O, y^{*c} \in G$, 显然 $R_G = R, y^{*0} = y^*, y^{*c} = y^* \cap G = y^* \cap G, y^{*c} \subseteq y^*$ 。

定理 1 设 (O, A, R) 为形式背景, $\forall G \subseteq O, G \neq f$, 总有

$$L(O, A, R) \geqslant L(G, A, R_G).$$

证明 $\forall (x, y) \in L(G, A, R_G)$, 则 $(y^*, y^{**}) \in L(O, A, R)$ 。下证 $y = y^{**}$ 。

首先 $y \subseteq y^{**}$, 其次 $x = y^{*c} \subseteq y^*$, 则有 $y = x^* \supseteq y^{**}$, 从而 $y \supseteq y^{**}$ 。故 $y = y^{**}$ 。

推论 1 设 $G \subseteq F \subseteq O, G \neq f, F \neq f$, $F \neq f$, R_F 定义类同 y^{*c}, R_G , 则有

$$L(F, A, R_F) \geqslant L(G, A, R_G).$$

定义 3 对于形式背景 (G, A, R_G) , 若存在对象集 F , 且 $G \subseteq F \subseteq O$, 使得 $L(F, A, R_F) \cong L(G, A, R_G)$, 则称 F 是 (G, A, R_G) 的相容集。若 $\forall u \in O - F, L(F + \{u\}, A, R_{F+\{u\}})$ 不同构于 $L(G, A, R_G)$, 记作 $L(F + \{u\}, A, R_{F+\{u\}}) \not\cong L(G, A, R_G)$, 称 F 是 (G, A, R_G) 的一个数据扩展。 G 称为数据基本集。

注 (1) G 本身是 (G, A, R_G) 的相容集;

(2) $L(F + \{u\}, A, R_{F+\{u\}}) \not\cong L(G, A, R_G)$ 是指 $L(F + \{u\}, A, R_{F+\{u\}}) \geqslant L(G, A, R_G)$ 或 $L(F + \{u\}, A, R_{F+\{u\}}) \leqslant L(G, A, R_G)$ 不能同时满足。

例 1 给定形式背景 (O, A, R) , 其中 $O = \{1, 2, 3, 4, 5, 6, 7, 8\}, A = \{a, b, c, d, e, f, g, h\}$ 其二元关系如表 1 所示。

表 1 二元关系

Table 1 Binary relation

	a	b	c	d	e	f
1	1	0	0	0	0	1
2	1	0	1	0	0	1
3	1	0	1	0	0	1
4	0	1	0	1	1	0
5	0	1	0	1	0	0
6	0	1	0	1	1	0

该形式背景有 6 个概念: $(23, af)$, $(123, af)$, $(46, bd)$, $(456, bd)$, (O, f) , (f, A) 。

设数据基本集 $G_1 = \{4\}$, 则形式背景 (G_1, A, R_{G_1}) 的相容集是 $\{4, 6\}$, 并且也是 (G_1, A, R_{G_1}) 的一个数据扩展。

关于数据扩展,有如下结论。

定理 2 设形式背景 (O, A, R) , $\forall G \subseteq O$, 则对于任何形式背景 (G, A, R_G) , 扩展必定存在。

证明 若对于 $\forall u \in O - G$, 都有 $L(G + \{u\}, A, R_{G + \{u\}}) \neq L(G, A, R_G)$, 则 G 本身就是扩展, 若存在 $u \in O - G$, 使得 $L(G + \{u\}, A, R_{G + \{u\}}) \cong L(G, A, R_G)$, 则考虑 $G_1 = G + \{u\}$, 若对于 $\forall u_1 \in O - G_1$, 都有 $L(G_1 + \{u_1\}, A, R_{G_1 + \{u_1\}}) \neq L(G_1, A, R_{G_1})$, 则 G_1 是扩展, 否则, 考虑 $G_2 = G_1 + \{u_1\}$, 重复上述过程, 由于 $O - G$ 是有限集, 这样总可以找到一个扩展。

现在给出相容集的判定定理。

定理 3 设形式背景 (G, A, R_G) , $G \subset F \subseteq O$, $G \neq f$, $F \neq f$, $C = F - G$, 则 F 是 (G, A, R_G) 的一个相容集 $\Leftrightarrow \forall K \subseteq C$, $K \neq f$, 有 $(K^{**} - C)^* = (K^{**} \cap G)^* = K^*$ 。

证明 (1) 必要性: 由 F 是相容集知, $L(G, A, R_G) \geq L(F, A, R_F)$, $\forall K \subseteq F - G$, $K \neq f$, 总有 $(K^{**}, K^*) \in L(F, A, R_F)$ 。因而 $\exists M \subseteq G$, 使得 $(M, K^*) \in L(G, A, R_G)$, 于是 $M^* = K^*$, 又 $M = K^{**} \cap G = K^{**} \cap C$, 且 $K^{**} - C = K^{**} + G - F = K^{**} \cap G$, 因此, $(K^{**} - C)^* = (K^{**} \cap G)^* = M^* = K^*$ 。

(2) 充分性: 若要证明 F 是相容集, 只需证明 $\forall (x, y) \in L(F, A, R_F)$, 有 $(x \cap G, y) \in L(G, A, R_G)$, 则 $L(G, A, R_G) \geq L(F, A, R_F)$ 。亦即证明 $(x \cap G)^* = y^* \cap G = x \cap G$ 。

首先, $y^* \cap G = y^* \cap G = x \cap G$ 。其次证明 $(x \cap G)^* = y$ 。显然 $x = (x \cap G) \cup (x \cap C)$ 。若 $x \cap C = f$, 则 $y = x^* = (x \cap G)^*$ 。

若 $x \cap C \neq f$, 则由于 $x \cap C \subseteq C$, 且由已知 $((x \cap C)^{**} \cap G)^* = (x \cap C)^*$, 于是 $(x \cap C) \subseteq x \Rightarrow (x \cap C)^{**} \subseteq x^{**} = y^* = x$, 则 $(x \cap C)^{**} \cap G \subseteq x \cap G$, 从而 $(x \cap C)^* = ((x \cap C)^{**} \cap G)^* \supseteq (x \cap G)^*$ 。于是, $y = x^* = (x \cap G)^* \cap (x \cap C)^* = (x \cap G)^*$ 。

综上, $(x \cap G)^* = y$ 。

推论 2 设形式背景 (G, A, R_G) , $G \subseteq F$, $F \neq f$, 若 F 相容集, 则 $G^* \subseteq (F - G)^*$ 。

证明 由 F 是相容集及定理 3 知, $((F - G)^{**} \cap G)^* = (F - G)^*$, 又 $(F - G)^{**} \cap G \subseteq G$, 所以 $((F - G)^{**} \cap G)^* \supseteq G^*$, 从而 $G^* \subseteq (F - G)^*$ 。

定理 4 设形式背景 (G, A, R_G) , $G \subseteq F$, $F \neq f$, $C = F - G$ 。则

F 是相容集 $\Leftrightarrow \forall K \subseteq C$, $K \neq f$, $\exists M \subseteq G$, $M \neq f$, 使得 $M^* = K^*$ 。

证明 必要性由定理 3 即得。

充分性: 由 $M^* = K^*$ 得 $M \subseteq M^{**} = K^{**}$, 且 $M \subseteq G$, $M \subseteq K^{**} \cap G$, $(K^{**} \cap G)^* \subseteq M^* = K^*$, 又 $(K^{**} \cap G)^* \supseteq K^* \cup G^* \supseteq K^*$, 从而 $(K^{**} \cap G)^* = K^*$, 故由定理 3 知 F 是相容集。

定理 5 设形式背景 (G, A, R_G) , $G \subseteq F$, $F \neq f$, $C = F - G$ 。则

F 是相容集 $\Leftrightarrow \forall a \in C$, $\exists M \subseteq G$, $M \neq f$, 有 $M^* = a^*$ 。

证明 必要性由定理 4 即得。

充分性: $\forall K \subseteq C$, $K \neq f$, 记 $K = \{a_t \mid t \in \tau\}$, 由已知, $\forall a_t \in K \subseteq C$, $\exists M_t \subseteq G$, $M_t \neq f$, 有 $M_t^* = a_t^*$, 则 $K^* = (\bigcup_{t \in \tau} a_t)^* = \bigcap_{t \in \tau} a_t^* = \bigcap_{t \in \tau} M_t^* = (\bigcup_{t \in \tau} M_t)^*$, 令 $M = (\bigcup_{t \in \tau} M_t)$, 则 $M \subseteq G$, $M \neq f$, $M^* = K^*$, 由定理 4 知 F 是相容集。

并非所有的 $G \subseteq O$ 都有 F 为其相容集, 且 $G \subset F$, 显然定理 5 有如下的逆否命题成立。

推论 3 设形式背景 (G, A, R_G) , 且 (G, A, R_G) 只有 G 本身为其相容集

$\Leftrightarrow \forall e \in O - G$, 不存在 $M \subseteq G$, 使得 $M^* = e^*$ 。

定理 6 设形式背景 (G, A, R_G) , $G \subseteq F$, $F \neq f$, $C = F - G$ 。则:

F 是相容集 $\Leftrightarrow \forall a \in C$, $(a^{**} - C)^* = (a^{**} \cap G)^* = a^*$ 。

证明 必要性由定理3即得。

充分性: $\forall a \in C, (a^{**} \cap G)^* = a^*$, 记 $M = a^{**} \cap G$, 则 $M \subseteq G, M \neq f$, 且 $M^* = a^*$, 则由定理5知 F 是相容集。

类似的, 定理6有如下的逆否命题成立。

推论4 设形式背景 (G, A, R_G) , 且 (G, A, R_G) 只有 G 本身为其相容集

$$\Leftrightarrow \forall a \in O - G, (a^{**} - C)^* = (a^{**} \cap G)^* \neq a^*.$$

例2 对于例1中的形式背景, 又设基础集 $G_2 = \{2, 3\}$, 则形式背景 (G_2, A, R_{G_2}) 的相容集只能是 $\{2, 3\}$ 本身, 事实上, 若设 $\{1, 2, 3\}$ 是其相容集, 则 $C = \{1\}, (1^{**} - C)^* = (2, 3)^* = \{a, c, f\} \neq 1^* = \{a, f\}$ 。从而 (G_2, A, R_{G_2}) 的扩展只能是 $G_2 = \{2, 3\}$ 本身。

定理7 设形式背景 $(G, A, R_G), G \subseteq F, F \neq f, C = F - G$, 则

$$F \text{ 是相容集} \Leftrightarrow L(G, A, R_G) \geq L(C, A, R_C).$$

证明 必要性: 由 F 是相容集知 $L(G, A, R_G) \geq L(F, A, R_F)$, 由于 $C \subseteq F$, 根据推论1有 $L(F, A, R_F) \geq L(C, A, R_C)$, 从而 $L(G, A, R_G) \geq L(C, A, R_C)$ 。

充分性: 由 $L(G, A, R_G) \geq L(C, A, R_C)$, 又 $\forall K \subseteq C, K \neq f$, 有 $(K^{**}, K^*) \in L(C, A, R_C)$, 故 $\exists M \subseteq G, M \neq f$, 使得 $(M, K^*) \in L(G, A, R_G)$, 从而 $M^* = K^*$, 由定理4知 F 是相容集。

3 结语

本文在概念格理论基础上对数据库中的数据构造进行研究, 得到了数据对象相容性判定定理, 这使得数据扩展有了理论依据。数据相容性问题, 特别是多种类型数据相容性问题的研究, 对数据库的管理与维护有着极其重要的实用价值。

参考文献:

- [1] Wille R. R-estructuring lattice theory: An approach based on hierarchies of concepts [M]. Dordrecht: Reidel, 1982: 445-470.
- [2] Ganter B, Wille R. Formal concept analysis is mathematics foundations [M]. Berlin: Springer-Verlag, 1999.
- [3] Yao Y Y. Concept lattice in rough set theory [R]. Canada: Proceedings of 2004 Annual Meeting of North American Fuzzy Information Processing Society, 2004: 796-801.
- [4] Zupan B, Bohrnec M. Learning by discovering concept hierarchies [J]. Artificial Intelligence, 1999, 109(1/2): 211-242.
- [5] Tonella U. Using a concept lattice of decompositions slices for program understanding and impact analysis [J]. IEEE Transactions on Software Engineering, 2003, 29(6): 495-509.
- [6] Arcalval G, Mdens T. Analyzing object-oriented application frameworks using concept analysis [R]. Lecture Notes in Computer Science 2426, 2002: 53-63.
- [7] Dekel U, Gil Y. Revealing class structure with concept lattices [R]. Canada: Proceedings of the 10th Working Conference on Reverse Engineering, 2003: 353-365.
- [8] Valtchev P, Missaoui R, Godin R, et al. Generating frequent item sets incrementally: Two novel approaches based on gabis lattice theory [J]. Journal of Experimental & Theoretical Artificial Intelligence, 2002, 14(2/3): 115-142.
- [9] 梁吉业, 王俊红. 基于概念格的规则产生集挖掘算法 [J]. 计算机研究与发展, 2004, 41(8): 1339-1344.
- [10] 谢志鹏, 刘宗田. 概念格与关联规则发现 [J]. 计算机研究与发展, 2000, 37(12): 1415-1421.
- [11] 范金城, 梅长林. 数据分析 [M]. 北京: 科学出版社, 2002.
- [12] 胡可云, 陆玉昌, 石纯一. 概念格及其应用进展 [J]. 清华大学学报, 2000, 40(9): 77-81.
- [13] 张文修, 魏玲, 祁建军. 概念格的属性约简理论与方法 [J]. 中国科学: E辑, 2005, 35(6): 628-639.
- [14] 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.
- [15] Michael A. Oracle数据库管理与维护技术手册 [M]. 江漫, 等译. 北京: 清华大学出版社, 2002.