控制与决策
Control and Decision

多粒度形式背景的不确定性度量与最优粒度选择

李金海, 贺建君

引用本文:
李金海,贺建君. 多粒度形式背景的不确定性度量与最优粒度选择[J]. 控制与决策, 2022, 37(5): 1299−1308.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.1365

多粒度形式背景的不确定性度量与最优粒度选择

李金海†, 贺建君

(1. 昆明理工大学 数据科学研究中心，昆明 650500；2. 昆明理工大学 理学院，昆明 650500)

摘 要: 多粒度形式概念分析是数据挖掘与知识发现的重要工具,但现有的多粒度形式概念分析理论中并未提出选择最优形式背景的标准,这导致只能对多个单粒度形式背景逐一研究其知识发现问题,因此无法应对含有多个粒度属性的形式背景. 鉴于此,对多粒度形式背景的粒度树上的属性块进行组合,将信息熵作为组合形式背景优劣的判别标准以评价最优粒度选择的性能. 首先,基于粒度树提出广义介粒度剪枝形式背景,它既能实现属性块内部跨粒度组合,又能实现属性块之间跨层组合;其次,给出广义介粒度剪枝形式背景的信息熵,以评价广义介粒度剪枝形式背景的优劣,并设计出最优粒度选择算法;接着,利用信息熵度量多粒度剪枝类属性块和粒度树的重要性;最后,通过实验分析表明基于信息熵的最优粒度选择和粒度树重要性度量方法是有效的.

关键词: 多粒度形式背景；多粒度类属性块；粒计算；信息熵；剪枝形式背景；最优粒度选择

中图分类号: TP18 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.1365 **开放科学(资源服务)标识码(OSID):**

引用格式: 李金海, 贺建君. 多粒度形式背景的不确定性度量与最优粒度选择[J]. 控制与决策, 2022, 37(5): 1299-1308.

Uncertainty measurement and optimal granularity selection for multi-granularity formal context

LI Jin-hai†, HE Jian-jun

(1. Data Science Research Center，Kunming University of Science and Technology，Kunming 650500，China；2. Faculty of Science，Kunming University of Science and Technology，Kunming 650500，China)

Abstract: Multi-granularity formal concept analysis is an important tool for data mining and knowledge discovery. However, there is no standard to select an optimal formal context in the existing multi-granularity formal concept analysis theory, which leads to the fact that multiple single-granularity formal contexts have to be studied separately one by one for achieving the task of knowledge discovery, leaving the formal contexts with multi-granularity attributes unexplored. In this paper, how to combine attribute blocks of the granularity tree of a multi-granularity formal context is studied, and information entropy is used as a criterion to judge whether a combined formal context is good or not, so as to evaluate the performance of the obtained optimal granularity selection results. Firstly, based on a granularity tree, the notion of a generalized meso-granularity pruning formal context is proposed, which can realize not only inter-layer cross-granularity combination but also cross-layer combination of attribute blocks. Secondly, the information entropy of a generalized meso-granularity pruning formal context is defined to evaluate its advantages and disadvantages, and an optimal granularity selection algorithm is designed. Then, the information entropy is used to measure the importance of the multi-granularity pruning class-attribute block and granularity tree. Finally, experimental analysis shows the effectiveness of the proposed methods of optimal granularity selection and importance measurement of a granularity tree based on information entropy.

Keywords: multi-granularity formal context；multi-granularity class-attribute block；granular computing；information entropy；pruning formal context；optimal granularity selection

0 引 言

粒计算是智能系统的一种重要研究方法,其原理是将复杂问题划分、抽象为若干简单的问题. 正是基于这一有效原理,粒计算快速成为数据挖掘、知识表示与发现以及不确定性信息处理等领域的重要工具. 1979年,Zadeh[1]提出模糊信息粒度,随后对其

收稿日期: 2020-10-01；**录用日期:** 2021-03-03.
基金项目: 国家自然科学基金项目(11971211).
责任编委: 胡清华.
†通讯作者. E-mail: jhlixjtu@163.com.

进一步完善并给出模糊信息粒化的概念;直到20年后,Lin[2]才正式使用“粒计算”这一术语. 随着学者们对粒计算研究的深入探讨,已发展出各种粒计算模型,如张铃等[3]建立的商空间理论、Pawlak[4]提出的粗糙集、Wille[5]给出的概念格以及Yao[6]提出的三支决策等. 截至目前,粒计算广泛应用于大数据挖掘[7]、多粒度认知[8]、机器学习[9]和医学图像识别[10]等方面,已成为计算科学的研究热点之一. 在信息粒度计算过程中,粒结构的构造对粒计算结果影响较大. 近年来,基于形式概念分析的粒计算理论和方法及其应用逐渐得到完善[11-14]. 传统的形式概念分析通常研究单一且固定的粒度框架下的数据挖掘与知识表示问题[15],不能有效处理现实中复杂的多粒化问题[16]. 受粒计算中“多粒度”思想的启发[17-18],一些学者也开始从多层次、多维度的角度研究形式概念分析. 针对这一问题,相应提出了多粒度标记形式背景[19],它将多个单粒度形式背景进行并置而成. 然而,这样的多粒度形式背景虽然相比单粒度形式背景其数据处理能力和实用性有所提高,但仍不能有效满足实际问题对数据跨层粒化的要求. 为此,李金海等[20-21]将粒度属性分层、分块,又提出介粒度形式背景和广义介粒度形式背景,使得形式背景的数据跨层粒度组合能力明显提高.

在处理复杂问题时,寻找合适的粒度层以有效解决问题是人们普遍比较关心的事情. 类似地,对于多粒度组合得到的形式背景,其最优粒度选择也是一个重要问题. 为此,吴伟志等[22-23]、顾沈明等[24]提出了协调的不完备多粒度标记决策系统的最优粒度选择方法, She等[25]针对多粒度形式背景研究了概念格的构造与概念知识转移,郝晨等[26]讨论了多标记背景下基于粒标记规则的最优标记选择. 这些方法都是通过代数方法描述最优粒度,不利于算法的快速实现. 实际上,信息论中的熵可以用来度量系统的无序性. 受此启发,梁吉业等[27]利用信息熵研究了信息系统中信息的不确定性. 考虑到熵可将一些实际问题用数值的方式进行直观刻画,因此利用熵来定量地判断属性粒度组合的优劣也是一个值得研究的课题.

综上所述,本文利用信息熵衡量多粒度形式背景的信息不确定性程度,以此作为属性粒度最优组合选择的评判标准. 具体地,将多粒度形式背景下的广义介粒度形式背景转化到粒度树上进行研究,定义剪枝形式背景的信息熵,给出基于信息熵选择最优粒度的算法以及度量多粒度剪枝类属性块和粒度树的重要性公式,并通过实验表明基于信息熵的最优粒度选择和粒度树重要性度量方法均是有效的.

1 预备知识

定义1[4] 形如$C=(U,A)$表示信息系统,其中U表示对象的非空集合,A表示属性的非空集合,对于任意属性$a\in A$,存在一个对应值集V_a,满足$a:U\to V_a$. 在给定的信息系统$C=(U,A)$中,若对于任意属性$a\in A$,值集V_a不为空,则称其为完备信息系统.

注意到,本文所讨论的信息系统均是完备的.

定义2[5] 形如(U,M,I)表示形式背景,其中I表示对象集U与属性集M之间的二元关系. 若对于任意$x\in U$和$m\in M$,满足$(x,m)\in I$,则表示对象x拥有属性m,反之表示对象x不拥有属性m.

将无全为1的行或列也无全为0的行或列的形式背景称为是正则的. 下文讨论的形式背景均正则.

定义3[5] 设(U,M,I)为形式背景,对于$X\subseteq U,B\subseteq M$,记$X^{\Delta}=\{m\in M:\forall x\in X,(x,m)\in I\}$, $B^{\Delta}=\{x\in U:\forall m\in B,(x,m)\in I\}$. 如果$X^{\Delta}=B,B^{\Delta}=X$,则称序对$(X,B)$为形式概念,$X$为概念的外延,$B$为概念的内涵.

设(U,M,I)为形式背景,$N\subseteq M,I_N=I\bigcap(U\times N)$,称$(U,N,I_N)$为$(U,M,I)$的属性子背景[5]. 为了与形式背景$(U,M,I)$的情形进行区别,子背景$(U,N,I_N)$上的概念诱导算子定义为

$$X^{\Delta_N}=\{m\in N:\forall x\in X,(x,m)\in I_N\},$$

$$B^{\Delta_N}=\{x\in U:\forall m\in B,(x,m)\in I_N\}.$$

性质1 设(U,N,I_N)为(U,M,I)的子背景,$X\subseteq U$,则$X^{\Delta\Delta}\subseteq X^{\Delta_N\Delta_N}$.

定义4[19] 设(U,M,I)为形式背景,I_a表示属性a拥有的对象. 对于$B\subseteq M$,若$U=\bigcup\limits_{a\in B}I_a$,且$I_a(a\in B)$两两不相交,则称$B$为$(U,M,I)$的单粒度类属性块.

上述定义表明单粒度类属性块能构成论域U的划分. 如果一个形式背景的属性集M可划分成多个单粒度类属性块,则它是可类属性块分割的. 注意到,本文后续讨论的形式背景都是可类属性块分割的.

定义5[20] 设(U,M_1,I_1)和(U,M_2,I_2)为两个不同粒度的形式背景,其类属性块分别为$B_{11},B_{12},\ldots,B_{1s}$和$B_{21},B_{22},\ldots,B_{2s}$. 如果单粒度类属性块$B_{1k}$中的布尔属性合并可以产生$B_{2k}$,则称$B_{1k}$是$B_{2k}$的特化属性集,或$B_{2k}$是$B_{1k}$的泛化属性集,记作$B_{1k}\succ B_{2k}$. 若对于任意$k\in\{1,2,\ldots,s\},B_{1k}\succ B_{2k}$均成立,则记作$(U,M_1,I_1)\succ(U,M_2,I_2)$.

定义6[19] 设$(U,M_i,I_i)(i\in\{1,2,\ldots,n\})$为$n$个单粒度形式背景,$B_{i1},B_{i2},\ldots,B_{is}$是$M_i$的类属性块,

其中$B_{1k}, B_{2k}, \ldots, B_{nk}(k \in \{1,2,\ldots,s\})$是不同粒度下的同类别属性块. 若$(U,M_1,I_1) \prec (U,M_2,I_2) \prec \ldots \prec (U,M_n,I_n)$,则称$\pi = \bigcup_{i=1}^{n}(U,M_i,I_i)$为多粒度形式背景.

定义7[20]　设$\pi = \bigcup_{i=1}^{n}(U,M_i,I_i)$为多粒度形式背景,$B_{i1}, B_{i2}, \ldots, B_{is}$是$M_i$的类属性块. 令$M_{\text{meso}} = B_{n_1 1} \bigcup B_{n_2 2} \bigcup \ldots \bigcup B_{n_s s}$,其中$n_1, n_2, \ldots, n_s \in \{1,2,\ldots,n\}$,$I_{\text{meso}} \subseteq U \times M_{\text{meso}}$,称$(U, M_{\text{meso}}, I_{\text{meso}})$为介粒度形式背景.

介粒度形式背景通过对数据进行划分与重组,使得不同的粒度层下的类属性信息可以跨粒度组合,但是每个类属性块中的信息均来自同一个粒度层.

定义8[28]　设(U,M,I)为形式背景,对于$a \in M$,若属性a及其特化属性能够形成一棵树,则称其为a的粒度树,记作T_a.

在画粒度树时,节点的放置遵循一定的规律. 通常某个属性节点向下连接的分支节点是该属性的特化属性,而向上连接的属性节点是其泛化属性.

2　多粒度形式背景的粒结构

注意到属性及粒化属性分布在不同的粒度层下,所以将属性之间的特化与泛化关系放在粒度树上进行研究,可使属性之间的粒度层次关系一目了然.

例1　给定一个信息系统$C = (U,A)$如表1所示,其中$U = \{x_1, x_2, \ldots, x_6\}$表示参加全国大学生数学建模竞赛的6个参赛队,$A = \{a_1, a_2\}$的元素分别代表获得奖项的等级以及奖金(单位为元). a_1、a_2各对应一个多值集合,具体为V_{a_1} ={国家级一等奖,国家级二等奖,省级一等奖,省级二等奖,省级三等奖},$V_{a_2} = \{3\,000, 5\,000, 10\,000, 30\,000\}$.

表1　信息系统$C = (U,A)$

U	a_1	a_2
x_1	国家级一等奖	30 000
x_2	国家级二等奖	10 000
x_3	省级一等奖	5 000
x_4	省级二等奖	3 000
x_5	省级三等奖	3 000
x_6	省级三等奖	3 000

下面先将信息系统转换成多个单粒度形式背景,再将它们进行并置形成多粒度形式背景.

将信息系统中的多值属性a_1分成两个子类: “国家级”和“省级”,分别记为b_1、b_2;将多值属性a_2也分成两个子类: “收入超过2万”和“收入不超过2万”,记为b_3、b_4,那么$C = (U,A)$可转化为如表2所示的形式背景(U,M_1,I_1). 其中:数字1表示对象拥有该属性,数字0表示对象不拥有该属性. 此外,还可将获奖等级和奖金b_1、b_2、b_3、b_4进一步细分,将b_1细分为国家级一等奖、国家级二等奖,记作c_1、c_2;将b_2细分为省级一等奖、省级二等奖或三等奖,记作c_3、c_4;将b_3细分为3万,记作c_5;将b_4细分为10 000、5 000、3 000,记作c_6、c_7、c_8. 那么形式背景(U,M_1,I_1)转化成表3.

表2　形式背景(U,M_1,I_1)

U	b_1	b_2	b_3	b_4
x_1	1	0	1	0
x_2	1	0	0	1
x_3	0	1	0	1
x_4	0	1	0	1
x_5	0	1	0	1
x_6	0	1	0	1

表3　形式背景(U,M_2,I_2)

U	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
x_1	1	0	0	0	1	0	0	0
x_2	0	1	0	0	0	1	0	0
x_3	0	0	1	0	0	0	1	0
x_4	0	0	0	1	0	0	0	1
x_5	0	0	0	1	0	0	0	1
x_6	0	0	0	1	0	0	0	1

表2、表3中的单粒度形式背景都有获奖等级与奖金两个类属性块,区别在于不同的单粒度形式背景所包含的属性粒度粗细有所差异. 为了通过粒度树讨论多粒度形式背景中属性粒化结构关系,本文约定:粒度树中位于最顶端的节点为父节点(最粗的属性,来源于信息系统),其下的节点均为子节点(特化属性),父节点第1次向下分支形成的所有节点组合为粒度树的第2层,第2次分支形成的所有节点为第3层,依此类推.

注意到,尽管多粒度形式背景中每一个属性节点都可以形成一棵粒度树,但只有部分粒度树的研究有意义. 为了刻画属性节点间完整的结构关系,本文只讨论每个属性所在的最大粒度树,即以最粗的属性(来源于信息系统)作为最顶端的父节点的粒度树.

性质2　设$C = (U,A)$为信息系统,对于$a \in A$,T_a为信息系统对应的多粒度形式背景$\pi = \bigcup_{i=1}^{n}(U,M_i,I_i)$的一棵属性粒度树,多粒度形式背景$\pi$可形成$|A|$棵属性粒度树.

性质3　设$\pi = \bigcup_{i=1}^{n}(U,M_i,I_i)$为多粒度形式背景,且$(U,M_i,I_i)$含有$s$个类属性快,那么$\pi$可形成$s$棵$n+1$层的属性粒度树.

例2　画出例1中多粒度形式背景的所有属性粒度树.

针对表1～表3中的数据,有$Ib_1 \bigcup Ib_2 = Ib_3 \bigcup Ib_4 = U$,因此表2有两个类属性块,分别记为$M_{11} = \{b_1, b_2\}, M_{12} = \{b_3, b_4\}$. 同理,表3中也有两个类属性块,分别记为$M_{21} = \{c_1, c_2, c_3, c_4\}, M_{22} = \{c_5, c_6, c_7, c_8\}$. 此外,$Ib_1 = Ic_1 \bigcup Ic_2, Ib_2 = Ic_3 \bigcup Ic_4, Ib_3 = Ic_5, Ib_4 = \bigcup_{k=6}^{8} Ic_k$. 因此,$M_{11} \prec M_{21}$且$M_{12} \prec M_{22}$,即$(U, M_1, I_1)$与$(U, M_2, I_2)$之间形成了粒度粗细关系,所以表2、表3可并置形成多粒度形式背景$\pi = \bigcup_{i=1}^{2} (U, M_i, I_i)$.

由于π包含两个单粒度形式背景,且每个单粒度形式背景均有2个类属性块,有$s = 2, n + 1 = 3$,即π可形成2棵3层的属性粒度树,如图1和图2所示.

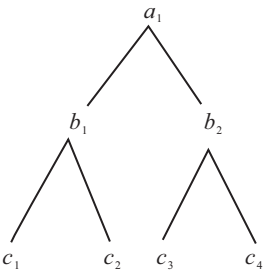

图1 获奖等级的属性粒度树

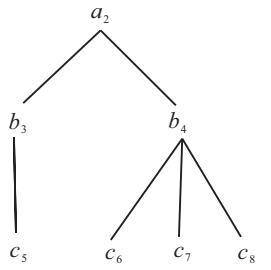

图2 奖金的属性粒度树

3 基于属性粒度树的剪枝形式背景

本节讨论如何通过属性粒度树的剪枝组合产生新的单粒度形式背景. 一般地,剪枝组合不能盲目进行,它需要遵循一定的规则且满足特定的语义解释. 下面主要讨论粒度树剪枝应遵循的规则.

传统的剪枝形式背景是对单个属性粒度树的同层进行剪枝,即每棵粒度树上剪取的节点均来自同一个粒度层. 然而,这种基于介粒度形式背景的剪枝方式对某些实际问题所需要的数据粒度组合要求无法满足. 比如,对于例2中获奖等级的属性粒度树,若学校规定以获奖等级的高低作为奖学金评定的一个指标,且要求只有获国家级奖的同学才能在奖学金评定中加分,则获省级奖不加分. 通常,国家级奖等级越高,加分越多. 那么,对国家级奖等级需要更细的描述信息,如到底是国家级一等奖还是国家级二等奖,而对省级奖则无需详细的描述,只要知道是否获得省级奖即可. 满足这个实际问题的数据粒度组合为$\{b_2, c_1, c_2\}$. 实际上,这种数据组合方式也形成一个类属性块,它是对获奖等级a_1的特定描述,由于这个属性块中的属性信息来自不同的粒度层,它是多粒度类属性块[21]. 根据多粒度类属性块中属性的特点,需要在一棵粒度树上剪取不同粒度层的节点来匹配多粒度类属性块的属性粒度组合要求.

此外,对多粒度类属性块的剪枝还有其他限制,虽然跨粒度层剪取属性节点进行组合可以大大提高数据匹配实际问题的能力,但是盲目的组合会增加计算量甚至组合出一些没有竞争优势的类属性块. 因此,属性进行分块时,如果分块太细,则会增加时间成本和信息获取代价;如果分块太粗,则不能很好地将数据与实际问题进行匹配. 一种合理的做法是,将属性粒度树上第2层包含的属性块个数作为分类的标准,这样既能提高数据的匹配能力又能降低各种成本. 本文约定:粒度树第2层的节点属性为子-父节点,子-父节点再往下的粒度属性按同层划分的规则将粒度树分割成多个子属性块. 如a_1的粒度树如图3所示,这里的d_1、d_2、d_3与c_1、c_2、c_3的语义对应相同,而d_4、d_5则是对属性c_4进一步划分得到的结果,它们分别表示“省级二等奖”“省级三等奖”. 按上述方法可将图3剪枝成6个子属性块:$\{b_1\}, \{b_2\}, \{c_1, c_2\}, \{c_3, c_4\}, \{d_1, d_2\}, \{d_3, d_4, d_5\}$.

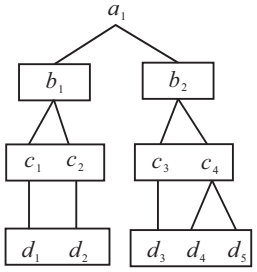

图3 子属性块划分示意图

需要指出的是,在多粒度形式背景的属性粒度树中,每棵粒度树上的剪枝结果(子属性块)必能组合出一个多粒度类属性块.

定义9 设$\pi = \bigcup_{i=1}^{n} (U, M_i, I_i)$可形成$s$棵$n+1$层的粒度树,令$B_{i1}, B_{i2}, \ldots, B_{is}$是$M_i$的类属性块. $B_{1k}, B_{2k}, \ldots, B_{nk} (k \in \{1, 2, \ldots, s\})$是第$k$棵粒度树上2至$n+1$层的属性,每个属性块包含的子属性块个数相同(记为$\gamma_k$), $B_{ik}^1, B_{ik}^2, \ldots, B_{ik}^{\gamma_k}$为$B_{ik}$的划分,若对于$k \in \{1, 2, \ldots, s\}, \{n_1, n_2, \ldots, n_{\gamma_k}\} \subseteq \{1, 2, \ldots, n\}, \{I_{n_1} b | b \in B_{n_1 k}^1\} \bigcup \{I_{n_2} b | b \in B_{n_2 k}^2\} \bigcup \ldots \bigcup \{I_{n_{\gamma_k}} b | b \in B_{n_{\gamma_k} k}^{\gamma_k}\}$构成$U$的划分,则称$D_k = \bigcup_{t=1}^{\gamma_k} B_{n_t k}^t$为第$k$棵粒度树上的一个多粒度剪枝类属

性块.

实际上,多粒度剪枝类属性块与现有的多粒度类属性块的分类机理和计算结果均相同,它只是借助粒度树得到多粒度类属性块而已,其目的是使得相关研究更加直观简洁.

例3　针对图2中奖金的属性粒度树,通过剪枝组合可形成4个多粒度剪枝类属性块,分别为$\{b_3,b_4\},\{b_3,c_6,c_7,c_8\},\{b_4,c_5\},\{c_5,c_6,c_7,c_8\}$.

设$D_k=\bigcup\limits_{t=1}^{\gamma_k}B^t_{n_tk}$和$H_k=\bigcup\limits_{t=1}^{\gamma_k}B^t_{n_t^*k}$为多粒度形式背景$\pi=\bigcup\limits_{i=1}^{n}(U,M_i,I_i)$第$k$棵粒度树上的两个多粒度剪枝类属性块. 若对于任意的$t\in\{1,2,\dots,\gamma_k\}$,有$n_t\leqslant n_t^*$,则称多粒度剪枝类属性块$D_k$比$H_k$的粒度粗,或$H_k$比$D_k$的粒度细,记作$D_k\prec H_k$.

定义10　设$\pi=\bigcup\limits_{i=1}^{n}(U,M_i,I_i)$可形成$s$棵$n+1$层的粒度树,令$B_{i1},B_{i2},\dots,B_{is}$是$M_i$的类属性块. $B_{1k},B_{2k},\dots,B_{nk}(k\in\{1,2,\dots,s\})$是第$k$棵粒度树上2至$n+1$层的属性,每个属性块包含的子属性块个数相同(记为$\gamma_k$),$B^1_{ik},B^2_{ik},\dots,B^{\gamma_k}_{ik}$为$B_{ik}$的划分,$D_k=\bigcup\limits_{t=1}^{\gamma_k}B^t_{n_tk}$为第$k$棵粒度树上的多粒度剪枝类属性块. 若$M_{\rm gmp}=\bigcup\limits_{k=1}^{s}D_k,I_{\rm gmp}\subseteq U\times M_{\rm gmp}$,则称$(U,M_{\rm gmp},I_{\rm gmp})$为$\pi$的一个广义介粒度剪枝形式背景.

本文约定:广义介粒度剪枝形式背景的粒度和等于各个多粒度剪枝类属性块的粒度之和,且多粒度剪枝类属性块的粒度又等于其子属性块的平均粒度. 因此,$(U,M_{\rm gmp},I_{\rm gmp})$的粒度和定义为

$$\mathrm{GS}(M_{\rm gmp})=\sum_{k=1}^{s}\frac{\sum\limits_{t=1}^{\gamma_k}n_t}{\gamma_k},$$

其中$\{n_1,n_2,\dots,n_{\gamma_k}\}\subseteq\{1,2,\dots,n\}$.

4　广义介粒度剪枝形式背景的信息熵

通常一个多粒度形式背景会包含多个广义介粒度剪枝形式背景,那么怎样衡量其好坏呢?如何选择最好的剪枝形式背景呢?本节将借助信息熵研究第1个问题,第2个问题在下一节讨论.

为了讨论方便,记广义介粒度剪枝形式背景$(U,M^\mu_{\rm gmp},I^\mu_{\rm gmp})$的所有对象概念为$\{(x^{\Delta_\mu\Delta_\mu},x^{\Delta_\mu})|x\in U\}$. 由于对象概念可以诱导出其余概念,其诱导公式为$(X,B)=\bigvee\limits_{x\in X}(x^{\Delta_\mu\Delta_\mu},x^{\Delta_\mu})$. 通过对象概念度量广义介粒度剪枝形式背景的信息量是合理的.

定义11　广义介粒度剪枝形式背景$(U,M^\mu_{\rm gmp},I^\mu_{\rm gmp})$的信息粒度定义为

$$\mathrm{IG}(M^\mu_{\rm gmp})=\frac{1}{|U|}\sum_{x\in U}\frac{|x^{\Delta_\mu\Delta_\mu}|}{|U|}.$$

信息粒度实际上是对信息粗细程度的一种刻画. 需要指出的是,不同的广义介粒度剪枝形式背景所包含的对象概念一般不同,但它们的信息粒度却可能相同,即两组数据不同但均值相同.

性质4　设$(U,M^\mu_{\rm gmp},I^\mu_{\rm gmp})$和$(U,M^\tau_{\rm gmp},I^\tau_{\rm gmp})$为多粒度形式背景$\pi=\bigcup\limits_{i=1}^{n}(U,M_i,I_i)$的两个广义介粒度剪枝形式背景,且$(U,M^\mu_{\rm gmp},I^\mu_{\rm gmp})\prec(U,M^\tau_{\rm gmp},I^\tau_{\rm gmp})$,那么$\mathrm{IG}(M^\tau_{\rm gmp})\leqslant\mathrm{IG}(M^\mu_{\rm gmp})$.

证明　因为$(U,M^\mu_{\rm gmp},I^\mu_{\rm gmp})\prec(U,M^\tau_{\rm gmp},I^\tau_{\rm gmp})$,所以$(U,M^\tau_{\rm gmp},I^\tau_{\rm gmp})$对象概念的外延均包含于$(U,M^\mu_{\rm gmp},I^\mu_{\rm gmp})$对象概念的外延中,因此$|x^{\Delta_\mu\Delta_\mu}|\geqslant|x^{\Delta_\tau\Delta_\tau}|(x\in U)$,根据定义11,有$\mathrm{IG}(M^\tau_{\rm gmp})\leqslant\mathrm{IG}(M^\mu_{\rm gmp})$. □

这意味着,属性粒度树剪取越细的分支组合,其信息粒度的值越小.

例4　计算表2中形式背景的信息粒度. 由于表2是由图1中粒度树T_{a_1}的第2层与图2中粒度树T_{a_2}的第2层进行属性组合得到的广义介粒度剪枝形式背景$(U,M^1_{\rm gmp},I^1_{\rm gmp})$,下式成立:

$$x_1^{\Delta_1\Delta_1}=\{x_1\},\ x_2^{\Delta_1\Delta_1}=\{x_2\},$$
$$x_3^{\Delta_1\Delta_1}=x_4^{\Delta_1\Delta_1}=x_5^{\Delta_1\Delta_1}=x_6^{\Delta_1\Delta_1}=\{x_3,x_4,x_5,x_6\}.$$

因此有

$$\mathrm{IG}(M^1_{\rm gmp})=\frac{1}{6}\Big(\frac{1}{6}+\frac{1}{6}+\frac{4}{6}+\frac{4}{6}+\frac{4}{6}+\frac{4}{6}\Big)=\frac{1}{2}.$$

定义12　广义介粒度剪枝形式背景$(U,M^\mu_{\rm gmp},I^\mu_{\rm gmp})$的信息熵定义为

$$\mathrm{IE}(M^\mu_{\rm gmp})=\frac{1}{|U|}\sum_{x\in U}\Big(1-\frac{|x^{\Delta_\mu\Delta_\mu}|}{|U|}\Big).$$

性质5　若(U,M^μ,I^μ)是$(U,M^\mu_{\rm gmp},I^\mu_{\rm gmp})$的子背景,则$\mathrm{IE}(M^\mu)\leqslant\mathrm{IE}(M^\mu_{\rm gmp})$.

证明　根据性质1可知$(U,M^\mu_{\rm gmp},I^\mu_{\rm gmp})$对象概念的外延均包含于$(U,M^\mu,I^\mu)$对象概念的外延中,因此$\mathrm{IE}(M^\mu)\leqslant\mathrm{IE}(M^\mu_{\rm gmp})$. □

上述性质表明,信息熵随着信息的增多而变大.

5　基于广义介粒度剪枝形式背景的最优粒度选择

由于多粒度形式背景中含有若干广义介粒度剪枝形式背景,而实际应用中通常是效益至上,追求满足其信息描述要求但代价最小的数据粒度组合. 为此,本文在信息熵的基础上结合“粒度和”研究最优粒度组合的选取问题. 这是因为数据的粒度和这一指标大致反映了广义介粒度剪枝形式背景的粒度粗

细情况,即数据的粒度和越小,某种程度上表明粒度越粗,其信息收集或采样所付出的代价会越小.

定义13 设$(U, M^{\mu}_{\rm gmp}, I^{\mu}_{\rm gmp})$和$(U, M^{\tau}_{\rm gmp}, I^{\tau}_{\rm gmp})$为多粒度形式背景$\pi = \bigcup\limits_{i=1}^{n}(U, M_i, I_i)$的两个广义介粒度剪枝形式背景,若${\rm IE}(M^{\mu}_{\rm gmp}) = {\rm IE}(M^{\tau}_{\rm gmp})$且${\rm GS}(M^{\mu}_{\rm gmp}) < {\rm GS}(M^{\tau}_{\rm gmp})$,则称$(U, M^{\mu}_{\rm gmp}, I_{\rm gmp}{}^{\mu})$的粒度组合优于$(U, M^{\tau}_{\rm gmp}, I^{\tau}_{\rm gmp})$.

这意味着,更优的广义介粒度剪枝形式背景具有相同的信息熵但粒度更粗. 换言之,更优的广义介粒度剪枝形式背景以更少的代价满足客户对信息熵提出的要求.

例5 以例1中的信息系统为例,在图1的基础上,粒度树T_{a_1}的剪枝方式共有4种: $B_{11} = \{b_1, b_2\}$, $B_{12} = \{b_1, c_3, c_4\}$, $B_{13} = \{c_1, c_2, b_2\}$, $B_{14} = \{c_1, c_2, c_3, c_4\}$. 同理,在图2的基础上,粒度树$T_{a_2}$的剪枝方式也有4种: $B_{21} = \{b_3, b_4\}$, $B_{22} = \{b_3, c_6, c_7, c_8\}$, $B_{23} = \{b_4, c_5\}$, $B_{24} = \{c_5, c_6, c_7, c_8\}$. 将它们进行两两组合可得16种广义介粒度剪枝形式背景,分别记为$S^{\mu} = (U, M^{\mu}_{\rm gmp}, I^{\mu}_{\rm gmp})$,如表4所示. 其中$\mu = 4(i-1)+j$,它反映了属性信息的具体组合情况,这里$i$是第1棵粒度树上子属性块的第2个下标变量,$j$是第2棵粒度树上子属性块第2个下标变量.

表4 例1的各广义介粒度剪枝背景的信息熵与粒度和

剪枝背景	信息熵	粒度和	剪枝背景	信息熵	粒度和
S^1	1/2	2	S^9	1/2	2.5
S^2	2/3	2.5	S^{10}	2/3	3
S^3	1/2	2.5	S^{11}	1/2	3
S^4	2/3	3	S^{12}	2/3	3.5
S^5	2/3	2.5	S^{13}	2/3	3
S^6	2/3	3	S^{14}	2/3	3.5
S^7	2/3	3	S^{15}	2/3	3.5
S^8	2/3	3.5	S^{16}	2/3	4

根据表4可知,在信息熵最大的情况下,广义介粒度剪枝形式背景$(U, M^{\mu}_{\rm gmp}, I^{\mu}_{\rm gmp})$$(\mu = 2, 5)$的属性组合粒度最粗. 由定义13可知,这2个广义介粒度剪枝形式背景均是最优的.

然而,面对大规模数据集时,一般需要启发式搜索策略以获得一个最优粒度组合,见算法1.

算法1 随机剪枝下的最优粒度选择.

输入: 多粒度形式背景$\pi = \bigcup\limits_{i=1}^{n}(U, M_i, I_i)$;

输出: 最优的广义介粒度剪枝形式背景.

step 1: 确定π中最粗的单粒度形式背景中属性块个数s,以此为基准将剩余的单粒度形式背景也分成s块,记为$B_{j1}, B_{j2}, \ldots, B_{js}, j \in \{1, 2, \ldots, n\}$.

step 2: 对于任意$k \in \{1, 2, \ldots, s\}$,令$\gamma_k = |B_{1k}|$,以确定子属性块的个数.

step 3: 对于任意$k \in \{1, 2, \ldots, s\}, j \in \{1, 2, \ldots, n\}$,将$B_{jk}$分成$\gamma_k$类,即$B^1_{jk}, B^2_{jk}, \ldots, B^{\gamma_k}_{jk}$.

step 4: 记$\max = {\rm IE}(M_n)$.

step 5: 令$(U, M_{\rm gmp}, I_{\rm gmp}) = (U, M_1, I_1)$.

step 6: 利用定义12计算${\rm IE}(M_{\rm gmp})$,若$\max = {\rm IE}(M_{\rm gmp})$,则结束算法,输出$(U, M_{\rm gmp}, I_{\rm gmp})$.

step 7: 令$Z = \varnothing$;

step 8: 令$k = 1$.

step 9: 令$t = 1$.

step 10: 令$M^{\mu}_{\rm gmp}$表示将$M_{\rm gmp}$的第k个多粒度剪枝类属性块D_k更新为$(D_k - B^t_{n_t k}) \bigcup B^t_{(n_t+1)k}$之后得到的结果.

step 11: 利用定义12计算${\rm IE}(M^{\mu}_{\rm gmp})$,若$\max = {\rm IE}(M^{\mu}_{\rm gmp})$,则结束算法,输出$(U, M^{\mu}_{\rm gmp}, I^{\mu}_{\rm gmp})$.

step 12: 执行$Z \leftarrow Z \bigcup {\rm IE}(M^{\mu}_{\rm gmp})$.

step 13: 若$t < \gamma_k$,则令$t = t + 1$,返回step 10.

step 14: 若$k < s$,则令$k = k + 1$,返回step 9.

step 15: 从Z中选出信息熵最大的广义介粒度剪枝形式背景$(U, M^{*}_{\rm gmp}, I^{*}_{\rm gmp})$,令$(U, M_{\rm gmp}, I_{\rm gmp}) = (U, M^{*}_{\rm gmp}, I^{*}_{\rm gmp})$,返回step 7.

注意到,在上述算法中,当多个子属性块下降后的信息熵相等时,随机选择一个子属性块进行下降. 在最坏的情况下,需要对各层的每个子属性块逐一进行下降,故核心步骤(step 6 ~ step 15)需要运行$ns\max\{\gamma_k\}$次. 由于计算一次信息熵的代价是$O(|U|^2\max\{|M_i|\})$,算法1的时间复杂度为$O(ns|U|^2\max\{\gamma_k|M_i|\})$.

6 多粒度剪枝类属性块和粒度树的重要性

文献[29]给出了一种属性重要性的度量方法,本节将其推广到多粒度形式背景中,以讨论多粒度剪枝类属性块和粒度树的重要性.

定义14 设$S^{\mu} = (U, M^{\mu}_{\rm gmp}, I^{\mu}_{\rm gmp})$为多粒度形式背景$\pi = \bigcup\limits_{i=1}^{n}(U, M_i, I_i)$的广义介粒度剪枝形式背景,$D^{\mu}_k (k \in \{1, 2, \ldots, s\})$是$S^{\mu}$的一个多粒度剪枝类属性块,$D^{\mu}_k$在$S^{\mu}$中的重要性定义为

$$ {\rm SIG}(M^{\mu}_{\rm gmp}|D^{\mu}_k) = {\rm IE}(M^{\mu}_{\rm gmp}) - {\rm IE}(M^{\mu}_{\rm gmp} - D^{\mu}_k). $$

定义14表明,多粒度剪枝类属性块相对于广义介粒度剪枝形式背景的重要性是通过移除它之后所引起的信息熵变化大小来衡量的. 如果重要性为零,则表明该多粒度剪枝类属性块是冗余的.

例6 计算例4中$(U, M^1_{\rm gmp}, I^1_{\rm gmp})$多粒度剪枝类属性块$D_2 = \{b_3, b_4\}$的重要性. 由例4可知,$(U, M^1_{\rm gmp}, I^1_{\rm gmp})$的信息熵为${\rm IE}(M^1_{\rm gmp}) = 1/2$,同时对于

$M^1_{\rm gmp}-D_2$,其信息熵为$\mathrm{IE}(M^1_{\rm gmp}-D_2)=4/9$,因此$\mathrm{SIG}(M^1_{\rm gmp}|D_2)=1/18$.

下面进一步讨论属性粒度树的重要性. 若在多粒度形式背景π中直接定义广义介粒度剪枝形式背景的粒度树的重要性,则需要求出含有该粒度树的所有广义介粒度剪枝形式背景,这是NP-hard问题,不容易解决. 因此,本文将在多粒度形式背景π的n个单粒度形式背景上定义属性粒度树的重要性.

定义15　设$\pi=\bigcup\limits_{i=1}^{n}(U,M_i,I_i)$为多粒度形式背景,且$\pi$可形成$s$棵$n+1$层的粒度树,那么第$k$棵粒度树$T_k$的重要性定义为

$$\mathrm{SIG}(T_k)=\frac{\sum\limits_{i=1}^{n}\mathrm{SIG}(M_i|B_{ik})}{n}.$$

不难发现,当粒度树上各个子属性块相对于各自的单粒度形式背景的重要性均为零时,整棵粒度树是冗余的.

例7　计算例2中粒度树T_{a_2}(见图2)的重要性. 由例2的计算结果和定义15可知,该粒度树的重要性为

$$\mathrm{SIG}(T_{a_2})=\frac{\mathrm{SIG}(M_1|M_{12})+\mathrm{SIG}(M_2|M_{22})}{2}=\frac{1}{36}.$$

它表明奖金这一属性的重要性不高.

算法2　属性粒度树的重要性.

输入:含有s棵$n+1$层粒度树的多粒度形式背景$\pi=\bigcup\limits_{i=1}^{n}(U,M_i,I_i)$;

输出:s个粒度树的重要性.

step 1:确定单粒度形式背景的类属性块B_{ik}.

step 2:计算每个单粒度形式背景(U,M_i,I_i)的信息熵$\mathrm{IE}(M_i)$.

step 3:计算子背景$(U,M_i-B_{ik},I_{M_i-B_{ik}})$的信息熵$\mathrm{IE}(M_i-B_{ik})$.

step 4:根据定义14和定义15计算粒度树$T_1,T_2,\ldots,T_s$的重要性.

算法2的时间复杂度为$O(ns|U|^2\max\{|M_i|\})$.

7　实验分析

本节通过实验评估算法1和算法2的性能,以表明广义介粒度剪枝方法相比其他粒度组合方法具有一定的优势,以及粒度树重要性度量方法是合理的.

7.1　实验环境及其相关说明

实验从UCI中选取6个数据集,分别为lris、Haberman's Survival、Ecoli、Balance Scale、Yeast、Tic-Tac-Toe Endgame[30]. 详细信息见表5.

由于6个数据集均以多值属性或连续属性的形式呈现,在对其进行多粒度数据分析之前,需要进行数据预处理,即将信息系统转化为形式背景. 具体步骤如下:先将原始数据转化为0-1布尔值的形式背景,这里的原始数据大致可分为2种类型:一是连续属性,二是多值属性. 针对第1种情况,根据数据分布特点设置合适的分段区间,如在[0,1]区间上连续取值的情形可将其分为$[0,0.2),[0.2,0.4),[0.4,0.6),[0.6,0.8),[0.8,1]$,每一个分段区间被重新看作一个属性(实验中针对不同的数据其分段方式有所差异);针对第2种情况,将多值属性的取值分别视作一个新的属性. 依照此方式,预处理后的属性个数如表6第3列所示. 需要指出的是,预处理后得到的数据作为最细粒度下的形式背景. 在此基础上,借鉴文献[20]中多粒度数据的形成方法,将数据集中相邻的布尔属性进行合并以产生多粒度形式背景. 为了更好地对比广义介粒度剪枝形式背景与介粒度形式背景之间的差异,本文将不同数据集由粗到细划分为4个粒度.

表5　实验数据集

数据集	对象个数	属性个数
lris	150	4
Haberman's Survival	306	3
Ecoli	336	8
Balance Scale	625	4
Yeast	1 484	8
Tic-Tac-Toe Endgame	958	9

表6　预处理后的数据集

数据集	对象个数	属性个数	类属性块个数
lris	150	14	4
Haberman's Survival	306	13	3
Ecoli	336	31	8
Balance Scale	625	20	4
Yeast	1 484	28	8
Tic-Tac-Toe Endgame	958	27	9

实验中,用$|L|$表示多粒度形式背景的粒度层数,将6个数据集转化成多粒度形式背景后,每个数据集可形成3种多粒度形式背景(取局部或整体的区别),即$|L|=2,|L|=3,|L|=4$. 在这3种情况下分别对比广义介粒度剪枝形式背景和介粒度形式背景的最优粒度组合情况.

此外,由于算法1计算最大信息熵得到的最优粒度组合可能不止一个,此时需要增加一个随机选择机制,这可能导致每次运行输出的最优粒度略有不同. 为此,本文选择重复实验10次,将其均值作为最优粒度组合的输出结果. 广义介粒度剪枝形式背景的最优粒度组合的粒度和依据公式

$$\mathrm{GS}(M_{\rm gmp})=\sum_{k=1}^{s}\frac{\sum\limits_{t=1}^{\gamma_k}n_t}{\gamma_k}$$

进行计算. 对于介粒度形式背景,由于其类属性块的特点是所有信息均来自同一粒度层,在计算其最优粒度组合的粒度和时,可直接选取广义介粒度剪枝形式背景中多粒度类属性块的最大粒度层得到介粒度形式背景的类属性块的粒度和. 因此,介粒度形式背景的最优粒度组合数的计算公式为

$$\mathrm{GS}(M_{\mathrm{meso}})=\sum_{k=1}^{s}\max\{n_1,n_2,\ldots,n_{\gamma_k}\}.$$

需要指出的是,表6中每个数据集有多少个类属性块,其数据集相应的便形成多少棵粒度树.

7.2 结果分析

对6个数据集的3种多粒度形式背景依次计算其广义介粒度剪枝形式背景(generalized meso-granularity pruning formal context, GMPFC)和介粒度形式背景(meso-granularity formal context, MFC)的最优粒度组合的粒度和. 为了表述方便,将Iris、Haberman's Survival、Ecoli、Balance Scale、Yeast、Tic-Tac-Toe Endgame预处理后得到的数据集分别命名为数据集1、2、3、4、5、6. 对比同一组数据在信息熵同样大(且最大)的情况下,广义介粒度剪枝形式背景与介粒度形式背景的最优粒度组合的粒度和. 表7给出了3种不同粒度层数下两种最优粒度组合方法的粒度和对比结果.

表7 MFC与GMPFC最优粒度组合的粒度和

数据集	$\|L\|=2$		$\|L\|=3$		$\|L\|=4$	
	MFC	GMPFC	MFC	GMPFC	MFC	GMPFC
1	6	5	8	6	12	9
2	4	4	6	5	11	9
3	11	10	16	11	32	27
4	8	6	12	8	16	16
5	13	11	15	14	33	30
6	10	10	22	18	35	32

由表7可以得到如下2个结论:

1) 信息熵相等的前提下,从广义介粒度剪枝形式背景中选出的最优粒度组合大多都优于从介粒度形式背景中选出的最优粒度组合,个别情况下两者的最优粒度组合的粒度和相同;

2) 随着粒度层数$|L|$的增大,介粒度形式背景的粒度和的增长速度要比广义介粒度剪枝形式背景的粒度和快.

另外,通过算法2,计算6个数据集的粒度树重要性. 需要指出的是,根据定义15计算出的粒度树重要性较小,不利于对比各粒度树的重要性差异. 为此,将多粒度形式背景所有属性树的重要性进行归一化. 这种处理方法是合理的,可以使粒度树的重要性在同一标准下进行对比,具体结果如表8～表10所示. 值得一提的是,表中横线"—"的作用是补齐表格各列的数值,原因是各个数据集的粒度树棵数不同,这里以粒度树棵数最大值(9棵,分别记为$T_1,T_2,\ldots,T_9$)制作表格. 此外,表8～表10还表明,第1个数据集在3种情况下各个粒度树的重要性均相等,出现这种情况与先前数据预处理和数据合并方式有关. 第3个数据集第4棵粒度树的重要性在3种情况下恒为0,这表明该粒度树上的类属性块存在冗余. 其他值反映了相应粒度树相对于多粒度形式背景的重要性程度.

表8 $|L|=2$时6个数据集的粒度树重要性

数据集	粒度树重要性/%								
	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
1	25.0	25.0	25.0	25.0	—	—	—	—	—
2	63.1	1.8	35.1	—	—	—	—	—	—
3	53.8	8.8	1.2	0.0	14.0	7.0	1.8	13.5	—
4	32.2	21.5	32.2	14.1	—	—	—	—	—
5	14.8	37.2	22.6	7.0	1.6	1.6	8.4	6.7	—
6	9.0	11.9	9.0	11.9	7.4	11.9	9.0	11.9	17.9

表9 $|L|=3$时6个数据集的粒度树重要性

数据集	粒度树重要性/%								
	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
1	25.0	25.0	25.0	25.0	—	—	—	—	—
2	67.8	2.0	30.1	—	—	—	—	—	—
3	57.7	8.6	0.6	0.0	12.2	7.4	1.2	12.3	—
4	35.8	22.0	28.8	13.5	—	—	—	—	—
5	17.7	32.1	20.9	6.5	1.6	1.3	7.9	12.0	—
6	8.5	10.6	8.5	10.6	8.5	12.8	10.6	12.8	17.0

表10 $|L|=4$时6个数据集的粒度树重要性

数据集	粒度树重要性/%								
	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
1	25.0	25.0	25.0	25.0	—	—	—	—	—
2	71.1	2.9	26.0	—	—	—	—	—	—
3	54.9	9.0	0.6	0.0	11.8	8.3	4.1	11.1	—
4	38.6	22.6	25.6	13.3	—	—	—	—	—
5	19.6	31.0	20.1	8.1	1.5	1.2	7.2	11.1	—
6	8.6	11.4	8.6	11.4	8.6	11.4	11.4	11.4	17.1

8 结 论

为了研究在何种粒度层下能够合理解决问题,本文基于信息熵提出数据重构的最优粒度选择方法. 创新之处是从属性粒度树的角度定义多粒度类

属性块,通过信息熵度量广义介粒度剪枝形式背景的粒度组合优劣,在此基础上讨论了多粒度剪枝类属性块与粒度树的重要性. 实验结果显示,广义介粒度剪枝形式背景的数据重构能力较强,匹配实际问题对数据粒度的需求比介粒度形式背景要好. 此外,实验结果与文献[21]得到的广义介粒度方法的粒度组合能力优于介粒度方法的结论是一致的,同时表明了基于信息熵的形式背景度量方法的合理性.

尽管基于信息熵的最优粒度选择方法为多粒度数据的粒度选择提供了新的分析工具,但是依然存在一些不足和需要进一步探讨的问题:

1) 本文利用启发式算法实现最优粒度组合的搜索,当算法出现多个子属性块下降后信息熵相等时,随机选择其中一个子属性块进行下降,这种随机设置方式很可能导致最优粒度组合结果不唯一,从而得到局部最优解而非全局最优解;

2) 为了应对NP-hard问题,在单粒度形式背景中给出了属性粒度树重要性的计算公式,而不是通过广义多粒度剪枝形式背景直接进行定义,这种处理方式的合理性与有效性有待评估;

3) 文中提出了多粒度剪枝类属性块的重要性度量方法,如何利用它继续研究多粒度数据的属性约简问题是一个有意义的课题;

4) 本文讨论的最优粒度选择问题是在假定各多粒度子属性块被选择几率均等的前提下进行的,即不带先验知识的数据粒度选择,但实际问题中客户可能对某些属性有特殊要求,对这种带先验知识的数据粒度选择仍需进一步考虑.

参考文献(References)

[1] Zadeh L A. Fuzzy sets and information granularity[M]. Advances in Fuzzy Systems — Applications and Theory. Amsterdam: World Scientific, 1996: 433-448.

[2] Lin T Y. Granular computing on binary relations I: Data mining and neighborhood systems[C]. Rough Sets in Knowledge Discovery. Heidelberg: Physica-Verlag, 1998: 107-121.

[3] 张铃, 张钹. 基于商空间的问题求解[M]. 北京: 清华大学出版社, 2014: 45-103.
(Zhang L, Zhang B. Quotient space based problem solving[M]. Beijing: Tsinghua University Press, 2014: 45-103.)

[4] Pawlak Z. Rough sets[J]. International Journal of Computer & Information Sciences, 1982, 11(5): 341-356.

[5] Wille R. Restructuring lattice theory: An approach based on hierarchies of concepts[C]. Ordered Sets. Dordrecht: Reidel, 1982: 445-470.

[6] Yao Y Y. Three-way decisions with probabilistic rough sets[J]. Information Sciences, 2010, 180(3): 341-353.

[7] 梁吉业, 钱宇华, 李德玉, 等. 大数据挖掘的粒计算理论与方法[J]. 中国科学: 信息科学, 2015, 45(11): 1355-1369.
(Liang J Y, Qian Y H, Li D Y, et al. Theory and method of granular computing for big data mining[J]. Scientia Sinica: Informationis, 2015, 45(11): 1355-1369.)

[8] 苗夺谦, 张清华, 钱宇华, 等. 从人类智能到机器实现模型 —— 粒计算理论与方法[J]. 智能系统学报, 2016, 11(6): 743-757.
(Miao D Q, Zhang Q H, Qian Y H, et al. From human intelligence to machine implementation model: Theories and applications based on granular computing[J]. CAAI Transactions on Intelligent Systems, 2016, 11(6): 743-757.)

[9] 陈德刚, 徐伟华, 李金海, 等. 粒计算基础教程[M]. 北京: 科学出版社, 2019: 117-123.
(Chen D G, Xu W H, Li J H, et al. Elements of granular computing[M]. Beijing: Science Press, 2019: 117-123.)

[10] 徐伟华, 李金海, 魏玲, 等. 形式概念分析理论与应用[M]. 北京: 科学出版社, 2016: 49-51.
(Xu W H, Li J H, Wei L, et al. Formal concept analysis: Theory and application[M]. Beijing: Science Press, 2016: 49-51.)

[11] 李金海, 魏玲, 张卓, 等. 概念格理论与方法及其研究展望[J]. 模式识别与人工智能, 2020, 33(7): 619-642.
(Li J H, Wei L, Zhang Z, et al. Concept lattice theory and method and their research prospect[J]. Pattern Recognition and Artificial Intelligence, 2020, 33(7): 619-642.)

[12] Yao Y Y. Interpreting concept learning in cognitive informatics and granular computing[J]. IEEE Transactions on Systems, Man, and Cybernetics — Part B: Cybernetics, 2009, 39(4): 855-866.

[13] Wei L, Wan Q. Granular transformation and irreducible element judgment theory based on pictorial diagrams[J]. IEEE Transactions on Cybernetics, 2016, 46(2): 380-387.

[14] Wu W Z, Leung Y, Mi J S. Granular computing and knowledge reduction in formal contexts[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(10): 1461-1474.

[15] 李美争, 王国胤. 三支近似概念格中基于对象-概念辨识矩阵的属性约简方法[J]. 控制与决策, 2016, 31(10): 1779-1784.
(Li M Z, Wang G Y. Object-concept discernibility matrix based approach to attribute reduction in three-way approximate concept lattice[J]. Control and Decision,

2016, 31(10): 1779-1784.)

[16] 张清华, 周玉兰, 滕海涛. 基于粒计算的认知模型[J]. 重庆邮电大学学报: 自然科学版, 2009, 21(4): 494-501.
(Zhang Q H, Zhou Y L, Teng H T. Cognition model based on granular computing[J]. Journal of Chongqing University of Posts and Telecommunications: Natural Science Edition, 2009, 21(4): 494-501.)

[17] Xie J P, Yang M H, Li J H, et al. Rule acquisition and optimal scale selection in multi-scale formal decision contexts and their applications to smart city[J]. Future Generation Computer Systems, 2018, 83: 564-581.

[18] Xu W H, Li W T, Zhang X T. Generalized multigranulation rough sets and optimal granularity selection[J]. Granular Computing, 2017, 2(4): 271-288.

[19] 李金海, 吴伟志, 邓硕. 形式概念分析的多粒度标记理论[J]. 山东大学学报: 理学版, 2019, 54(2): 30-40.
(Li J H, Wu W Z, Deng S. Multi-scale theory in formal concept analysis[J]. Journal of Shandong University: Natural Science, 2019, 54(2): 30-40.)

[20] 李金海, 李玉斐, 米允龙, 等. 多粒度形式概念分析的介粒度标记方法[J]. 计算机研究与发展, 2020, 57(2): 447-458.
(Li J H, Li Y F, Mi Y L, et al. Meso-granularity labeled method for multi-granularity formal concept analysis[J]. Journal of Computer Research and Development, 2020, 57(2): 447-458.)

[21] 李金海, 贺建君, 吴伟志. 多粒度形式概念分析的类属性块优化[J]. 山东大学学报: 理学版, 2020, 55(5): 1-12.
(Li J H, He J J, Wu W Z. Optimization of class-attribute block in multi-granularity formal concept analysis[J]. Journal of Shandong University: Natural Science, 2020, 55(5): 1-12.)

[22] 吴伟志, 陈颖, 徐优红, 等. 协调的不完备多粒度标记决策系统的最优粒度选择[J]. 模式识别与人工智能, 2016, 29(2): 108-115.
(Wu W Z, Chen Y, Xu Y H, et al. Optimal granularity selections in consistent incomplete multi-granular labeled decision systems[J]. Pattern Recognition and Artificial Intelligence, 2016, 29(2): 108-115.)

[23] 吴伟志, 杨丽, 谭安辉, 等. 广义不完备多粒度标记决策系统的粒度选择[J]. 计算机研究与发展, 2018, 55(6): 1263-1272.
(Wu W Z, Yang L, Tan A H, et al. Granularity selections in generalized incomplete multi-granular labeled decision systems[J]. Journal of Computer Research and Development, 2018, 55(6): 1263-1272.)

[24] 顾沈明, 顾金燕, 吴伟志, 等. 不完备多粒度决策系统的局部最优粒度选择[J]. 计算机研究与发展, 2017, 54(7): 1500-1509.
(Gu S M, Gu J Y, Wu W Z, et al. Local optimal granularity selections in incomplete multi-granular decision systems[J]. Journal of Computer Research and Development, 2017, 54(7): 1500-1509.)

[25] She Y H, He X L, Qian T, et al. A theoretical study on object-oriented and property-oriented multi-scale formal concept analysis[J]. International Journal of Machine Learning and Cybernetics, 2019, 10(11): 3263-3271.

[26] 郝晨, 范敏, 李金海, 等. 多标记背景下基于粒标记规则的最优标记选择[J]. 模式识别与人工智能, 2016, 29(3): 272-280.
(Hao C, Fan M, Li J H, et al. Optimal scale selection in multi-scale contexts based on granular scale rules[J]. Pattern Recognition and Artificial Intelligence, 2016, 29(3): 272-280.)

[27] 梁吉业, 钱宇华. 信息系统中的信息粒与熵理论[J]. 中国科学E辑: 信息科学, 2008, 38(12): 2048-2065.
(Liang J Y, Qian Y H. Information granularity and entropy theory in information system[J]. Science in China Series E: Information Sciences, 2008, 38(12): 2048-2065.)

[28] Belohlavek R, De Baets B, Konecny J. Granularity of attributes in formal concept analysis[J]. Information Sciences, 2014, 260: 149-170.

[29] Huang C C, Li J H, Dias S M. Attribute significance, consistency measure and attribute reduction in formal concept analysis[J]. Neural Network World, 2016, 26(6): 607-623.

[30] Lichman M. UCI machine learning repository[EB/OL]. (2020-07-22)[2020-09-01]. http://archive.ics.uci.edu/ml.

作者简介

李金海 (1984–), 男, 教授, 博士生导师, 从事大数据环境下的数据挖掘技术、概念认知学习、粒计算、智能系统分析与集成等研究, E-mail: jhlixjtu@163.com;

贺建君 (1992–), 女, 硕士生, 从事多粒度数据分析、粒计算与形式概念分析的研究, E-mail: jianjunhe2018@163.com.

(责任编辑: 郑晓蕾)