领域知识形式背景规范化理论研究*

王 凯,朱文婕
(蚌埠医学院,安徽 蚌埠 233030)

【摘 要】领域知识形式背景是描述某个学科领域中概念及概念间关系的重要知识载体,在很大程度上决定了知识表达的规模和精确程度。本文针对临床诊断领域形式背景中知识表示的完备与精简问题,以高血压疾病为载体,综合考虑概念格内对象与属性间的不同层次关系,区别对待具有不同重要性的背景属性,提出形式背景属性约简理论与方法,给出消除背景冗余的II度规范理论,并在此基础上寻找到解决形式背景缺值的满值化方法。

【关键词】形式背景 约简 缺值
【中图分类号】TP18 【文献标识码】A 【文章编号】1673-1891(2014)01-0066-04

1 引言

概念格理论是由R.Wille于上世纪80年代提出的针对某一特定领域进行知识表述与分析的数学工具,体现了概念间泛化与特化的关系,生动地展示了数据集合中对象、属性间的二元关系。该理论在知识发现、数据分析与数据挖掘等方面均得到了广泛的应用。形式背景作为概念格理论的重要组成部分,是生成概念格结构的数据基础,其规模的大小和内容的准确度在很大程度上将决定最终格结构的计算效率,乃至整个格内节点信息的准确性。当一个形式背景过于庞大或包含许多重复信息时,通过其所表达的领域知识必然是低效的,甚至是冗余的。因此有必要研究形式背景的相关理论,提高知识表达的精确性,为后期大规模融合领域知识提供可靠的理论依据与解决方法。

文献[1]总结前人的研究成果,提出概念格的属性约简理论,寻找形式背景上的最小属性集,给出了形式背景上约简知识的表示方法;文献[2]定义了形式背景中的上、下近似算子,利用相似矩阵的相关性质,给出了基于相似矩阵的属性约简判定方法;文献[3]通过刻画属性中重要的指标,提出属性约简的判定定理;文献[4]给出形式背景中协调集的四类判定定理,丰富了协调集的判定,优化了形式背景的知识表达能力。

本文着重探讨形式背景的运算理论、约简理论以及缺值背景的满值化方法,从知识的精简表示角度优化了概念格理论。

2 形式背景基本概念

形式背景K=(G,M,I),其中G为对象集合,M为属性集合,I是G与M之间的一个二元关系。

2.1 并置与叠置

存在K1=(G1,M1,I1)、K2=(G2,M2,I2),令Gi={i}×Gi,Mi={i}×Mi,Ii={((i,g),(i,m))|(g,m)∈Ii},其中{i}×Gi,{i}×Mi是为了保证集合间相交为空集,并置为了解决形式背景间对象域相同,而属性域不相同时的形式背景连接问题;类似地,叠置是解决对象不同,而属性域相同的情况,故有如下定义:

1)若G1=G2=G,则K1与K2间的并置:K1|K2=(G,M1∪M2,I1∪I2);

2)若M1=M2=M,则K1与K2间的叠置:$\frac{K_1}{K_2}$=(G1∪G2,M,I1∪I2)

2.2 多值属性背景单值化

多值属性背景由四元组(G,M,W,I)表示,其中G与M的含义与上述情况保持一致,W为具体的属性值,I是由G、M和W间的三元关系。在实际的应用中,具体的属性是由其相应的值表示,例如某日天气是多云,而另日为晴,这必然会出现多值属性的情况。对于多值形式背景而言,必须先要将其转化成为单值背景,才能获得相应的形式概念。本文针对属性本身,采用概念缩放的方法,将具体的属性值转化为该概念的每个属性,用以解释相应的属性,其目的是将多元背景值转化为二元背景来表达,利用属性增加的手段来换取关系(对象与属性间的关系)的二元化。例如可以将多值背景表1按上述方法转化为二元背景表2。

表1 多值背景

Context : 多值形式背景	a	b	c
1	3.5	红	大
2	3.9	黄	小

表2 二元背景

	a_3.5	a_3.9	b_红	b_黄	c_大	c_小
1	X		X		X	
2		X		X		X

收稿日期:2013-10-16
*基金项目:蚌埠医学院科研项目(项目编号:BYKy1304)。
作者简介:王凯(1985-),男,安徽蚌埠人,硕士,助教,研究方向:概念格,医学知识库融合,本体等。

3 形式背景约简

3.1 形式背景简化

形式背景中蕴含着丰富的知识信息,降低形式背景的规模可以有效地提高概念格的生成效率,降低知识挖掘的复杂性。由文献[5]可知,简化的形式背景与原始背景是同构的,不会影响形式概念的质量,也不会降低其内涵与外延的数量。本文将形式背景简化的程度分为两个层次:I 度规范化和 II 度规范化,分别定义如下:

定义 1 当且仅当形式背景 K=(G,M,I) 满足如下条件,称之为 I 度规范:

1) $\forall g, h \in G$,若 $g' = h'$,必有 g=h;2) $\forall m,n \in M$,若 $m' = n'$,必有 m=n。

定义 2 当且仅当形式背景 K=(G,M,I) 满足如下条件,称之为 II 度规范:

1) 任意 $g \in G$,属性集 g' 不等于其他对象的属性交集;2) 任意 $g \in G$,属性集 $g' \neq \varnothing$

定义 1 主要是从对象及属性冗余的角度出发,合并形式背景中相同的对象和属性,使其达到基本的简洁;定义 2 从对象及属性相互表达的角度,去除所有能够用其他对象(属性)的交集来表达的外延(内涵),下面将举例来说明。

存在形式背景如表 3 所示,其中对象集 G1{Patient1,Patient 6}具有相同的属性集{Attr1,Attr2,Attr3,Attr4,Attr5 },属性集 M1{ Attr3,Attr 4}具有相同的对象集{ Patient1,Patient3,Patient4,Patient 6},为了实现 I 度规范,应将对象集 G1 合并为一行,属性集 M1 合并为一列,得到的 I 度规范形式背景如表 4 所示。对象{ Patient 5}所具有的属性{Attr2},是对象{ Patient 4}、{ Patient 7}对应属性集的交集,故可以将该对象删除,得到满足 II 度规范的形式背景,如表 5 所示。

表3　冗余形式背景

	Attr 1	Attr 2	Attr 3	Attr 4	Attr 5
Patient 1	×	×	×	×	×
Patient 2	×	×			
Patient3		×	×	×	×
Patient 4	×	×	×	×	
Patient 5		×			
Patient6	×	×	×	×	×
Patient7		×			×

表4　I 度规范形式背景

	Attr 1	Attr 2	Attr 3-Attr 4	Attr 5
Patient 1-Patient6	×	×	×	×
Patient 2	×	×		
Patient 3		×	×	×
Patient 4	×	×	×	
Patient 5		×		
Patient 7		×	×	×

表5　II 度规范形式背景

	Attr 1	Attr 2	Attr 3-Attr 4	Attr 5
Patient 1-Patient6	×	×	×	×
Patient 2	×	×		
Patient 3		×		×
Patient 4(5)	×	×	×	
Patient 7		×	×	

如表 3～5 所示,不难发现表 3 由于形式背景存在冗余,所形成的格结构如图 1 所示;化简后得到的 II 度规范形式背景,所形成的格结构如图 2 所示,二者的格结构具有同构性,形式背景的约简并未改变原有的概念特征,均蕴含 9 个概念集。

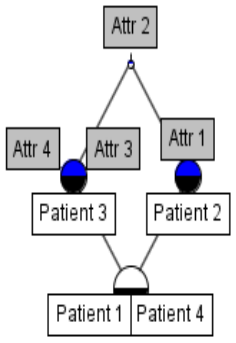

图1　与表3相对的概念格结构

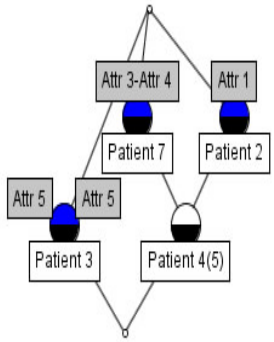

图2　与表5相对的概念格结构

3.2 形式背景的属性约简

由于形式背景是由对象、属性以及它们之间的多元关系组成,而属性的冗余是造成背景规模庞大的主要原因,故可以通过研究属性的约简来实现形式背景的精简。本文着重讨论属性值的约简。

从粗糙集合理论的角度看,属性约简即是在不改变原有知识分类能力的前提下,取出若干与原内容不相关或是关联程度不高的属性。属性约简可以有效地降低知识表示的规模,凸显有价值的信息,对提高形式背景的知识清晰度意义重大。在讨论本内容之前,为了便于说明,先给出相关定义。

定义 3 对于给定的知识库 K=(U, R),其中 U 为非空的对象集,称为论域;R 是 U 上的一族等价关系。若 $P \subseteq R$,且 P 不为空集,则 P 中所有等价关系的交集也是一个等价关系,称为 P 上的不可区分(indiscernibility)关系。对于每个子集 $X \subseteq U$ 和一个等价关系 R,给出如下定义:

$\underline{R}X = \cup\{Y \in U/R \mid Y \subseteq X\}$,$\overline{R}X = \cup\{Y \in U/R \mid Y \cap X \neq \varnothing\}$

定义 4 若 P 为一族等价关系,且 $R \in P$,IND(P-{R})=IND(P),则称关系 R 为 P 的不必要属性集。

定义 5 若 U 为论域,P 为定义在 U 上的等价关系,则称 P 中所有必要关系的集合为 P 的核心属性

集。

由定义3-5可知，对于存在依赖关系的关系族，存在不必要属性的部分可以进行约简，并且仅可以对该部分约简，不能够约简核心属性集。

定义6　对于形式背景K=(G，M，I)，任意的x，y∈G，称sim (x, y)为x,y的相似属性集，值为sim (x, y)=fA(x)∩fB(x)，所有对象的相似属性集的并称为该形式背景的相似属性矩阵。

形式背景K的约简集应为所有形式背景的属性约简集的并，又由核心属性集的定义可知，若a为核心属性则必有sim (x, y)={a}，故K的约简集∇可表示为：∇ = ∧{∨(ai : ai∈sim(x, y)}。在形式背景K=(G，M，I)中，对象集G={b1,b2,b3,b4,b5}，属性集M={a1,a2,a3,a4,a5}，关系I如表6所示。

表6　形式背景

	a1	a2	a3	a4	a5
b1	1	0	1	0	1
b2	0	1	0	0	0
b3	0	1	0	1	0
b4	0	0	0	1	0
b5	1	0	1	0	0

不为空的相似属性集分别为：

sim(b1,b1)={a1, a3, a5}；

sim(b1,b5)={a1,a3},sim(b2,b3)=sim(b2,b2)={a2}；

sim(b3,b3)={a2,a4}；

sim(b3,b4)=sim(b4,b4)={a4}；

sim(b5,b5)={a1,a3} 则∇ ={(a1∨ a3∨ a5)∧(a1∨a3)∧(a2)∧(a2∨a4)∧(a4)∧(a1∨ a3)}，进一步化简可得∇ =(a1∧ a2∧ a4)∨(a2∧ a3∧ a4)，至此将该形式背景化简为两个彼此独立的约简集{a1,a2,a4}与{a2,a3,a4}，取二者的交集可得核心属性集{a2,a4}。

4　形式背景满值化

形式概念分析理论所研究的对象具有确定的、精确的二元关系，是建立在等价关系上的数据结构。然而大多数时候由于对数据的理解或是采样方法的限制，造成数据之间存在很多的不确定性，论域中的知识不是由等价关系构成的，而是建立在一种削弱的二元关系上，使得某些对象与属性的关系呈现未知的状态。

处理缺值形式背景通常采用的方法是填补法。该方法将形式背景中未知的关系用0或1加以填充，考虑对象与属性间所有的可能，从而将其转化成完整的形式背景。虽然该方法操作起来简便易行，但存在着以下不足：容易带来知识的冗余和错误，填充后的形式背景虽然涵盖了其所有的可能，但其中含有许多无用的甚至于错误的信息，而且知识的结构过于庞大，不利于后期的挖掘。

针对于上述方法的缺点，本文着眼于缺值关系本身，从属性的角度，将其与对象的不确定性关系加以扩展，对缺值的属性按其在不同对象中的不同缺值分别进行扩充，最后得到完整的形式背景。该方法最大的好处是可以在不明显扩大形式背景规模的前提下，充分保持原有的背景信息。

对于形式背景K=(G，M，I)，扩展后的属性集Me={(m，1)|m∈M}∪{(m，*)|m∈M}，相应地，原有对象与新增属性集Me间的关系集Ie也需要做相应地调整，调整后的关系集Ie满足以下约束：1) (g，(m，1))∈Ie⇔(g，m)∈I and (g，(m，*))∈Ie，则调整后的形式背景是原背景的完备化扩展。

例如有初始样本背景如表7所示，显然其对象与属性间的关系存在缺值，如对象1该采用怎样的诊治方法存在未知。从本质上说，在将多值关系转化为二元关系后，如果假设所有的缺值背景都建立在二元的前提下，那么该问题就转化成 如何用一个二元的形式来表达一个三元关系。现将转化机理描述如下。

表7　缺值形式背景

Context：初始样本背景

	诊治方法	温度	湿度	分布情况	危害程度
1	\	高	较小	广泛	严重
2	常规	高	较大	很小	一般
3	\	高	较大	\	严重
4	复杂	低	较大	\	一般
5	常规	低	较小	广泛	严重

转化机理：

若(属性a, 1)值为1:表明该对象与属性a的二元关系为定值；

若(属性a, 1)值为0:

i) 如果(属性a, 1)与(属性a,*)取值相反，表明该对象与属性a的二元关系为缺值；

ii)如果(属性a, 1)与(属性a,*)取值相同，表明该对象与属性a的二元关系为定值，且其二元关系与i)中情况相反。

按上述思路，对于对象1的定值关系而言，则其属性值为‘常规’的值赋为1(表8中X代表关系值1，灰色方框代表关系值0)，而将值为‘复杂’的关系赋为0，但为了区别与缺值关系相混淆，将(诊治方法)分解成(诊治方法，1)和(诊治方法，*)，将其分别赋0。对于对象1的缺值关系而言，将(诊治方法，1)和

(诊治方法,*)分别赋0和1。这样就用一个二元的形式背景来表示复杂的三元关系。

表8　满值化形式背景

Context : 满值化形式背景

	(诊治方法,1)	(诊治方法,*)	温度	湿度	(分布情况,1)	(分布情况,*)	危害程度
1		X	X		X		X
2	X		X	X			
3		X	X	X		X	X
4				X		X	
5	X				X		X

5 结论与展望

本文针对形式背景在知识表达方面所存在的问题,分别从约简和缺值角度,提出形式背景的II度规范理论,从背景属性的重要性角度,进行属性的约简,并且将不完备的形式背景用简洁的形式满值化,以上研究对于下一步简化知识的表示与复用具有十分重要的意义。

注释及参考文献:

[1]张文修. 概念格的属性约简理论与方法[J]. 中国科学E辑信息科学, 2005, 35(6):628-639

[2]李同军, 张文修. 基于粗糙集的形式背景属性约简及属性特征[J].计算机科学, 2006, 33(9):178-180

[3]吕跃进, 李金海. 概念格属性约简的启发式算法[J].计算机工程与应用,2009, 45(2):154-157

[4]张东晓, 王国俊. 概念格属性约简的判定[J].计算机工程与应用,2007, 43(22):165-168

[5]QIU G F, CHEN J. Discriminative criteria for different kinds of attributes in knowledge discovery. Proceeding of the 11th World Congress of International Fuzzy Systems Association. Beijing: Tsinghua University Press, 2005:1387-1390

Research on the Formal Context and Related Theories Oriented Domain Knowledge of Hypertension Disease

WANG Kai, ZHU Wen-jie

(Bengbu Medical College, Bengbu, Anhui 233030)

Abstract: Formal context is the basis of concept lattice theory. And it largely determines the scale and accuracy of the expression of domain knowledge. In this paper, the knowledge of completeness and reduction in formal context is mainly discussed, jointly considering the different hierarchies relationship between objects and attributes in concept lattice. The attribute of different importance is handled differently. The theory and methods about the approximation reduction of formal context is proposed. The degree II of gauge theory dealing with the redundancy is also prospected, depending on which the way to solve the missing-value context is analyzed.

Key words: formal context; reduction; missing value