



广义多粒度决策形式背景的属性约简

王攀¹, 王欣怡², 李金海²

1. 昆明理工大学 信息工程与自动化学院, 云南 昆明 650500; 2. 昆明理工大学 理学院, 云南 昆明 650500

摘要 目前存在的大部分决策形式背景属性约简方法只针对单粒度决策属性进行研究, 较少考虑多粒度决策属性的情况。针对该问题, 给出了广义多粒度决策形式背景属性约简方法, 以实现信息系统属性约简, 即先对多粒度属性进行粒度选择, 再对冗余类属性块进行删除。为了刻画冗余类属性块, 在广义多粒度决策形式背景中引入条件信息熵, 以区分属性对于信息系统的重要程度, 继而提出属性约简方法及相应实现算法。实验部分先对数据采用本文方法进行约简, 再对约简后的数据进行分类预测, 并将其与其他算法的分类精度进行对比, 所得结果验证了本文提出的属性约简方法的有效性。

关键词 形式概念分析; 广义多粒度决策形式背景; 条件信息熵; 粒度选择; 属性约简

中图分类号: O29; TP18 **DOI**: 10.16152/j.cnki.xdxbr.2026-03-011

Attribute reduction in generalized multi-granularity formal decision contexts

WANG Pan¹, WANG Xinyi², LI Jinhai²

1. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China;

2. Faculty of Science, Kunming University of Science and Technology, Kunming 650500, China

Abstract Currently, most of the methods for attribute reduction in formal decision contexts only consider single-granularity decision attributes, rarely investigating multi-granularity decision attributes. Motivated by this, this paper presents a method for attribute reduction in generalized multi-granularity formal decision contexts to achieve attribute reduction in information systems. Specifically, it first selects granularity for multi-granularity attributes and then removes redundant class attribute blocks. In order to characterize redundant class attribute blocks, this paper introduces conditional information entropy in generalized multi-granularity formal decision contexts, thereby distinguishing the importance of attributes to the information system, and subsequently proposes an attribute reduction method as well as its corresponding implementation algorithm. In the experiments, the data are first reduced using our method, and then the reduced data are used for achieving classification tasks. Our method is further compared with other existing algorithms in terms of the classification accuracy. The obtained results validate the effectiveness of the proposed attribute reduction method.

Keywords formal concept analysis; generalized multi-granularity formal decision context; conditional information entropy; granularity selection; attribute reduction

收稿日期: 2026-01-15 修回日期: 2026-01-30

基金项目: 国家自然科学基金(12171388); 云南省基础研究计划杰出青年项目(202401AV070009)。

第一作者: 王攀, 男, 硕士研究生, 从事粒计算和数据挖掘研究, 1147691216@qq.com。

通信作者: 李金海, 男, 教授, 从事大数据分析、概念认知学习、智能系统分析与集成研究, jhlixjtu@163.com。

Wille 等提出形式概念分析^[1-2]的方法,该方法通过简化复杂数据将其转化为 0-1 数值表,以挖掘数据背后潜藏的关联关系。形式概念分析中的核心数据结构是概念格。随着几十年相关研究的不断深入,该理论已被广泛应用于规则挖掘和概念知识发现^[3-4],并与粒计算^[5-6]、多粒度^[7-9]及粗糙集^[10]等众多理论相辅相成。除了经典研究外,形式概念分析还被运用于新的方向和领域,包括机器学习^[11]、概念认知学习^[12]、数据挖掘^[13]等。

形式概念分析是一种用于数据分析和知识发现的方法,该方法可以完成对数据进行概念提取的操作,对于理解复杂系统和知识表示具有重要意义。在形式概念分析中,数据被表示为形式背景,包括对象和属性,以及它们之间的关联。通过在形式背景中定义概念,可以揭示出数据中存在的模式和关联。在该领域中,属性约简吸引了许多学者的关注。张文修等给出了概念格的属性约简理论以及协调集和约简的判定定理,并通过引入可辨识属性矩阵从而实现概念格的属性约简^[14]。魏玲等利用决策形式背景强弱协调性研究概念格属性约简问题^[15]。Wu 等提出了决策形式背景的粒规则挖掘方法,并基于可辨识矩阵及布尔函数,讨论了属性约简问题^[16]。李进金等利用交式可约元概念,得出属性约简个数的计算公式,完善了属性约简理论,并给出概念格生成算法^[17]。

Shannon 等提出信息熵,其目的是处理信息不确定性的度量问题^[18]。随后,学者将其运用在粗糙集中,以此对属性重要度和知识精度等问题进行研究。该工具为探讨知识约简提供了有益的手段^[19-22]。借助条件信息熵的原理,王国胤等研究了一种针对决策表进行知识简化的高效方法^[23]。同时,信息熵的应用在形式概念分析领域也得到了重视。Singh 等提出了形式概念分析的属性约简方法,该方法基于信息熵的加权概念格实现^[24]。李金海等将信息熵应用于多粒度决策形式背景,对属性约简的课题进行了进一步探讨^[25]。

在形式概念分析领域,目前的研究大多是在单粒度层下进行的,多粒度决策形式背景的属性约简也只是针对条件属性进行了多粒度层研究,并未将多粒度思想延伸至决策属性,这阻碍了通过删除冗余数据以提升分类质量的研究。鉴于此,本文通过广义多粒度决策形式背景属性约简

研究如何删除冗余属性块以显著提升分类器的性能。

1 相关知识

定义 1^[1] 设 $N = (K, V, L)$ 为形式背景, $K = \{k_1, k_2, \dots, k_z\}$ 为非空有限对象集, $V = \{v_1, v_2, \dots, v_q\}$ 为非空有限属性集, L 为笛卡尔积 $K \times V$ 上的二元关系。若 $(k, v) \in L$ 时,用 kLv 表示对象 k 具有属性 v 。

定义 2^[1] 在形式背景 $N = (K, V, L)$ 中,对于 $X \subseteq K$ 和 $C \subseteq V$,定义算子:

$$X^* = \{v \in V \mid \forall k \in X, kLv\} \quad (1)$$

$$C^* = \{k \in K \mid \forall v \in C, kLv\} \quad (2)$$

若 $X^* = C, C^* = X$,则称序对 (X, C) 为一个形式概念,简称概念,其中 C 表示概念的内涵, X 表示概念的外延。

定义 3^[2,16] 对于形式背景 (K, V, L) ,若 $W \subseteq V$,则称 (K, W, L_W) 为 (K, V, L) 的一个子背景,其中 $L_W = L \cap (K \times W)$ 。在子背景 (K, W, L_W) 上的算子定义为

$$X^{*W} = \{v \in W \mid \forall k \in X, kLv\}, \forall X \subseteq K,$$

$$C^{*W} = \{k \in K \mid \forall v \in C, kLv\}, \forall C \subseteq W.$$

可得 $X^{*W} = X^* \cap W$,特别地, $X^{*V} = X^*$ 。

性质 1^[16] 设形式背景 (K, W, L_W) 为 (K, V, L) 的子背景,则 $X^{**} \subseteq X^{*W^{*W}}$ 。

定义 4^[26] 设 (K, V, L) 与 (K, D, J) 为形式背景,且 $V \cap D = \emptyset$,则称 $F = (K, V, L, D, J)$ 为决策形式背景,其中 V 表示条件属性集, D 表示决策属性集, (K, V, L) 是 F 的条件子背景, (K, D, J) 是 F 的决策子背景。

例 1 表 1 是一个信息系统 $B = (K, V), K = \{k_1, k_2, \dots, k_6\}$ 代表 6 个排球体育生, $V = \{a_1, a_2, a_3, a_4, a_5\}$, a_1 表示身高, a_2 表示 100 m 短跑用时, a_3 表示弹跳高度, a_4 表示运动员级别, a_5 表示体校排球队免试资格。多值集合由属性 a_1, a_2, a_3, a_4, a_5 的取值形成。

本文将要构建的广义多粒度决策形式背景,需要先将表 1 进行转换,将多值属性的取值当作 0-1 布尔值属性,拆分得到不同的形式背景,且都是单粒度的。然后,进行合并操作,引入多粒度决策属性至形式背景中。

以下给出对表 1 信息系统进行转换获得决策形式背景的详细过程,其拆分结果见表 2,具体拆

分过程如下。将 a_1 拆分为两个子类,其中 b_1 表示身高大于或等于180 cm, b_2 表示身高小于180 cm;将 a_2 拆分为两个子类,其中 b_3 表示100 m短跑用时13 s及以下, b_4 表示100 m短跑用时超过13 s;将 a_3 拆分为两个子类,其中 b_5 表示弹跳高度3.4 m以上, b_6 表示弹跳高度低于或等于3.4 m; d_1 由 a_4 诱导得到,表示是否获得一级或二级运动员证书; d_2 与 a_5 相同,表示是否被体校排球队免试录取。需要指出的是,在表1的信息系统 B 中,由于 a_1 、 a_2 、 a_3 表示排球体育生自身身体条件和体测的数据信息,而 a_4 、 a_5 表示排球体育生的级别和免试资格,它们之间存在一定的逻辑推理关系,且找出具体的逻辑推理关系对后续选拔培养优秀的排球体育生能够提供有益的参考,因此,本文将 a_1 、 a_2 、 a_3 视作条件属性, a_4 、 a_5 视作决策属性。在表2中的0-1布尔属性值,数字1代表该对象符合所在列的条件,0表示不符合。另外,决策属性 $D_1 = \{d_1, d_2\}$,它代表第1决策粒度。

表1 信息系统 $B = (K, V)$

Tab. 1 Information system $B = (K, V)$					
K	a_1	a_2	a_3	a_4	a_5
k_1	186	11.3	3.07	二级运动员	免试
k_2	193	10.9	3.22	一级运动员	免试
k_3	178	11.9	3.01	一级运动员	无免试
k_4	177	12.7	2.58	二级运动员	无免试
k_5	189	13.2	3.43	一级运动员	免试
k_6	185	11.8	2.89	无运动员证书	免试

表2 决策形式背景 (K, V_1, L_1, D_1, J_1)

Tab. 2 Formal decision context (K, V_1, L_1, D_1, J_1)								
K	b_1	b_2	b_3	b_4	b_5	b_6	d_1	d_2
k_1	1	0	1	0	0	1	1	1
k_2	1	0	1	0	0	1	1	1
k_3	0	1	1	0	0	1	1	0
k_4	0	1	1	0	0	1	1	0
k_5	1	0	0	1	1	0	1	1
k_6	1	0	1	0	0	1	0	1

为了进一步得到第2决策粒度,将多值属性 d_1 拆分为两个子类,其中 q_1 表示是否获得一级运动员证书, q_2 表示是否获得二级运动员证书, q_3 与 d_2 相同,详见表3。

表3 决策形式背景 (K, V_1, L_1, D_2, J_2)

Tab. 3 Formal decision context (K, V_1, L_1, D_2, J_2)

K	b_1	b_2	b_3	b_4	b_5	b_6	q_1	q_2	q_3
k_1	1	0	1	0	0	1	0	1	1
k_2	1	0	1	0	0	1	1	0	1
k_3	0	1	1	0	0	1	1	0	0
k_4	0	1	1	0	0	1	0	1	0
k_5	1	0	0	1	1	0	1	0	1
k_6	1	0	1	0	0	1	0	0	1

综上所述,表2和表3两个不同的决策粒度组合在一起,形成了多粒度决策属性,是形成广义多粒度决策形式背景的必要条件。在后续讨论中,本文将省略第1决策粒度下的决策形式背景的数据,因为将决策属性 q_1 、 q_2 、 q_3 替换为 d_1 、 d_2 即可得到该部分数据。

定义5^[16] 设 $F = (K, V, L, D, J)$ 为决策形式背景,对于 $\forall k \in K$,若 $k^{*V*V} \subseteq k^{*D*D}$,则将它定义为协调决策形式背景。

定义6^[16] 设 $F = (K, V, L, D, J)$ 为决策形式背景,其中 W 是 V 的一个子集, $\forall k \in K$,如果 $k^{*W*W} \subseteq k^{*D*D}$,则称 W 为 F 的协调属性集; $\forall E \subset W$,若能找到至少一个 k 属于 K ,使得 $k^{*E*E} \not\subseteq k^{*D*D}$,则称 W 为 F 的一个约简。

2 广义多粒度决策形式背景的信息熵

本节对信息熵进行讨论,给出广义多粒度决策形式背景的条件信息熵。

2.1 广义多粒度决策形式背景

定义7^[27] 设 $N = (K, V, L)$ 为形式背景,且 $Q = \{v_1, v_2, \dots, v_q\} \subseteq V$,若对象集 Lv_i 表示属性 v_i ($i \in \{1, 2, \dots, q\}$)所包含的所有对象, Lv_i ($v_i \in Q$)构成 K 的划分,且 Q 中的布尔属性在语义上归属于同一类,则称 Q 为 (K, V, L) 的类属性块。

例2 对于表2中的条件子背景 $N = (K, V_1, L_1)$,其中 $\{Lb_1, Lb_2\}$ 是 K 的一个划分,则 $Q_{11} = \{b_1, b_2\}$ 是 (K, V_1, L_1) 的类属性块。

定义8^[8] 设 (K, V_1, L_1) 和 (K, V_2, L_2) 为形式背景,它们的属性集划分为 $Q_{11}, Q_{12}, \dots, Q_{1s}$ 和 $Q_{21}, Q_{22}, \dots, Q_{2s}$,若单粒度类属性块 Q_{2n} 的布尔属性由 Q_{1n} 合并而成,则 Q_{1n} 比 Q_{2n} 的粒度细,记为 $Q_{1n} < Q_{2n}$ 。如果 $\forall n \in \{1, 2, \dots, s\}$ 有 $Q_{1n} < Q_{2n}$,

则 (K, V_1, L_1) 比 (K, V_2, L_2) 的粒度细,记为 $(K, V_1, L_1) < (K, V_2, L_2)$ 。

定义 9^[8] 设 $(K, V_i, L_i) (i \in \{1, 2, \cdots, p\})$ 为 p 个单粒度形式背景,假设 V_i 的类属性块为 $Q_{i1}, Q_{i2}, \cdots, Q_{is}$,其中 $Q_{1n}, Q_{2n}, \cdots, Q_{pn} (n \in \{1, 2, \cdots, s\})$ 为不同粒度下的同类别类属性块,若 $Q_{pn} < Q_{(p-1)n} < \cdots < Q_{1n}$,则称

$$\Phi = \bigcup_{i=1}^p (K, V_i, L_i) \tag{3}$$

为多粒度形式背景。

定义 10^[28] 设 $\Phi = \bigcup_{i=1}^p (K, V_i, L_i)$ 为多粒度形式背景, $Q_{i1}, Q_{i2}, \cdots, Q_{is}$ 是 V_i 的划分, $Q_{1n}, Q_{2n}, \cdots, Q_{pn} (n \in \{1, 2, \cdots, s\})$ 属于同一类别且包含布尔属性的子类个数为 $\lambda_n, Q_{in}^1, Q_{in}^2, \cdots, Q_{in}^{\lambda_n}$ 都是 Q_{in} 的划分, $Q_n = \bigcup_{r=1}^{\lambda_n} Q_{p,n}^r$ 和 $P_n = \bigcup_{r=1}^{\lambda_n} Q_{p',n}^r$ 属于 Φ ,都在其第 n 类别下。若 $\forall k \in \{1, 2, \cdots, \lambda_n\}$,有 $p_r \leq p'_r$,则 P_n 比 Q_n 的粒度细,记为 $P_n < Q_n$ 。

定义 11^[29] 设 $\Phi = \bigcup_{i=1}^p (K, V_i, L_i)$ 和 $\phi = \bigcup_{j=1}^z (K, D_j, J_j)$ 为多粒度形式背景,则称

$$\Pi = (\bigcup_{i=1}^p (K, V_i, L_i), \bigcup_{j=1}^z (K, D_j, J_j)) \tag{4}$$

为广义多粒度决策形式背景。

若在广义多粒度决策形式背景 Π 中,存在某一粒度层,使 (K, V_i, L_i, D_j, J_j) 协调,则称 Π 是部分协调的广义多粒度决策形式背景(Partially Consistent Generalized Multi-granularity Decision Context, PCGMDC)。本文重点讨论此类数据集。

例 3 对表 3 中的属性,进行更细的划分。先把属性 b_1 细分成 2 个子属性, c_1 表示身高大于或等于 190 cm, c_2 表示身高大于 180 cm 且小于 190 cm;把属性 b_3 也细分成 2 个子属性, c_4 表示 100 m 短跑用时小于或等于 12 s, c_5 表示 100 m 短跑用时大于 12 s 且小于 13 s;把属性 b_6 也细分成 2 个子属性,其中 c_8 表示弹跳高度大于 3.2 m 且小于 3.4 m, c_9 表示弹跳高度低于或等于 3.2 m。通过上述细分操作后,表 3 可转化为表 4 所示的决策形式背景 (K, V_2, L_2, D_2, J_2) 。

令 $Q_{21} = \{c_1, c_2, c_3\}, Q_{22} = \{c_4, c_5, c_6\}, Q_{23} = \{c_7, c_8, c_9\}$,那么 Q_{21}, Q_{22}, Q_{23} 便构成了 V_2 的划分,其都是 (K, V_2, L_2) 的类属性块。在第 1 条件粒度层下 $Q_{12} = \{b_3, b_4\}$,根据定义 8 以及例 3 的讨论可知, Q_{12}, Q_{22} 属于同一类别,但是 Q_{12} 比 Q_{22} 的粒

度更粗。以相同的方式可以讨论其它的粒度粗细关系。

表 4 决策形式背景 (K, V_2, L_2, D_2, J_2)

Tab. 4 Formal decision context (K, V_2, L_2, D_2, J_2)

K	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	q_1	q_2	q_3
k_1	0	1	0	1	0	0	0	0	1	0	1	1
k_2	1	0	0	1	0	0	0	1	0	1	0	1
k_3	0	0	1	1	0	0	0	0	1	1	0	0
k_4	0	0	1	0	1	0	0	0	1	0	1	0
k_5	0	1	0	0	0	1	1	0	0	1	0	1
k_6	0	1	0	1	0	0	0	0	1	0	0	1

根据定义 11 和表 4 知, $(\bigcup_{i=1}^2 (K, V_i, L_i),$

$\bigcup_{j=1}^2 (K, D_j, J_j))$ 为广义多粒度决策形式背景,且是不协调的,因为在第 1 粒度层下存在 $k_1^* v_1^* v_1 \not\subseteq k_1^* d_j^* d_j$,在第 2 粒度层下也存在 $k_1^* v_2^* v_2 \not\subseteq k_1^* d_j^* d_j$ 。所以还要对广义多粒度决策形式背景进行更细的划分,才能得到所需的数据集。

2.2 广义多粒度决策形式背景的信息熵

文献[25]提出多粒度形式背景的信息熵,在此基础上,本文拓展至广义多粒度决策形式背景。

定义 12 在 $PCGMDC \Pi = (\bigcup_{i=1}^p (K, V_i, L_i), \bigcup_{j=1}^z (K, D_j, J_j))$ 中,其第 i 条件粒度层 $N_i = (K, V_i, L_i)$ 的信息熵定义为

$$IE(V_i) = \frac{1}{|K|} \sum_{k \in K} \left(1 - \frac{|k^* v_i^* v_i|}{|K|} \right) \tag{5}$$

性质 2 在 $PCGMDC \Pi = (\bigcup_{i=1}^p (K, V_i, L_i), \bigcup_{j=1}^z (K, D_j, J_j))$ 中,其信息熵大于它所包含子背景的信息熵。

证明 通过性质 1 可知广义多粒度决策形式背景 $\Pi = (\bigcup_{i=1}^p (K, V_i, L_i), \bigcup_{j=1}^z (K, D_j, J_j))$ 的对象概念的外延均包含于其子背景对象概念的外延中,因此,它的信息熵大于它所包含子背景的信息熵。

定义 13 在 $PCGMDC \Pi = (\bigcup_{i=1}^p (K, V_i, L_i), \bigcup_{j=1}^z (K, D_j, J_j))$ 中,其第 j 决策粒度层下的子背景 (K, D_j, J_j) 关于第 i 条件粒度层下的子背景 (K, V_i, L_i) 的条件信息熵定义为

$$CIE(D_j | V_i) = \frac{1}{|K|} \sum_{k \in K} \left(\frac{|k^* v_i^* v_i| - |k^* v_i^* v_i \cap k^* d_j^* d_j|}{|K|} \right) \tag{6}$$

性质3 在 $PCGMDC \Pi = (\bigcup_{i=1}^p (K, V_i, L_i),$

$\bigcup_{j=1}^z (K, D_j, J_j))$ 中,若 $W_i \subseteq V_i$,则 W_i 为第 i 条件粒度层和第 j 决策粒度层 (K, V_i, L_i, D_j, J_j) 的协调集当且仅当 $CIE(D_j | W_i) = 0$ 。

证明 充分性。如果 $CIE(D_j | W_i) = 0$,那么 $\forall k \in K$,有 $k^{*W_i * W_i} \subseteq k^{*D_j * D_j}$,由定义6知, W_i 是 (K, V_i, L_i, D_j, J_j) 的协调集。

必要性。已知 W_i 是 (K, V_i, L_i, D_j, J_j) 的协调集,因此决策形式背景 $(K, W_i, L_{W_i}, D_j, J_j)$ 是协调的,所以通过定义5可得, $\forall k \in K$,有 $k^{*W_i * W_i} \subseteq k^{*D_j * D_j}$,由定义13知 $CIE(D_j | W_i) = 0$ 。

3 广义多粒度决策形式背景属性约简

本节基于平均条件信息熵、最粗最细协调决策形式背景条件信息熵,对广义多粒度决策形式背景进行讨论,给出属性约简方法及其算法。

3.1 基于平均条件信息熵的属性约简

定义14 在 $PCGMDC \Pi = (\bigcup_{i=1}^p (K, V_i, L_i),$

$\bigcup_{j=1}^z (K, D_j, J_j))$ 中,对于第 j 决策粒度层,如果协调条件粒度层总共有 t_j 个,在每个协调条件粒度层下,条件信息熵为 $CIE(D_j | V_i) (i = 1, 2, \dots, t_j)$,那么协调粒度层平均条件信息熵定义为

$$CIE_{ave}(\sum_{j=1}^z D_j | \sum_{i=1}^{t_j} V_i) = \frac{1}{z} \sum_{j=1}^z \frac{1}{t_j} \sum_{i=1}^{t_j} CIE(D_j | V_i) \quad (7)$$

定义15 在 $PCGMDC \Pi = (\bigcup_{i=1}^p (K, V_i, L_i),$

$\bigcup_{j=1}^z (K, D_j, J_j))$ 中,对于第 j 决策粒度层,如果存在 t_j 个协调条件粒度层,在每一协调粒度层下都有 $W_i \subseteq V_i$,使协调粒度层平均条件信息熵满足

$$CIE_{ave}(\sum_{j=1}^z D_j | \sum_{i=1}^{t_j} W_i) = CIE_{ave}(\sum_{j=1}^z D_j | \sum_{i=1}^{t_j} V_i),$$

则称 $(W_1, W_2, \dots, W_{t_j})$ 为 Π 的协调粒度层协调集;如果 $\forall i \in \{1, 2, \dots, t_j\}$,存在 $E_i \subseteq W_i$,且有 $E_j \subset W_j$ 使得

$$CIE_{ave}(\sum_{j=1}^z D_j | \sum_{i=1}^{t_j} E_i) \neq$$

$$CIE_{ave}(\sum_{j=1}^z D_j | \sum_{i=1}^{t_j} V_i),$$

那么称 $(W_1, W_2, \dots, W_{t_j})$ 为 Π 的协调粒度层约简集。

值得强调的是,在约简 V_i 的过程中, W_i 相较于 $V_i (i = 1, 2, \dots, t_j)$ 删除的类属性块属于同一类别,这样才能保证不同粒度层下其删除的属性是一致的。同时,也说明了被删除的类属性块在不同粒度层下都是不必要的。

性质4 在 $PCGMDC \Pi = (\bigcup_{i=1}^p (K, V_i, L_i),$

$\bigcup_{j=1}^z (K, D_j, J_j))$ 中,若 $W_i \subseteq V_i$,则 $(W_1, W_2, \dots, W_{t_j})$ 为 Π 的协调粒度层协调集当且仅当在每一协调粒度层下, W_i 都是 (K, V_i, L_i, D_j, J_j) 的协调集。

证明 充分性。如果 $(W_1, W_2, \dots, W_{t_j})$ 是 Π 的协调粒度层协调集,则

$$\begin{aligned} CIE_{ave}(\sum_{j=1}^z D_j | \sum_{i=1}^{t_j} W_i) &= \\ \frac{1}{z} \sum_{j=1}^z \frac{1}{t_j} \sum_{i=1}^{t_j} CIE(D_j | W_i) &= \\ \frac{1}{z} \sum_{j=1}^z \frac{1}{t_j} \sum_{i=1}^{t_j} CIE(D_j | V_i) &= \\ CIE_{ave}(\sum_{j=1}^z D_j | \sum_{i=1}^{t_j} V_i) &= 0. \end{aligned}$$

因为条件信息熵 $CIE(D_j | W_i) \geq 0$,所以 $CIE(D_j | W_i) = 0$,通过性质3可知在每一粒度层下 W_i 都是 (K, V_i, L_i, D_j, J_j) 的协调集。

必要性。如果在每一粒度层下 W_i 都是 (K, V_i, L_i, D_j, J_j) 的协调集,那么通过性质3可知 $CIE(D_j | W_i) = 0$,所以

$$\begin{aligned} CIE_{ave}(\sum_{j=1}^z D_j | \sum_{i=1}^{t_j} W_i) &= \\ \frac{1}{z} \sum_{j=1}^z \frac{1}{t_j} \sum_{i=1}^{t_j} CIE(D_j | W_i) &= \\ \frac{1}{z} \sum_{j=1}^z \frac{1}{t_j} \sum_{i=1}^{t_j} CIE(D_j | V_i) &= \\ CIE_{ave}(\sum_{j=1}^z D_j | \sum_{i=1}^{t_j} V_i) &= 0, \end{aligned}$$

即 $(W_1, W_2, \dots, W_{t_j})$ 是 Π 的协调粒度层协调集。

定义16 在 $PCGMDC \Pi = (\bigcup_{i=1}^p (K, V_i, L_i),$

$\bigcup_{j=1}^z (K, D_j, J_j))$ 中,对于第 j 决策粒度层,如果协调条件粒度层个数有 t_j 个, $B_1, B_2, \dots, B_{t_j}$ 依次为其各协调条件粒度层下的属性真子集,那么在每一粒度层下 $Q_i \subseteq V_i - B_i$ 相对于 B_i 的外重要度,记作

$$SIG_{outer}^{V_i}(Q_i \mid B_i) =$$
$$CIE(\sum_{j=1}^z D_j \mid B_i) - CIE(\sum_{j=1}^z D_j \mid B_i \cup Q_i)。$$

所以在 Π 中,新增的属性集的协调粒度层平均外重要度定义为

$$\overline{SIG}_{outer}(\sum_{i=1}^{t_j}(Q_i \mid B_i)) = \frac{1}{t_j} \sum_{i=1}^{t_j} SIG_{outer}^{V_i}(Q_i \mid B_i)$$

(8)

定义 17 在 $PCGMDC \Pi = (\bigcup_{i=1}^p (K, V_i, L_i), \bigcup_{j=1}^z (K, D_j, J_j))$ 中, $V_1, V_2, \cdots, V_{t_j}$ 是各协调条件粒度层下全部属性集,若 $\forall Q_i \subset V_i$,则各粒度层下 Q_i 对于 V_i 内重要度记为

$$SIG_{inner}^{V_i}(V_i \mid Q_i) =$$
$$CIE(\sum_{j=1}^z D_j \mid (V_i - Q_i)) - CIE(\sum_{j=1}^z D_j \mid V_i)。$$

在 Π 中,被删除的属性集的协调粒度层平均内重要度记为

$$\overline{SIG}_{inner}(\sum_{i=1}^{t_j}(V_i \mid Q_i)) = \frac{1}{t_j} \sum_{i=1}^{t_j} SIG_{inner}^{V_i}(V_i \mid Q_i)$$

(9)

属性集的增减是以类属性块为单位的。若这些类属性块属于同一类,便促成了信息系统属性约简过程。为了方便,记 $Core(V_1, V_2, \cdots, V_{t_j})$ 为所有必要类属性块(在任何属性约简过程中都不能被删除的类属性块)组成的集合,这些类属性块在协调粒度层的平均内重要度均大于 0。

例 4 对表 4 中的属性进行更细的划分。首先,将属性 c_9 细分成 2 个子属性, e_9 表示弹跳高度小于 3.2 m 且大于 3.0 m, e_{10} 表示弹跳高度小于或等于 3.0 m。通过上述细分,表 4 可转化为表 5 所示的决策形式背景 (K, V_3, L_3, D_2, J_2) 。

由表 5 知, $(\bigcup_{i=1}^3 (K, V_i, L_i), \bigcup_{j=1}^2 (K, D_j, J_j))$ 为 $PCGMDC$,即在决策形式背景 (K, V_3, L_3, D_j, J_j) 的第 3 条件粒度层下是协调的。

表 5 决策形式背景 (K, V_3, L_3, D_2, J_2)

Tab. 5 Formal decision context (K, V_3, L_3, D_2, J_2)

K	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}	q_1	q_2	q_3
k_1	0	1	0	1	0	0	0	0	1	0	0	1	1
k_2	1	0	0	1	0	0	0	1	0	0	1	0	1
k_3	0	0	1	1	0	0	0	0	1	0	1	0	0
k_4	0	0	1	0	1	0	0	0	0	1	0	1	0
k_5	0	1	0	0	0	1	1	0	0	0	1	0	1
k_6	0	1	0	1	0	0	0	0	0	1	0	0	1

为了获取协调粒度层约简集,对表 5 的属性进一步拆分。将属性 e_{10} 细分成 2 个子属性, f_{10} 表示弹跳高度大于 2.8 m 且小于或等于 3.0 m, f_{11} 表示弹跳高度小于或等于 2.8 m。通过上述细分,表 5 可转化为表 6 所示的决策形式背景 (K, V_4, L_4, D_2, J_2) 。

表 6 决策形式背景 (K, V_4, L_4, D_2, J_2)

Tab. 6 Formal decision context (K, V_4, L_4, D_2, J_2)

K	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	q_1	q_2	q_3
k_1	0	1	0	1	0	0	0	0	1	0	0	0	1	1
k_2	1	0	0	1	0	0	0	1	0	0	0	1	0	1
k_3	0	0	1	1	0	0	0	0	1	0	0	1	0	0
k_4	0	0	1	0	1	0	0	0	0	0	1	0	1	0
k_5	0	1	0	0	0	1	1	0	0	0	0	1	0	1
k_6	0	1	0	1	0	0	0	0	0	1	0	0	0	1

可以看出,在决策形式背景 (K, V_4, L_4, D_j, J_j) 中的第 4 条件粒度层下它是协调的,即 $(\bigcup_{i=1}^4 (K, V_i, L_i), \bigcup_{j=1}^2 (K, D_j, J_j))$ 为 $PCGMDC$ 。

下面通过定义 13 的方法得到广义多粒度决策形式背景中各协调条件粒度层的条件信息熵,并获取约简集。未经过处理的决策形式背景记为 (K, V_i, L_i, D_j, J_j) ,因为在计算的过程中,需要通过删掉任一类属性块后计算其平均条件信息熵并产生约简集,所以去掉类属性块 Q_{ij} 后形成的决策形式背景记作 $(K, V_i^{Q_{xy}}, L_i^{Q_{xy}}, D_j, J_j)$,其中, x 代表不同条件粒度层, y 代表不同类属性块。

例 5 计算 $PCGMDC \Pi = (\bigcup_{i=1}^4 (K, V_i, L_i), \bigcup_{j=1}^2 (K, D_j, J_j))$ 的协调粒度层约简集。注意 Π 的第 3 和 4 条件粒度层是协调的,故 $CIE(D_j \mid V_3) = 0$, $CIE(D_j \mid V_4) = 0$ 。记决策子背景 $\bar{N} = (K, D_j, J_j)$,根据表 5 ~ 表 6 可得:

$$k_1^{*d_1 * d_1} = k_2^{*d_1 * d_1} = k_5^{*d_1 * d_1} = \{k_1, k_2, k_5\};$$
$$k_3^{*d_1 * d_1} = k_4^{*d_1 * d_1} = \{k_1, k_2, k_3, k_4, k_5\};$$
$$k_6^{*d_1 * d_1} = \{k_1, k_2, k_5, k_6\};$$
$$k_1^{*d_2 * d_2} = \{k_1\};$$
$$k_2^{*d_2 * d_2} = k_5^{*d_2 * d_2} = \{k_2, k_5\};$$
$$k_3^{*d_2 * d_2} = \{k_2, k_3, k_5\};$$
$$k_4^{*d_2 * d_2} = \{k_1, k_4\};$$
$$k_6^{*d_2 * d_2} = \{k_1, k_2, k_5, k_6\}。$$

删除第 3 条件粒度层下形式背景 $N_3 = (K,$

V_3, L_3) 中的类属性块 $Q_{31} = \{e_1, e_2, e_3\}$ 后, 条件子背景 $N'_3 = (K, V_3^{e_1e_2e_3}, L_3^{e_1e_2e_3})$ 满足:

$$\begin{aligned} k_1^* \vee_3^{e_1e_2e_3} * \vee_3^{e_1e_2e_3} &= k_3^* \vee_3^{e_1e_2e_3} * \vee_3^{e_1e_2e_3} = \{k_1, k_3\}; \\ k_2^* \vee_3^{e_1e_2e_3} * \vee_3^{e_1e_2e_3} &= \{k_2\}; \\ k_4^* \vee_3^{e_1e_2e_3} * \vee_3^{e_1e_2e_3} &= \{k_4\}; \\ k_5^* \vee_3^{e_1e_2e_3} * \vee_3^{e_1e_2e_3} &= \{k_5\}; \\ k_6^* \vee_3^{e_1e_2e_3} * \vee_3^{e_1e_2e_3} &= \{k_6\}. \end{aligned}$$

(K, D_1, J_1) 关于 $N'_3 = (K, V_3^{e_1e_2e_3}, L_3^{e_1e_2e_3})$ 的条件信息熵为

$$CIE(D_1 | (V_3 - Q_{31})) = \frac{1}{|K|}.$$

$$\sum_{k \in K} \left(\frac{|k^* \vee_3^{e_1e_2e_3} * \vee_3^{e_1e_2e_3}| - |k^* \vee_3^{e_1e_2e_3} * \vee_3^{e_1e_2e_3} \cap k^* D_1 * D_1|}{|K|} \right) = \frac{1}{36};$$

(K, D_2, J_2) 关于 $N'_3 = (K, V_3^{e_1e_2e_3}, L_3^{e_1e_2e_3})$ 的条件信息熵为

$$CIE(D_2 | (V_3 - Q_{31})) = \frac{1}{|K|}.$$

$$\sum_{k \in K} \left(\frac{|k^* \vee_3^{e_1e_2e_3} * \vee_3^{e_1e_2e_3}| - |k^* \vee_3^{e_1e_2e_3} * \vee_3^{e_1e_2e_3} \cap k^* D_2 * D_2|}{|K|} \right) = \frac{1}{18}.$$

同样的, 删除第4条件粒度层下形式背景 $N_4 = (K, V_4, L_4)$ 中的类属性块 $Q_{41} = \{f_1, f_2, f_3\}$ 后, 条件子背景 $N'_4 = (K, V_4^{f_1f_2f_3}, L_4^{f_1f_2f_3})$ 满足:

$$\begin{aligned} k_1^* \vee_4^{f_1f_2f_3} * \vee_4^{f_1f_2f_3} &= k_3^* \vee_4^{f_1f_2f_3} * \vee_4^{f_1f_2f_3} = \{k_1, k_3\}; \\ k_2^* \vee_4^{f_1f_2f_3} * \vee_4^{f_1f_2f_3} &= \{k_2\}; \\ k_4^* \vee_4^{f_1f_2f_3} * \vee_4^{f_1f_2f_3} &= \{k_4\}; \\ k_5^* \vee_4^{f_1f_2f_3} * \vee_4^{f_1f_2f_3} &= \{k_5\}; \\ k_6^* \vee_4^{f_1f_2f_3} * \vee_4^{f_1f_2f_3} &= \{k_6\}. \end{aligned}$$

(K, D_1, J_1) 关于条件子背景 $N'_4 = (K, V_4^{f_1f_2f_3}, L_4^{f_1f_2f_3})$ 的条件信息熵为

$$CIE(D_1 | (V_4 - Q_{41})) = \frac{1}{36}.$$

(K, D_2, J_2) 关于条件子背景 $N'_4 = (K, V_4^{f_1f_2f_3}, L_4^{f_1f_2f_3})$ 的条件信息熵为

$$CIE(D_2 | (V_4 - Q_{41})) = \frac{1}{18}.$$

所以, 在删除类属性块 Q_{31} 和 Q_{41} 后, 各决策子背景 $\bar{N}$ 关于条件子背景的平均条件信息熵为 $\frac{1}{24} > 0$, 故不能删除。

当去掉类属性块 $Q_{32} = \{e_4, e_5, e_6\}$ 后, 各决策

子背景 $\bar{N}$ 关于条件子背景 $N'_3 = (K, V_3^{e_4e_5e_6}, L_3^{e_4e_5e_6})$ 的信息粒满足:

$$\begin{aligned} k_1^* \vee_3^{e_4e_5e_6} * \vee_3^{e_4e_5e_6} &= \{k_1\}; \\ k_2^* \vee_3^{e_4e_5e_6} * \vee_3^{e_4e_5e_6} &= \{k_2\}; \\ k_3^* \vee_3^{e_4e_5e_6} * \vee_3^{e_4e_5e_6} &= \{k_3\}; \\ k_4^* \vee_3^{e_4e_5e_6} * \vee_3^{e_4e_5e_6} &= \{k_4\}; \\ k_5^* \vee_3^{e_4e_5e_6} * \vee_3^{e_4e_5e_6} &= \{k_5\}; \\ k_6^* \vee_3^{e_4e_5e_6} * \vee_3^{e_4e_5e_6} &= \{k_6\}. \end{aligned}$$

第1和第2决策粒度层下的决策子背景 $\bar{N}$ 关于条件子背景 $N'_3 = (K, V_3^{e_4e_5e_6}, L_3^{e_4e_5e_6})$ 的条件信息熵分别为

$$\begin{aligned} CIE(D_1 | (V_3 - Q_{32})) &= 0, \\ CIE(D_2 | (V_3 - Q_{32})) &= 0. \end{aligned}$$

删除类属性块 $Q_{42} = \{f_4, f_5, f_6\}$ 后, 第1和第2决策粒度层下的决策子背景 $\bar{N}$ 关于条件子背景 $N'_4 = (K, V_4^{f_4f_5f_6}, L_4^{f_4f_5f_6})$ 的条件信息熵分别为

$$\begin{aligned} CIE(D_1 | (V_4 - Q_{42})) &= 0, \\ CIE(D_2 | (V_4 - Q_{42})) &= 0. \end{aligned}$$

所以, 在删除类属性块 Q_{32} 和 Q_{42} 后, 各决策子背景 $\bar{N}$ 关于条件子背景的平均条件信息熵为0, 故能够删除。

删除类属性块 $Q_{33} = \{e_7, e_8, e_9, e_{10}\}$ 后, 第1和第2决策粒度下的决策子背景 $\bar{N}$ 关于条件子背景 $N'_3 = (K, V_3^{e_7e_8e_9e_{10}}, L_3^{e_7e_8e_9e_{10}})$ 的条件信息熵为

$$\begin{aligned} CIE(D_1 | (V_3 - Q_{33})) &\neq 0, \\ CIE(D_2 | (V_3 - Q_{33})) &\neq 0. \end{aligned}$$

所以, 在 Q_{33} 被删除后, 平均条件信息熵大于0, 故不能删除。

综上, 删除类属性块 Q_{32} 和 Q_{42} 不会影响平均条件信息熵的变化, 即 $(Q_{31} \cup Q_{33}, Q_{41} \cup Q_{43})$ 为 Π 的协调粒度层约简集。

3.2 最粗最细协调决策形式背景

最粗最细协调决策形式背景指所有条件粒度比此背景粗, 且所有决策粒度比此背景细的决策形式背景为不协调的。

定义 18 在 $PCGMDC \Pi = (\bigcup_{i=1}^p (K, V_i, L_i))$,

$\bigcup_{j=1}^z (K, D_j, J_j)$ 中, 若 $W_{\text{cor}} \subseteq V_{\text{cor}}$, 在最粗最细协调决策形式背景

$$F = (K, V_{\text{cor}}, L_{\text{cor}}, D_{\text{fin}}, J_{\text{fin}}) \quad (10)$$

中的条件信息熵满足

$$CIE(D_{\text{fin}} | W_{\text{cor}}) = CIE(D_{\text{fin}} | V_{\text{cor}}),$$

则 W_{cor} 为 Π 的最粗最细协调粒度层协调集; 若

$\forall E_{\text{cor}} \subset W_{\text{cor}}$, 有

$$CIE(D_{\text{fin}} | E_{\text{cor}}) \neq CIE(D_{\text{fin}} | V_{\text{cor}}),$$

则 W_{cor} 为广义多粒度决策形式背景 Π 的最粗最细协调粒度层约简集。

定义 19 $PCGMDC \Pi = (\bigcup_{i=1}^p (K, V_i, L_i),$

$\bigcup_{j=1}^z (K, D_j, J_j))$ 在最粗最细协调粒度层 $F = (K, V_{\text{cor}}, L_{\text{cor}}, D_{\text{fin}}, J_{\text{fin}})$ 中删除任一属性集 $Q \subseteq V_{\text{cor}}$ 后, 得到 $F' = (K, V_{\text{cor}}^Q, L_{\text{cor}}^Q, D_{\text{fin}}, J_{\text{fin}})$, 定义属性集 Q 在 Π 中的内重要度为

$$\begin{aligned} SIG_{\text{inner}}^{\text{cor}}(V_{\text{cor}} | Q) = \\ CIE(D_{\text{fin}} | (V_{\text{cor}} - Q)) - CIE(D_{\text{fin}} | V_{\text{cor}}) \end{aligned} \quad (11)$$

性质 5 在 $PCGMDC \Pi = (\bigcup_{i=1}^p (K, V_i, L_i),$

$\bigcup_{j=1}^z (K, D_j, J_j))$ 中, $\{Q_{\text{cor}}^1, Q_{\text{cor}}^2, \dots, Q_{\text{cor}}^s\}$ 为最粗最细协调决策形式背景下条件属性集 V_{cor} 的划分。令 W_{cor} 为 V_{cor} 的最粗最细协调粒度层约简集, 记为 $\bigcup_{n=1}^r \{Q_{\text{cor}}^n\}$ 。若提升决策粒度, 则不改变条件属性划分的情况下, 得到的决策形式背景依然是协调的。

3.3 属性约简算法

在本节中, 主要对约简算法进行讨论。

性质 6 在 $PCGMDC \Pi = (\bigcup_{i=1}^p (K, V_i, L_i),$

$\bigcup_{j=1}^z (K, D_j, J_j))$ 中, $F = (K, V_{\text{cor}}, L_{\text{cor}}, D_{\text{fin}}, J_{\text{fin}})$ 是最粗最细协调决策形式背景, $\forall Q \subseteq V_{\text{cor}}, Q$ 为 Π 的必要类属性块当且仅当 $SIG_{\text{inner}}^{\text{cor}}(V_{\text{cor}} | Q) > 0$ 。

证明 充分性。如果 Q 为 Π 的必要类属性块, 那么

$$SIG_{\text{inner}}^{\text{cor}}(V_{\text{cor}} | Q) > 0,$$

否则

$$SIG_{\text{inner}}^{\text{cor}}(V_{\text{cor}} | Q) = 0,$$

即

$$CIE(D_{\text{fin}} | (V_{\text{cor}} - Q)) - CIE(D_{\text{fin}} | V_{\text{cor}}) = 0。$$

因为 F_{cor} 是协调的, 所以

$$CIE(D_{\text{fin}} | V_{\text{cor}}) = 0,$$

那么

$$CIE(D_{\text{fin}} | (V_{\text{cor}} - Q)) = \frac{1}{|K|} \cdot$$

$$\sum_{k \in K} \left(\frac{|k^* V_{\text{cor}}^Q * V_{\text{cor}}^Q| - |k^* V_{\text{cor}}^Q * V_{\text{cor}}^Q \cap k^* D_{\text{fin}} * D_{\text{fin}}|}{|K|} \right) = 0。$$

$\forall k \in K$, 有 $k^* V_{\text{cor}}^Q * V_{\text{cor}}^Q \subseteq k^* D_{\text{fin}} * D_{\text{fin}}$, 根据定义 6 可得 $V_{\text{cor}} - Q$ 为 F_{cor} 的协调集。所以, 属性集 Q 不是必要

类属性块, 这与已知事实矛盾, 故

$$SIG_{\text{inner}}^{\text{cor}}(V_{\text{cor}} | Q) > 0。$$

必要性。当 $SIG_{\text{inner}}^{\text{cor}}(V_{\text{cor}} | Q) > 0$ 时, 可得

$$CIE(D_{\text{fin}} | (V_{\text{cor}} - Q)) - CIE(D_{\text{fin}} | V_{\text{cor}}) > 0。$$

由于 F_{cor} 在最粗最细粒度层下的决策形式背景下是协调的, 因此

$$CIE(D_{\text{fin}} | V_{\text{cor}}) = 0,$$

所以

$$CIE(D_{\text{fin}} | (V_{\text{cor}} - Q)) = \frac{1}{|K|} \cdot$$

$$\sum_{k \in K} \left(\frac{|k^* V_{\text{cor}}^Q * V_{\text{cor}}^Q| - |k^* V_{\text{cor}}^Q * V_{\text{cor}}^Q \cap k^* D_{\text{fin}} * D_{\text{fin}}|}{|K|} \right) > 0。$$

可以看出, 存在某个对象 k_0 , 使得

$$|k_0^* V_{\text{cor}}^Q * V_{\text{cor}}^Q| - |k_0^* V_{\text{cor}}^Q * V_{\text{cor}}^Q \cap k_0^* D_{\text{fin}} * D_{\text{fin}}| > 0, \text{ 那}$$

么, $k_0^* V_{\text{cor}}^Q * V_{\text{cor}}^Q \not\subseteq k_0^* D_{\text{fin}} * D_{\text{fin}}$ 。

因此, $V_{\text{cor}} - Q$ 不是 F_{cor} 的协调集。换言之, 把属性集 Q 去掉后对 F_{cor} 的协调性产生了影响, 所以 Q 是必要类属性块。

定义 20 在 $PCGMDC \Pi = (\bigcup_{i=1}^p (K, V_i, L_i),$

$\bigcup_{j=1}^z (K, D_j, J_j))$ 中, 若 $W_j \subset V_j$, 则 $Q_j \subseteq V_j - W_j$ 相对于 W_j 的最粗最细协调粒度层外重要度定义为

$$\begin{aligned} SIG_{\text{outer}}^{\text{cor}}(Q_j | W_j) = \\ CIE(D_{\text{fin}} | W_j) - CIE(D_{\text{fin}} | W_j \cup Q_j) \end{aligned} \quad (12)$$

对于基于平均条件信息熵的属性约简方法, 只经过例子中说明的这几个步骤得到的 W_i (约简集) 一般不是最终的约简集, 需要再计算其余类属性块的外重要度, 将外重要度最大的加入 W_i , 直到 W_i 为协调粒度层协调集; 最后再计算内重要度去除 W_i 中的冗余类属性块, 得到最终的协调粒度层约简集 W_i 。最粗最细协调决策形式背景的约简只考虑单一粒度, 所以没有引入平均条件信息熵, 直接计算内重要度得到决策形式背景的所有必要类属性块, 此时得到的约简集 W_{cor} 一般也不是最终的约简集, 需要再利用内外重要度进行动态调整, 后续步骤与基于平均条件信息熵的属性约简方法类似。为了进一步理解其语义, 下文分成 2 种算法以反映本文所提出方法的属性约简过程, 具体如下。

算法 1 基于协调粒度层平均条件信息熵的约简

输入 Π 以及它的协调粒度层

$$(\bigcup_{i=1}^p (K, V_i, L_i), \bigcup_{j=1}^z (K, D_j, J_j));$$

输出 约简集 $(W_1, W_2, \dots, W_{t_j})$

step 1 $(W_1, W_2, \dots, W_{t_j}) = (\emptyset, \emptyset, \dots, \emptyset)$;

step 2 对 H 中各协调粒度层的平均条件信息熵进行计算

$$CIE_{ave}(\sum_{j=1}^z D_j | \sum_{i=1}^{t_j} V_i);$$

step 3 遍历 V_i , 删除 $Q_{1j}, Q_{2j}, \dots, Q_{t_j}$ 后, 计算平均条件信息熵

$$CIE_{ave}(\sum_{j=1}^z D_j | \sum_{i=1}^{t_j} (V_i - Q_{ij}));$$

step 4 若

$$CIE_{ave}(\sum_{j=1}^z D_j | \sum_{i=1}^{t_j} (V_i - Q_{ij})) \neq$$

$$CIE_{ave}(\sum_{j=1}^z D_j | \sum_{i=1}^{t_j} (V_i)),$$

则 $W_i \leftarrow W_i \cup Q_{ij}$;

step 5 若 j 小于类属性块的数量, 返回 step 3; 若大于则转到 step 7。

step 6 如果 $Q_{ij} \subseteq V_i - W_i$ 且

$$\overline{SIG}_{outer}(\sum_{i=1}^{t_j} (Q_{ih} | W_i)) =$$

$$\max_{Q_{ij} \subseteq V_i - W_i} \{\overline{SIG}_{outer}(\sum_{i=1}^{t_j} (Q_{ij} | W_i))\},$$

那么, $W_i \leftarrow W_i \cup Q_{ih}$;

step 7

$$CIE_{ave}(\sum_{j=1}^z D_j | \sum_{i=1}^{t_j} W_i) \neq$$

$$CIE_{ave}(\sum_{j=1}^z D_j | \sum_{i=1}^{t_j} V_i),$$

返回 step 6, 直到

$$CIE_{ave}(\sum_{j=1}^z D_j | \sum_{i=1}^{t_j} W_i) =$$

$$CIE_{ave}(\sum_{j=1}^z D_j | \sum_{i=1}^{t_j} V_i)$$

停止;

step 8 遍历 $(W_1, W_2, \dots, W_{t_j})$, 计算 Q_{ij} 的内重要度;

step 9 如果

$$\overline{SIG}_{inner}(\sum_{i=1}^{t_j} (W_i | Q_{ij})) = 0,$$

那么 $W_i \leftarrow W_i - Q_{ij}$;

step 10 输出约简集 $(W_1, W_2, \dots, W_{t_j})$ 。

算法 2 最粗最细协调决策形式背景的约简

输入 广义多粒度决策形式背景中包含的最粗最细协调决策形式背景

$$F = (K, V_{cor}, L_{cor}, D_{fin}, J_{fin});$$

输出 约简集 W_{cor}

step 1 $Core_{cor}(V_{cor}) = \emptyset$;

step 2 对属性集 V_{cor} 中的每一类属性块 Q_j , 执行遍历以算出各自的内重要度;

step 3 如果 $SIG_{inner}^{cor}(V_{cor} | Q_j) > 0$, 那么,

$$Core_{cor}(V_{cor}) \leftarrow Core_{cor}(V_{cor}) \cup Q_j;$$

step 4 设 $V_{cor} - Core_{cor}(V_{cor})$ 中包含的类属性块为 $Q_1, Q_2, \dots, Q_r$, 令 $W_{cor} = Core_{cor}(V_{cor})$;

step 5 若

$$SIG_{outer}^{cor}(Q_h | W_{cor}) =$$

$$\max_{Q_j \subseteq V_{cor} - Core_{cor}(V_{cor})} \{SIG_{outer}^{cor}(Q_j | W_{cor})\},$$

则 $W_{cor} \leftarrow W_{cor} \cup Q_h$;

step 6 若

$$CIE(D_{fin} | W_{cor}) \neq CIE(D_{fin} | V_{cor}),$$

返回 step 5, 直到

$$CIE(D_{fin} | W_{cor}) = CIE(D_{fin} | V_{cor})$$

停止;

step 7 对 W_{cor} 进行遍历, 对其每一类属性块 Q_j 的内重要度进行计算;

step 8 如果 $SIG_{inner}^{cor}(W_{cor} | Q_j) = 0$, 那么, $W_{cor} \leftarrow W_{cor} - Q_j$;

step 9 输出 $W_{cor} \circ$

4 实验及结果分析

4.1 数据集

在本节中, 通过实验及对比分析验证在删除冗余类属性块后是否会显著提升分类器的性能。选取 UCI (<https://archive.ics.uci.edu/>) 数据库中的 5 个真实公开数据集, 分别为 Sensor、Obsity、Credit、Pop 以及 Wpbc, 它们的具体信息如表 7 所示。

一般情况下, 选取的数据集并不会是标准的 0-1 布尔值, 需要对这 5 个数据集进行相应的预处理。因此, 这里的首要任务是构建部分协调的广义多粒度决策形式背景。这一过程需要先把原始数据集全部转成形式背景, 在进行实验之前, 根据 5 个不同数据集的特点将其属性值转变为 0-1 布尔值, 以满足要求。为了确保所得形式背景的协调性, 在原始数据集中删除部分影响形式背景协调性的对象。然后将上述处理后的数据集转

变为 0-1 布尔值,详细处理过程如下:①如果数据集所包含的属性取值为多值属性,把这些多值属性的取值当作 0-1 布尔值属性,具体见本文所给出的例子,并且通过合适的处理方式把相同的类属性块所包含的相似属性进行合并;②如果数据集所包含的属性取值是连续型时,把连续型属性分成多个区间,也就是对数据集属性进行分段,此时可以把这些区间分完后的属性视作 0-1 布尔值属性。比如,把区间[10, 100]当作连续型属性,它可以分成[10, 20),[20, 40),[40, 70),[70, 80),[80, 100]7 个不同的分段属性,然后

分别把这 7 个分段属性当作 0-1 布尔值属性,即可完成连续型属性到 0-1 布尔值属性的转变。

在完成对原始数据集转换成 0-1 布尔值属性的操作后,对处理后的数据集进行广义多粒度决策形式背景的最优粒度选择^[29],其最优粒度选择产生的结果见表 8。

在本节中,将使用到前文中所提出的算法 2 对已完成处理后的 5 个数据集进行属性约简,除去冗余属性,获得协调粒度层约简集,并与文献[25]的算法、文献[29]的算法以及最细粒度原始数据进行对比分析。

表 7 实验数据集的基本信息

Tab. 7 The basic information of experimental datasets

数据集	领域	样本	属性	类别
Sensor	Computer Science	5 456	24	4
Obsity	Health and Medicine	2 111	16	7
Credit	Business	690	15	2
Pop	Physical	540	18	2
Wpbc	Health and Medicine	198	33	2

表 8 实验数据集预处理后的属性粒度组合

Tab.8 The combination of attribute granularities of experimental datasets after pre-processing

数据集	$V_1 \times V_2 \times \cdots \times V_q$	最优粒度组合
Sensor	$2 \times 2 \times 1 \times \cdots \times 1$	(2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
Obsity	$1 \times 2 \times 1 \times 2 \times 1 \times \cdots \times 1$	(1,2,1,2,1,1,1,1,1,1,1,1,1,1,1,1)
Credit	$1 \times 2 \times 1 \times \cdots \times 2 \times 1 \times \cdots \times 1$	(1,2,1,1,1,1,1,1,1,1,2,1,1,1,1)
Pop	$1 \times 1 \times 2 \times 1 \times 2 \times 1 \times 1 \times \cdots \times 1$	(1,1,2,1,2,1,1,1,1,1,1,1,1,1,1,1)
Wpbc	$1 \times 1 \times 1 \times 2 \times 2 \times 1 \times \cdots \times 1$	(1,1,1,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)

4.2 实验结果

文献[25]的算法是对选择出最细的决策属性的形式背景进行属性约简。文献[29]的算法是对多粒度决策形式背景进行研究,提出了广义多粒度形式背景的最优粒度选择,其目标是确定最合适的粒度,使用最粗的条件属性和最细的决策属性来优化属性粒度的组合。最细粒度原始数据是指还没有进行预处理的数据,故将其记为最细粒度原始数据。本文算法是将粒度选择(文献[29]的算法)与属性约简(本文算法 2)相结合,完成属性约简。

对选取的 5 个真实数据集和 6 种分类方法进行对比分析,对比方法包括 Fine Gaussian SVM (FGSVM)、Medium Gaussian SVM (MGSVM)、Cu-

bic SVM (CSVM)、Weighted KNN (WKNN)、Simple Tree (STree)、Linear Discriminant Analysis (LDA)。最终产生的分类结果与文献[25]、文献[29]以及最细粒度原始数据的分类结果进行比较,它们的分类精度详见表 9。在表 9 中,由于数据集的类型问题,“\”代表该数据集在此算法下无法进行分类。

实验表明,数据在进行本文约简方法后再进行分类时,其分类精度能在大多数数据集中优于仅进行文献[25]或[29]数据处理的分类精度以及最细粒度原始数据的分类精度。因此,这证明了本文提出的广义多粒度决策形式背景的属性约简方法是有效的,它能够有效提高分类方法的性能。

表 9 各算法在数据集上的分类精度
Tab. 9 The classification accuracy of each algorithm on the datasets

数据集	算法	分类精度/%					
		FGSVM	MGSVM	CSVM	WKNN	STree	LDA
Sensor	本文算法	94.6	97.7	97.9	97.3	87.7	91.8
	文献[25]算法	93.2	96.7	97.2	96.2	88.2	91.6
	文献[29]算法	92.5	97.8	93.0	97.2	87.7	84.6
	最细原始数据	89.5	92.2	95.3	93.6	88.4	79.2
Obsity	本文算法	96.6	97.1	97.3	96.7	96.3	94.4
	文献[25]算法	96.2	96.7	96.3	96.4	96.2	95.4
	文献[29]算法	96.4	96.9	96.9	96.7	96.0	94.6
	最细原始数据	86.4	95.7	97.0	\	67.3	\
Credit	本文算法	81.4	94.5	95.5	92.5	93.9	93.9
	文献[25]算法	75.3	93.7	94.1	90.2	93.9	93.9
	文献[29]算法	74.3	93.5	96.3	87.1	93.9	95.1
	最细原始数据	68.9	92.6	90.8	\	93.9	\
Pop	本文算法	94.8	94.8	91.2	95.0	94.8	95.0
	文献[25]算法	94.8	94.8	93.1	94.6	94.8	95.2
	文献[29]算法	94.8	94.8	93.4	95.0	94.4	95.2
	最细原始数据	94.5	94.5	96.4	94.5	93.7	95.4
Wpbc	本文算法	87.3	87.9	87.3	88.4	87.9	85.0
	文献[25]算法	87.3	87.9	87.3	88.4	87.9	86.1
	文献[29]算法	87.3	86.7	87.3	87.9	86.1	86.1
	最细原始数据	87.3	87.3	80.3	87.9	86.1	76.9

5 结论

形式概念分析发展至今,对形式背景的属性约简研究逐渐深入,目前已开展的研究基本上都是针对单粒度以及多粒度条件粒度层进行的,很少有对包含多粒度决策属性的数据集进行此类研究,而在本文中,则是对广义多粒度决策形式背景的属性约简进行讨论,它既考虑到条件属性的多粒度又考虑到决策属性的多粒度,并将其与其他方法进行对比分析。具体地,所提方法首先对部分协调的广义多粒度决策形式背景进行粒度选择,再对选择后的单粒度决策形式背景进行冗余类属性块的删除,保留必要的属性信息,完成对信息系统的属性约简。相较于现有的方法,利用条

件信息熵对广义多粒度决策形式背景的属性约简进行研究,能够更好的适应复杂的数据环境,解决属性冗余问题。这对以后有关形式概念分析的属性约简的研究提供了参考。

对于广义多粒度决策形式背景的属性约简的研究,本文虽然给出了相应方法,但仍有一些不足之处:①在形式背景属性约简的过程中,可能出现不止一个结果的情形,因此在得到多个结果时,怎样找到一个合适的约简集,使得对分类结果产生更有利的作用是值得研究的;②本文的属性约简是在一个固定信息系统中进行的,但在实际情况中,为了满足不同用户的需求,可以增加一些偏好属性的重要度或减少一些排斥属性的重要度,研究偏好意义下的属性约简也是有价值的;③动态数据环境下的广义多粒度决策形式背景的属性约

简也是一个后续值得探讨的问题。

参考文献

- [1] WILLE R. Restructuring lattice theory: An approach based on hierarchies of concepts[C]//RIVAL I, ed. Ordered Sets. Berlin, Germany: Springer, 1982: 445-470.
- [2] GANTER B, WILLE R. Formal concept analysis: Mathematical foundations [M]. New York: Springer-Verlag, 1999.
- [3] LI J H, MEI C L, LV Y J. Incomplete decision contexts: Approximate concept construction, rule acquisition and knowledge reduction[J]. International Journal of Approximate Reasoning, 2013, 54(1): 149-165.
- [4] WU W Z, NIU D R, LI J H, et al. Rule acquisition in generalized multi-scale information systems with multi-scale decisions[J]. International Journal of Approximate Reasoning, 2023, 154: 56-71.
- [5] MULKAR-MEHTA R, HOBBS J, HOVY E. Granularity in natural language discourse[C]//Proc of the 9th International Conference on Computational Semantics. Stroudsburg, USA: ACL, 2011: 360-364.
- [6] WEI L, WAN Q. Granular transformation and irreducible element judgment theory based on pictorial diagrams[J]. IEEE Transactions on Cybernetics, 2017, 46(2): 380-387.
- [7] 曾望林,折延宏. 面向对象的多粒度形式概念分析[J]. 计算机科学, 2018, 45(10): 51-53.
ZENG W L, SHE Y H. Object-oriented multigranulation formal concept analysis[J]. Computer Science, 2018, 45(10): 51-53.
- [8] 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.
- [9] 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.
- [10] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341-356.
- [11] KUZNETSOV S O. Machine learning on the basis of formal concept analysis[J]. Automation and Remote Control, 2001, 62(10): 1543-1564.
- [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] 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.
- [14] 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.
ZHANG W X, WEI L, QI J J. Attribute reduction theory and approach to concept lattice[J]. Science in China Series F: Information Sciences, 2005, 48(6): 713-726.
- [15] 魏玲,祁建军,张文修. 决策形式背景的概念格属性约简[J]. 中国科学: E辑, 2008, 38(2): 195-208.
WEI L, QI J J, ZHANG W X. Attribute reduction theory of concept lattice based on decision formal contexts[J]. Science in China Series F: Information Sciences, 2008, 51(7): 910-923.
- [16] WU W Z, LEUNG Y, MI J S. Granular computing and knowledge reduction in formal contexts[J]. IEEE Transaction on Knowledge and Data Engineering, 2009, 10(21): 1461-1474.
- [17] 李进金,张燕兰,吴伟志,等. 形式背景与协调决策形式背景属性约简与概念格生成[J]. 计算机学报, 2014, 37(8): 1768-1774.
LI J J, ZHANG Y L, WU W Z, et al. Attribute reduction for formal context and consistent decision formal context and concept lattice generation[J]. Chinese Journal of Computers, 2014, 37(8): 1768-1774.
- [18] SHANNON C E, WEARER W. The mathematical theory of communication. Champaign [M] Chicago: University of Illinois Press, 1971.
- [19] HU Q H, XIE Z X, YU D R. Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation[J]. Pattern Recognition, 2007, 40(12): 3509-3521.
- [20] WANG F, LIANG J Y, QIAN Y H. Attribute reduction: A dimension incremental strategy[J]. Knowledge-Based Systems, 2013, 39: 95-108.

[21] DAI J H, HU H, ZHENG G J, et al. Attribute reduction in interval-valued information systems based on information entropies[J]. *Frontiers of Information Technology & Electronic Engineering*, 2016, 17(9): 919-928.

[22] THUY N N, WONGTHANAVASU S. On reduction of attributes in inconsistent decision tables based on information entropies and stripped quotient sets[J]. *Expert Systems with Applications*, 2019, 137: 308-323.

[23] 王国胤,于洪,杨大春. 基于条件信息熵的决策表约简[J]. *计算机学报*, 2002, 25(7): 759-766.
WANG G Y, YU H, YANG D C. Decision table reduction based on conditional information entropy[J]. *Chinese Journal of Computers*, 2002, 25(7): 759-766.

[24] SINGH P K, CHERUKURI A K, LI J H. Concepts reduction in formal concept analysis with fuzzy setting using Shannon entropy[J]. *International Journal of Machine Learning and Cybernetics*, 2017, 8(1): 179-189.

[25] 李金海,周新然. 多粒度决策形式背景的属性约简[J]. *模式识别与人工智能*, 2022, 35(5): 387-400.
LI J H, ZHOU X R. Attribute reduction in multi-granularity formal decision contexts[J]. *Pattern Recognition and Artificial Intelligence*, 2022, 35(5): 387-400.

[26] 张文修,仇国芳. 基于粗糙集的不确定决策[M]. 北京:清华大学出版社, 2005.

[27] 刘之茗,李金海. 多粒度形式背景的表示[J]. *海南热带海洋学院学报*, 2019, 26(5): 59-64.
LIU Z M, LI J H. The Representation of multi-granularity formal contexts[J]. *Journal of Hainan Tropical Ocean University*, 2019, 26(5): 59-64.

[28] 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.

[29] WANG F, LI J H, YU C C. Optimal scale selection approach for classification based on generalized multi-scale formal context[J]. *Applied Soft Computing*, 2024, 152: 111277.

(编辑 张欢)