

一种区间数分解与定标算法及其扩展形式背景的概念格生成方法

刘耀华 周 文 刘宗田

(上海大学计算机工程与科学学院 上海 200072)

摘 要 现有的概念格模型无法处理既包含以布尔值表示的信息,又包含以标量、模糊数及区间数表示的信息。因此,针对包含所有这些类型信息的扩展形式背景提出它的处理方法,在此基础上,生成经扩展的概念格,是一项有意义的工作。提出了一种新的区间数分解与定标算法,以处理含有多种类型的扩展形式背景,并给出了相应的扩展格生成算法。最后,实验表明,该方法具有良好的效果。

关键词 概念格,概念格构造算法,形式概念分析,模糊概念格,区间概念格

中图法分类号 TP311 **文献标识码** A

Interval Scaling Algorithm and its Concept Lattice Construction from Extended Formal Context

LIU Yao-hua ZHOU Wen LIU Zong-tian

(School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China)

Abstract The existing concept lattice model is unable to process data which contains not only fuzzy information but also scalar and Boolean information. The extended formal context includes many kinds of information such as scalar, fuzzy, Boolean, and interval. Therefore, how to build concept lattice from extended formal context, is a meaningful study. An interval scaling algorithm is proposed here to deal with the extended formal context and a corresponding concept lattice construction method is produced. In the end of this paper, the experimental results show that this algorithm is useful.

Keywords Concept lattice, Construction algorithm, Formal concept analysis, Fuzzy concept lattice, Interval concept lattice

形式概念分析(Formal Concept Analysis, FCA)是一种基于序论的数据分析方法,是研究知识表示的领域,其核心数据结构——概念格从本质上描述了对对象和属性之间的联系,在一定程度上是对客观世界的一种高度简化的描述形式。Wille 基于这种简化,系统地研究了概念的有序性质、格代数性质以及概念格与形式背景的同构性质,开创了形式概念分析这一新的研究领域^[1]。经过近十多年的研究,概念格已广泛地应用于知识管理^[2]、机器学习^[3]、信息检索^[4]和软件工程^[5]等诸多领域。

然而经典的概念格模型仅能处理布尔值。近年来,已经有人研究将概念格的表示能力扩展到了标量值、模糊集和粗糙集上,并提出了扩展的概念格模型。针对信息的模糊性和不确定性,研究一种新的扩展形式背景的概念格及其描述信息,具有重要的实用意义。

本文提出了一种新的区间数定标算法及其扩展形式背景的概念格生成方法,能处理包括布尔、模糊数、枚举、区间类型的形式背景,扩展后的概念格可以处理大多数的常用数据,并支持规则提取、概念聚类、属性约简等运算。

1 形式概念分析及其扩展

对于模糊、不确定信息上的形式概念分析研究,国外已有了一些研究成果。

Wolff 提出了一种基于模糊信息的表示法^[6],将传统的形式背景中的属性用模糊语言变量值表示,依据标度分类形式背景的对象,构造基于标度的格。

Girard 和 Ralambondrainy 提出了基于模糊量词集的方法,通过在数据类型定义中插入模糊量词集处理模糊信息,并研究了由模糊量词描述的形式背景的概念格的构造^[8,9]。实际上,精确信息与模糊信息常并存于一个多值背景中,此方法仅将模糊属性加上标识,处理时仍将多值背景转换为单值背景来计算格,有一定的局限性。

为解决现实中出现的区间值表示问题,区间分析被广泛研究。Burusco 等引入了形式概念分析的方法,解决区间数的表示和处理问题,引入了 L-fuzzy 概念理论处理区间数问题。周文和刘宗田等^[16]提出了一种基于区间值分解的属性定标方法,进而在获得的区间值形式背景基础上,构建了一种

到稿日期:2008-11-18 返修日期:2009-02-16 本文受国家自然科学基金(60575035),上海高校选拔培养优秀青年教师科研专项基金(shu-07027),上海市重点学科建设项目(J50103)资助。

刘耀华(1982—),男,硕士研究生,主要研究方向为数据挖掘等,E-mail:yao618@163.com;周文(1979—),女,博士,讲师,主要研究方向为人工智能等;刘宗田(1946—),男,教授,博士生导师,主要研究方向为人工智能、数据挖掘等。

新的区间数概念格模型,并提出了一种渐进的区间格生成算法,将区间格模型应用于网格资源匹配中,用以对网格中的资源进行建模。

本文针对文献[16]给出了一种新的区间分析与属性定标方法,在结合区间分析与属性定标的基础上改进了基于区间值分解的属性定标方法;改进了能处理连续隶属度值的形式背景的模糊概念格的属性定标方法;在对枚举型形式背景的属性定标时,提出了枚举类型和枚举区间类型两个概念,以处理特定的形式背景。

2 扩展的形式背景

本节将给出一种形式背景表示,该形式背景可允许多种类型数据,将包含布尔类型、数字类型、枚举类型及区间类型的信息。

定义 1(扩展形式背景) 该扩展的形式背景可以支持布尔类型、数字类型、枚举类型、区间类型的数据,因此其形式背景表示为 $K=(G, M, T, F)$,其中 G 为对象集, M 为属性集, T 为属性的类型集,映射 F 称为隶属度函数。 M 与 T 是相互对应的, T 可以为布尔值 (Bool)、模糊数 (FuzzyNumber)、枚举 (Enum)、枚举区间 (EnumRange) 和区间数 (Interval) 这 5 种类型。

定义 2(扩展形式背景参数) 对于属性集 M 、属性的类型集 T 和对象集 G ,由于不同属性的类型不同,所以它们的参数也不同。 Bool 类型的属性没有参数; FuzzyNumber 类型的属性需要确定一组参数 $N'=\{N_1, N_2, \dots\}$; Enum 类型的属性不需要参数; EnumRange 属性需要多组参数 $E'=\{E_1, E_2, \dots\}$, $E_1=\{e_1, e_4, \dots\}$, $E_2=\{e_2, e_3, \dots\}$; Interval 类型的属性不需要参数。

定义 3(FuzzyNumber 类型的属性参数意义) 对于 FuzzyNumber 类型的属性参数 N' 和对象集 G ,对于该属性下的对象 o ,如果 $o > N_m$ 并且 $o \leq N_n$ (满足 $m+1=n$),则对象 o 在新的形式背景上的由该 FuzzyNumber 类型属性衍生出来的新属性 $(N_m \sim N_n)$ 上将被置为 1,而在由该 FuzzyNumber 类型属性衍生出来的其他新属性上将被置为 0。

定义 4(EnumRange 类型的属性参数意义) 对于 EnumRange 类型的属性参数 E' , $E_1, E_2, \dots$ 等和对象集 G ,对于该属性下的对象 o ,如果 $o \subseteq E_m$ (满足 $m+1=n$),则对象 o 在新的形式背景上的由该 EnumRange 类型属性衍生出来的新属性 (E_m) 上将被置为 1,而在由该 EnumRange 类型属性衍生出来的其他新属性上将被置为 0。

3 一种新的区间数定标算法

本节将给出一种新的区间数分解定标算法,它是对文献[16]中的区间数分解算法的改进。

分解算法如下:

- 1) 当该列数据类型为 Interval 类型时,遍历该列数据,得到一组区间数;
- 2) 分解该组数据区间(因为其为 Interval 类型,故有最小最大值),得到两个最值数据;
- 3) 判断这两个数据是否被一初始为空的数据集合包含,如果不被包含,则将其插入到该集合;
- 4) 遍历完成后,对得到的数据集合进行排序,将依次

分解出数据集合长度减 1 个数据区间,得到的数据区间即为衍生属性。

相应的定标方法设计如下:

- 1) 当该列数据类型为 Interval 类型时,拆分当前要定标的数据区间(因为其为 Interval 类型,故有最小最大值),得到两个最值数据;
- 2) 对于上述分解算法得到的多个数据区间,如果存在这样的数据区间 W 被当前数据区间完全包含(比较两个数据区间的最值),则修改当前行对应的衍生属性列(数据区间 W)的数据为 1。

4 扩展形式背景的格生成原理及算法

由于扩展后的形式背景无法直接被现有的概念格生成算法所应用,故要对其进行转换和处理。原理是根据每一个属性的类型及其类型的对应参数将该属性进行转换,衍生出多个不同的属性,并依据新的属性集转换原始对象集数据。

4.1 算法伪码

- 1) 分别读取属性集 M 和属性的类型集 T 。
- 2) 新建一个二维数据区域 L (M 的长度为列数,对象集 U 的长度+1 为行数),将 M 存放至 L 的第一行,其他位置数据全部置为 0。
- 3) 按列的方式开始遍历对象集合 U ,得到一系列数据 O_c ,并通过该列的索引 i 分别访问 M 和 T ,得到 O_c 的属性 M_c 和类型 T_c ;
- 4) 根据 T_c 的类型衍生属性 M_c 至一或多个,并将衍生过的新属性插入并替换至 L 。
 - a) 当 T_c 为 Bool 类型时,无衍生属性。
 - b) 当 T_c 为 FuzzyNumber 类型时,根据其参数 N' 内的数据将属性衍生多个,如 $(N_1 \sim N_2), (N_3 \sim N_4) \dots (N_x \sim N_y)$,如果 N' 内数据只有 1 个,不衍生新属性。
 - c) 当 T_c 为 Enum 类型时,遍历该列 O_c 得到 O_{cm} 。如果 $O_{cm}! \subseteq \{E_1, E_2, \dots, E_m\}$,则衍生一个新属性 E_n ,如 $(E_1), (E_2) \dots (E_n)$ 。
 - d) 当 T_c 为 EnumRange 类型时,根据其参数 E' 内的数据将属性衍生多个,如 $(E_1), (E_2) \dots (E_n)$ 。
 - e) 当 T_c 为 Interval 类型时,遍历该列 O_c 得到 O_{cm} ,拆分 O_{cm} 数据(因为其为 Interval 类型,故有最小最大值),得到 O_{cm}' 和 O_{cm}'' 。分别判断这两个数据是否被 $I=\{I_1, I_2, \dots, I_n\}$ (初始为空)包含。如果不被包含,则将其插入 I 。遍历完成后,对 I 进行排序得到 I' ,并衍生多个新属性 $(I_1' \sim I_2'), (I_3' \sim I_4') \dots (I_x' \sim I_y')$ 。
- 5) 按行的方式遍历列 O_c ,得到单个数据 o 。如果数据 o 符合衍生过的属性条件(包含或在该范围内),则修改该数据在 L 中对应位置的数据为 1。
 - a) 当 T_c 为 Bool 类型时,直接对 L 中数据赋值。
 - b) 当 T_c 为 FuzzyNumber 类型时,根据新属性 $(N_1 \sim N_2), (N_3 \sim N_4) \dots (N_x \sim N_y)$,如果 $o > N_m$ 并且 $o \leq N_n$ (满足 $m+1=n$),则修改 L 中当前行的 $(N_m \sim N_n)$ 属性列数据为 1。
 - c) 当 T_c 为 Enum 类型时,根据新属性 $(E_1), (E_2) \dots (E_n)$,如果 $o = E_m$,则修改 L 中当前行的 (E_m) 属性列数据为 1。
 - d) 当 T_c 为 EnumRange 类型时,根据新属性 $(E_1),$

$(E_2), \dots, (E_n)$ 以及参数 $E_1 = \{e_1, e_4, \dots\}, E_2 = \{e_2, e_3, \dots\}, \dots$, 如果 $\alpha \subseteq E_m$, 则修改 L 中当前行的 (E_m) 属性列数据为 1。

e) 当 T_C 为 Interval 类型时, 拆分数据 α (因为其为 Interval 类型, 故有最小最大值), 得到 α' 和 α'' , 并根据新属性 $(I_1' \sim I_2'), (I_3' \sim I_4') \dots (I_x' \sim I_y')$, 如果 $\alpha' \leq I_m'$ 并且 $\alpha'' \geq I_n'$ (满足 $m+1=n$), 则修改 L 中当前行的 $(I_m' \sim I_n')$ 属性列数据为 1。

6) 循环遍历数据。遍历完成后 L 中数据即为转换过的传统形式概念的形式背景数据。

7) 应用 L 中数据进行概念格生成, 并设定最小支持度、可信度, 以提取关联规则。

4.2 算法复杂度

各属性类型的时间复杂度如下。

1) Bool 类型属性: $T(n) = n(n$ 为对象集合 U 包含的对象个数);

2) FuzzyNumber 类型属性: $T(n) = n * \lg(a)$ (n 为对象集合 U 包含的对象个数, a 为当前属性参数 N' 包含的值个数)

3) Enum 类型属性: $T(n) = n + n * b \approx n * b$ (n 为对象集合 U 包含的对象个数, b 为当前属性的枚举值总数);

4) EnumRange 类型属性: $T(n) = n * c$ (n 为对象集合 U 包含的对象个数, c 为当前属性参数 E' 包含的枚举集合个数);

5) Interval 类型属性: $T(n) = n + n * d \approx n * d$ (n 为对象集合 U 包含的对象个数, d 为当前属性下数据最小最大值集合 I 包含的值个数);

扩展形式背景的格生成算法的时间复杂度为 $T(n)' = (n * S_1) + (n * \lg(a) * S_2) + (n * b * S_3) + (n * c * S_4) + (n * d * S_5)$, 其中 S_1, S_2, S_3, S_4, S_5 分别表示 Bool, FuzzyNumber, Enum, EnumRange 和 Interval 类型属性的数目。由于硬件限制, 无法做出大数据量的属性实验, 故其时间复杂度没有约简 (理论约简值为 n)。

5 实验与分析

本实验采用 JAVA 语言编写, 应用了 ConExp 对转换过的形式背景进行概念格生成和关联规则提取, 实现了该扩展形式背景的格构造算法。实验数据采用 1000 组随机生成的样本数据, 该样本数据原始属性集为 9 个, 转换为传统形式背景后属性集为 30 个。为了提高数据真实性和程序的运算性能, 对随机生成的样本数据的其中几列数据设定了百分比取值概率。

模型的空间复杂度由表 1 和表 2 体现。表 1 是测试属性数 (原始为 9 个, 转换后 30 个) 不变的情况下对象数与格规模的关系。其中, 第一行为对象数, 初始情况取对象数为 200 个, 依次递增 200 个对象数, 到对象数为 1000 个时为止; 第二行为生成格的节点数。

表 1 对象数与格规模的关系

对象数	200	400	600	800	1000
格节点数	12518	30436	46923	62335	81954

表 2 是测试在对象数为 1000 不变时, 属性数与格规模的关系。第一行为原始属性数, 初始属性数为 2 个, 依次递增 2 个属性数, 到属性数为 9 个时为止; 第二行为转换过后的属性

数; 第三行为生成格的节点数。

表 2 属性数与格规模的关系

原始属性数	2	4	6	8	9
转换后属性数	4	11	16	25	30
格节点数	10	414	3137	41432	80013

算法的时间复杂性由表 3 和表 4 体现。表 3 给出测试属性数 (原始为 9 个, 转换后 30 个) 不变的情况下对象数与运行时间的关系。表 4 给出对象数 (1000 个) 不变的情况下属性数与运行时间的关系。

表 3 对象数与运行时间的关系

对象数	200	400	600	800	1000
转换所需时间(毫秒)	125	172	187	187	203
生成格所需时间(毫秒)	3906	7750	12640	17813	25750

表 4 属性数与运行时间的关系

属性数	2	4	6	8	9
转换后属性数	4	11	16	25	30
转换所需时间(毫秒)	78	171	172	203	203
生成格所需时间(毫秒)	297	547	1563	12422	26954

从表 1 到表 4 可以看出, 属性数的改变对运行时间的改变影响较大, 而对象数的变化对运行时间的影响相对较小。同样, 属性数的改变对生成格的规模影响较大, 而对象数的变化对生成格的规模的影响相对较小。

结束语 本文提出了一种新的区间数分解与定标算法及其扩展形式背景的格生成方法。与传统的形式背景的概念格相比, 扩展形式背景的概念格模型能处理的数据类型更多、更广泛, 也更实用。文中的试验也表明, 在属性数改变的情况下, 生成格的空间复杂度和时间复杂度均大幅上升。因此, 在提高概念格生成算法的性能的同时, 该模型还有待改进。在对于属性的转换上, 衍生出更少的新属性并能满足新形式背景的要求, 以此为入口, 可以提高程序的性能。

参考文献

- [1] Wille R. Concept lattices and conceptual knowledge systems[J]. Computers and Mathematics with Application, 1992, 23: 493-522
- [2] Krohn U, Davies N J, Weeks R. Concept lattices for knowledge management[J]. BT Technol J, 1999, 17(4): 118-116
- [3] Kuznetsov S O. Machine Learning on the Basis of formal Concept Analysis[J]. Automation and Remote Control, 2001, 62(10): 1543-1564
- [4] Carpineto C, Romano G. Information retrieval through hybrid navigation of lattice representations[J]. Int. J. Human-Computer Studies, 1996, 45: 553-578
- [5] Godin R, Mili H, Mineau G, et al. Design of class hierarchies based on concept(Galois) lattices[J]. Theory and Application of Object Systems, 1998, 4(2): 117-134
- [6] Wolff K E. Conceptual interpretation of fuzzy theory[C]//The 6th European Congress on Intelligent Techniques and Soft Computing, 1998, 1: 555-562
- [7] Burusco A, Fuentes R. The study of L-fuzzy concept lattices[J]. Mathware Soft Comput, 1994, 3: 209-218
- [8] Girard R, Ralambondrainy H. Conceptual classification from imprecise data[C]// Proceedings of Information Processing and Management of Uncertainty in Knowledge-Based System, 1996, 1: 247-252

- [9] Girard R, Ralambondrainy H. Conceptual classification from structured and fuzzy data[C]//Proceedings of the 6th IEEE International Conference on Fuzzy System. 1997;135-142
- [10] Bordat J P. Calcul pratique du treillis de galois d'une correspondance[J]. Math. Et Sci. Humaines, 24eme année, 1986, 96;31-47
- [11] Godin R, Missaoui R, Alaoui H. Incremental concept formation algorithms based on Galois(concept) lattices[J]. Computational Intelligence, 1995, 11(2);246-267
- [12] 谢志鹏, 刘宗田. 概念格的快速渐进式构造算法[J]. 计算机学报, 2002, 25(5);490-496
- [13] Njiwoua P, Mephu N E. A parallel algorithm to build concept lattice[C] // Proceedings of 4th Groningen Intl. Information

Technical Conference for Students. 1997;103-107

- [14] Liu Zong-tian, Li Lian-sheng, Zhang Qing. Research on a Union Algorithm of Multiple Concept Lattices[C]//Proceedings of 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Springer, 2003;533-540
- [15] Shen Xia-jiong, Liu Zong-tian, et al. Isomorphic generating of concept lattices[C]//IEEE International Conference on Granular Computing. 2005, 1;249-252
- [16] Zhou Wen, Liu Zong-tian, et al. An Interval Lattice Model for Grid Resource Searching[C]//Shi Y, et al., eds. ICCS 2007: Proceedings of the 7th International Conference on Computational Science, Part II. LNCS 4488, 2007;689-695

(上接第 163 页)

第 9 秒时并发的请求数达到 19 个,此时系统仍正常运行;但当继续增加测试压力,即并发的请求数达到 20 个时,将会导致有的请求得不到满足,因此,在本文的测试环境下,支持的最大用户并发数为 19 个。

5 相关工作

为了能有效地把 SIWS 进行组合形成 workflow,目前要进行两方面的研究工作:一种基于 Petri 网^[12];另一种基于有向非循环图(Directed acyclic graph, DAG)^[1,4,5,7]。文献[12]分析了 Web 服务组合之间的约束关系,提出了基于 Horn 子句集的建模方法,并给出了相应的 Petri 网模型,这对 Web 服务的最优化组合提供了一种参考方法。目前的研究表明, Petri 网是一种很好的过程建模和分析工具,但是用来描述 Petri 网状态的标记随着 Petri 状态的变换而不断运动,这也为 Petri 网流程的分析造成了一定困难^[12]。考虑到 OGC(Open Geo-Spatial Consortium)提出的 OWS(OGC Web Service)和 ISO19115 等国际标准对 SIWS 都有着很好的约束,地理 Web 服务之间有着规范的接口与协议,并且可以通过接口函数访问到相关服务的属性。因此,本文和文献[1,4,5,7]一样选择第二种方式,即把 SIWS 的组合抽象为具有一个开始节点和结束节点的 DAG。文献[4]把空间信息服务分成 5 个动态阶段,提出了一种动态服务组合参考模型 DSCRM,并应用到了国家 863 的 SIG 项目中,其实验表明 Web 服务组合的原理在空间信息领域有一定的实用性和可行性。文献[5]介绍了美国 NASA 在 SIWS 方面的研究,通过对 SIWS 分类,建立相应的 OGCRIM 类的层次模型,建立了基于 BPLe 的 Web 服务组合的服务链。文献[1,7]则从 OGC 对 SIWS 约束规范的角度出发,从理论上探讨了 SIWS 组合一般性原则,提出了在 BPLe4WS 标准上讨论动态 GIS 服务工作流的动态选择机制。上述研究对本文基于本体和规则的动态 workflow 模型有一定的启发作用。

结束语 本文建立了基于空间领域本体的面向 SIWS 的工作流模型 DWF MOR,引入空间领域本体消除了 SIWS 服务在形成 workflow 时在语义上的冲突;结合业务规则的设置对 OWS2 提出的 Web 服务模型进行了改进,设计了相应的工作流引擎,动态形成了 SIWS 的工作流,增强了 SIWS 的灵活

性。但是,本模型还具有一定的局限性,主要表现在:其一, SIWS 是在一个本地服务器中提供的,目前还不能实现用户对远程空间数据的透明访问;其二,系统支持的并发用户数目有限,可以考虑在客户端利用 P2P 网络来减轻中央服务器的压力。下一步,本项目将对此展开进一步研究。

参考文献

- [1] 贾文珏,李斌,龚健雅. 基于 workflow 技术的动态 GIS 服务链研究[J]. 武汉大学学报, 2005, 30(11);982-985
- [2] Degree. <http://degree.sourceforge.net/> [DB/OL], 2007-07-09/2008-10-20
- [3] Zai J H, Eckhard K, Lars D. Intelligent binding in the engineering of automation systems using ontology and Web services[J]. IEEE Transaction on Systems, Man, and Cybernetics—Part C: Application and Reviews, 2003, 33;403-412
- [4] 刘书雷,刘云翔,唐宇,等. 基于 workflow 的空间信息服务动态聚合技术研究[J]. 计算机科学, 2007, 34(13);126-128
- [5] Di L, Zhao P, Yang W, et al. Ontology-driven Automatic Geospatial-Processing Modeling based on Web-service Chaining[C] //Proceedings of the Sixth Annual NASA Earth Science Technology Conference. College Park, MD, USA, June 2006
- [6] Samper J J, Tomás V R, Carrillo E, et al. Visualization of Ontologies to Specify Semantic Descriptions of Services [J]. IEEE Transaction on knowledge and data engineering, 2008, 20(1);130-134
- [7] Alameh, N. Chaining geographic information Web services[J]. Internet Computing, IEEE 2003(7);22-29
- [8] Protégé. <http://protege.stanford.edu/overview/po-screenshots.html> [DB/OL]. 2006-07-09/2008-10-22
- [9] W3C 组织. SPARQL Query Language for RDF. <http://www.w3.org/TR/rdf-sparql-query/> [DB/OL]. 2006-10-04/2009-01-20
- [10] Klien E, et al. Ontology-based discovery of geographic information services—An application in disaster management[J]. Computers, Environment and Urban Systems, 2006(30);102-123
- [11] HPJena 官方网站. <http://jena.sourceforge.net/> [DB/OL]. 2007-01-17/2009-02-12
- [12] 汤宪飞,蒋昌俊,丁志军,等. 基于 Petri 网的语义 Web 服务自动组合方法[J]. 软件学报, 2007, 18(12);2991-3000