概念格中基于粗糙熵的属性约简方法

李美争[1,2] 李磊军[3,4] 米据生[3,4] 解 滨[1,2]

(河北师范大学信息技术学院 石家庄 050024)[1] (河北省网络与信息安全重点实验室 石家庄 050024)[2]

(河北师范大学数学与信息科学学院 石家庄 050024)[3]

(河北省计算数学与应用重点实验室 石家庄 050024)[4]

摘 要 属性约简是概念格理论的研究重点内容之一。通过将粗糙熵引入概念格理论中,定义了一种粗糙熵约简。首先,基于所有概念外延定义了形式背景的粗糙熵,并分析了它的性质;其次,定义了形式背景的粗糙熵约简,并揭示了粗糙熵约简与概念格约简之间的关系;在此基础上,基于属性重要度设计了计算粗糙熵的启发式算法,并通过实验验证了该算法的有效性。

关键词 概念格,属性约简,启发式算法,粗糙熵

中图法分类号 TP182 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.01.013

Rough Entropy Based Algorithm for Attribute Reduction in Concept Lattice

LI Mei-zheng[1,2] LI Lei-jun[3,4] MI Ju-sheng[3,4] XIE Bin[1,2]

(College of Information Technology, Hebei Normal University, Shijiazhuang 050024, China)[1]

(Hebei Key Laboratory of Network and Information Security, Shijiazhuang 050024, China)[2]

(College of Mathematics and Information Science, Hebei Normal University, Shijiazhuang 050024, China)[3]

(Hebei Key Laboratory of Computational Mathematics and Applications, Shijiazhuang 050024, China)[4]

Abstract Attribute reduction is one of the crucial issues in the theory study of concept lattice. In this paper, rough entropy was introduced to conduct a kind of attribute reduction. Firstly, rough entropy in a formal context was defined via the whole set of all concept extents, and the properties of rough entropy were analyzed. Secondly, a rough entropy based attribute reduction of a formal context was given, and the relationship between the rough entropy-based reduct and the concept lattice-based reduct was revealed. Based on this, a heuristic algorithm based on the attribute significance was proposed to compute a rough entropy-based reduct, and some numerical experiments were conducted to show the efficiency of the proposed methods.

Keywords Concept lattice, Attribute reduction, Heuristic algorithm, Rough entropy

概念格理论,也称为形式概念分析(Formal Concept Analysis,FCA),是德国数学家 Wille[1]于1982年提出的一种知识表示与知识发现的有力工具,已被广泛应用于人工智能[2]、软件工程[3-4]、信息检索[5]、规则提取[6-7]等领域。

属性约简是概念格理论研究的核心问题之一,其目的是寻找保持某种性质不变的最小属性子集。根据约简目标的不同,可以将基于概念格的属性约简分为3类[8]:1)基于概念格复杂性的属性约简;2)基于规则的属性约简;3)基于格结构的属性约简。

第一类属性约简的目标是降低概念格中的节点数量,删除无关紧要的概念节点,以利于重要知识的发现,在该过程中概念格的结构发生了变化[9-12]。

第二类属性约简的目标是保持特定的规则集不变。例如,Li等[13-14]提出了保持非冗余规则集不变的属性约简方法;Shao等[15]、Li等[16]分别从保持极大规则不变的角度出发研究了属性约简方法。

第三类属性约简的目标是保持特定的格结构不变,这方面的研究相对较多。张文修和魏玲首次系统地研究了保持经典概念格结构不变的属性约简理论与方法[17],并提出了一种基于辨识矩阵和辨识函数计算属性约简的布尔方法,可以获

到稿日期:2017-03-03 返修日期:2017-07-02 本文受国家自然科学基金项目(61502144,61573127,61672206,71571062),河北省高等学校自然科学基金项目(QN2017095,QN2016133),河北省高校创新团队领军人才培育计划项目(LJRC022),河北省博士后择优资助科研项目(B2016003013),河北师范大学博士基金项目(L2017B19,L2015B01)资助。

李美争(1984—),女,博士,讲师,CCF会员,主要研究方向为概念格;李磊军(1985—),男,博士,副教授,主要研究方向为粒计算、集成学习,E-mail:liljun1985@163.com(通信作者);米据生(1966—),男,博士,教授,博士生导师,主要研究方向为粒计算、近似推理;解 滨(1976—),男,博士,教授,主要研究方向为粒计算、近似推理。

得所有的属性约简。这种方法在其他概念格模型中也得到了广泛应用，例如文献[18-20]。也有学者研究属性约简的启发式算法，以便快速获得一个属性约简。例如，Wang 等[21]将封闭标签格作为算法输入，利用封闭标签的特性寻找核心属性，并将 $|a^*|$ 作为属性重要度，提出了一种启发式属性约简算法。吕跃进和李金海[22]则将全部外延的改变量作为启发式信息来设计属性约简算法。

粗糙熵是 Liang 等[23-24]为了衡量不完备信息系统中的不确定性，在相容关系下通过引入信息熵建立的一种不确定性度量。之后，黄兵等[25]将粗糙熵进一步推广到一般二元关系。本文将粗糙熵引入形式背景中，进一步定义了基于粗糙熵的属性约简，并给出了基于属性重要度的启发式粗糙熵约简算法。

本文第 1 节回顾概念格理论的基本知识；第 2 节将粗糙熵引入形式背景，并研究了它的相关性质；定义了形式背景的粗糙熵约简，并研究了它与概念格约简的关系；第 3 节首先定义了基于粗糙熵的属性重要度，并将属性重要度作为启发信息设计了属性约简算法；第 4 节通过实验验证了算法的有效性；最后总结全文并讨论了未来可能的工作方向。

1　概念格理论基础

形式背景、形式概念和概念格是概念格理论中的 3 个基本术语[26]。形式背景是一个有序三元组 $\mathbb{K}=(G,M,I)$，其中 G 是一个非空对象集合，M 是一个非空的属性集合，$I\subseteq G\times M$ 是一个从对象集 G 到属性集合 M 的二元关系，$(g,m)\in I$ 当且仅当对象 g 具有属性 m。本文假设对象集和属性集都是有限的。

当 I 满足下列两个条件时，称它是一个正则的关系（此时称 $\mathbb{K}$ 为正则的形式背景）：

(1) $\forall g\in G,\exists m,n\in M$，满足 $(g,m)\in I$，且 $(g,n)\notin I$；

(2) $\forall m\in M,\exists g,h\in G$，满足 $(g,m)\in I$，且 $(h,m)\notin I$。

定义 1[26]　设 $\mathbb{K}=(G,M,I)$ 是一个形式背景，对于任意的对象子集 X 和属性子集 B，分别定义运算：

$$X^*=\{m\in M|\ \forall g\in X,(g,m)\in I\}$$

$$B^*=\{g\in G|\ \forall m\in B,(g,m)\in I\}$$

$*$ 运算是形式背景中的基本运算。X^* 表示 X 中所有对象共同具有的所有属性；B^* 表示具有 B 中所有属性的对象集合。一般情况下，$\{g\}^*$ 简记为 g^*，$\{m\}^*$ 简记为 m^*。

性质 1[26]　设 $\mathbb{K}=(G,M,I)$ 是一个形式背景，如果 $X,X_1,X_2\subseteq G,B,B_1,B_2\subseteq M,T$ 是一个指标集，那么

(1) $X_1\subseteq X_2\Rightarrow X_2^*\subseteq X_1^*$，$B_1\subseteq B_2\Rightarrow B_2^*\subseteq B_1^*$；

(2) $X\subseteq X^{**}$，$B\subseteq B^{**}$；

(3) $X^*=X^{***}$，$B^*=B^{***}$；

(4) $X\subseteq B^*\Leftrightarrow B\subseteq X^*$；

(5) $(\bigcup_{t\in T}X_t)^*=\bigcap_{t\in T}X_t^*$，$(\bigcup_{t\in T}B_t)^*=\bigcap_{t\in T}B_t^*$。

定义 2[26]　设 $\mathbb{K}=(G,M,I)$ 是一个形式背景，$X\subseteq G$，$B\subseteq M$。若 $X^*=B$ 且 $B^*=X$，则称有序二元组 (X,B) 是一个（形式）概念，称 X 是这个概念的外延，B 是这个概念的内涵。

记背景 $\mathbb{K}$ 中所有概念的集合为 $L(\mathbb{K})$。$L(\mathbb{K})$ 上的偏序关系为：

$$(X_1,B_1)\leqslant(X_2,B_2)\Leftrightarrow X_1\subseteq X_2(\Leftrightarrow B_1\supseteq B_2)$$

记 $\mathscr{L}(\mathbb{K})=(L(\mathbb{K}),\leqslant)$，$\mathscr{L}(\mathbb{K})$ 是一个完备格，被称为 $\mathbb{K}$ 的概念格。两个概念的上下确界分别为：

$$(X_1,B_1)\wedge(X_2,B_2)=(X_1\cap X_2,(B_1\cup B_2)^{**})$$

$$(X_1,B_1)\vee(X_2,B_2)=((X_1\cup X_2)^{**},B_1\cap B_2)$$

例 1　表 1 列出了一个形式背景 $\mathbb{K}=(G,M,I)$，其中对象集 G 中有 4 个对象 g_1,g_2,g_3 和 g_4，对象集 M 中有 5 个属性 a,b,c,d 和 e，对象和属性具有关系则用“1”表示，否则用“0”表示。$\mathbb{K}$ 的概念格可以用 Hasse 图的形式来表示（见图 1）。

表 1　形式背景 $\mathbb{K}=(G,M,I)$

Table 1　A formal context $\mathbb{K}=(G,M,I)$

G/M	a	b	c	d	e
g_1	1	0	1	1	1
g_2	1	1	0	0	0
g_3	0	0	1	0	1
g_4	1	1	0	0	0

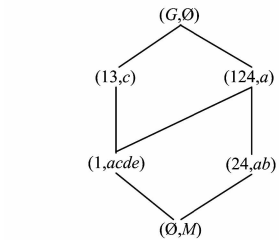

图 1　$\mathbb{K}$ 的概念格

Fig. 1　Concept lattice of $\mathbb{K}$

设 $\mathbb{K}=(G,M,I)$ 是一个形式背景，$D\subseteq M$。记 $I_D=I\cap(G\times D)$，则 $\mathbb{K}_D=(G,D,I_D)$ 也是一个形式背景，称它为 $\mathbb{K}$ 的一个子形式背景。

记 $Ext(\mathbb{K})=\{X\subseteq G|\ (X,B)\in L(\mathbb{K})\}$，它表示 $\mathbb{K}$ 的所有概念的外延集合。

定理 1[26]　设 $\mathbb{K}=(G,M,I)$ 是一个背景，$D\subseteq M$，则 $Ext(\mathbb{K})\supseteq Ext(\mathbb{K}_D)$ 成立。

定理 1 说明子形式背景的外延集合包含于原形式背景的外延集合。

定义 3[17]　设 $\mathbb{K}_1=(G,M_1,I_1)$ 和 $\mathbb{K}_2=(G,M_2,I_2)$ 是两个形式背景，$\mathscr{L}(\mathbb{K}_1)$ 和 $\mathscr{L}(\mathbb{K}_2)$ 是对应的概念格。若对于任意的 $(X_2,B_2)\in\mathscr{L}(\mathbb{K}_1)$，总存在 $(X_1,B_1)\in\mathscr{L}(\mathbb{K}_2)$，使得 $X_1=X_2$，则称 $\mathscr{L}(\mathbb{K}_1)$ 细于 $\mathscr{L}(\mathbb{K}_2)$，记作 $\mathscr{L}(\mathbb{K}_1)\leqslant\mathscr{L}(\mathbb{K}_2)$。

如果 $\mathscr{L}(\mathbb{K}_1)\leqslant\mathscr{L}(\mathbb{K}_2)$ 和 $\mathscr{L}(\mathbb{K}_2)\leqslant\mathscr{L}(\mathbb{K}_1)$ 同时成立，那么两个概念格同构，记作 $\mathscr{L}(\mathbb{K}_1)\cong\mathscr{L}(\mathbb{K}_2)$。

定义 4[17]　设 $\mathbb{K}=(G,M,I)$ 是一个形式背景，$D\subseteq M$。若 $\mathscr{L}(\mathbb{K})\cong\mathscr{L}(\mathbb{K}_D)$，则称 D 是 $\mathbb{K}$ 的一个协调集。更进一步，若任意的 $d\in D$，$\mathscr{L}(\mathbb{K}_D)\ncong\mathscr{L}(\mathbb{K}_{D-\{d\}})$，则称 D 是 $\mathbb{K}$ 的一个约简。

为了便于区分，本文将这种协调集/约简称为概念格协调集/约简。实际上，$\mathbb{K}$ 的一个概念格协调集 D 是保持 $\mathbb{K}$ 的外延集合不变的属性子集，即 $Ext(\mathbb{K})=Ext(\mathbb{K}_D)$。

2　形式背景的粗糙熵约简

本节首先定义形式背景的粗糙熵，并研究它的性质；然后

基于粗糙熵定义一种属性约简，并讨论这种约简与概念格约简之间的关系。

2.1 形式背景的粗糙熵

文献[25]在定义基于一般二元关系的粗糙熵时，首先给出了每个对象在这个关系下的邻域，然后统计每个对象的邻域个数，并对其取对数之后加权求和，从而得到知识的粗糙熵。在概念格中，概念的外延可以看作是外延中的每一个元素的邻域，基于这种邻域可以定义背景的一种粗糙熵。

定义 5 设$\mathbb{K}=(G,M,I)$是一个形式背景，$D\subseteq M$，$\forall g\in G$，称$N_D(g)=\{X\in Ext(\mathbb{K}_D)\mid g\in X\}$为$g$的关于属性子集$D$的邻域系，或者简称为$g$的$D$-邻域系。称$E_R(D)=\frac{1}{|G|}\sum_{g\in G}\log_2|N_D(g)|$为子形式背景$\mathbb{K}_D$的粗糙熵。

如果$X\in N_D(g)$，那么$g\in X$，根据性质 1 可得$g^{*D*D}\subseteq X$（$*D$表示$\mathbb{K}_D$中的$*$运算），这说明X实际上是大于(g^{*D*D},g^{*D})的某个概念的外延，因此g的D-邻域系是由子形式背景$\mathbb{K}_D$中所有大于(g^{*D*D},g^{*D})的概念外延构成的集合，它在一定程度上体现了概念格的深度，是对概念格结构的一种纵向描述。另一方面，粗糙熵$E_R(D)$考虑了每个对象的D-邻域系，这在一定程度上刻画了概念格的广度，是对概念格结构的一种横向描述。因此，粗糙熵实际上从纵向与横向两个方向综合度量了概念格的层次结构。

性质 2 设$\mathbb{K}=(G,M,I)$是一个形式背景，$D\subseteq M$，下列命题成立：

(1)$\forall D\subseteq M,\forall g\in G,G\in N_D(g)$；

(2)$E_R(D)\geqslant 0$；

(3)$E_R(D)=0\Leftrightarrow Ext(\mathbb{K}_D)=\{\varnothing,G\}$；

(4)若$\forall g\in G,|N_D(g)|=2$，则$E_R(D)=1$。

证明：(1)显然成立。

(2)由命题(1)可得。

(3)$E_R(D)=0\Leftrightarrow\forall g\in G,\log_2|N_D(g)|=0$

$\Leftrightarrow\forall g\in G,|N_D(g)|=1$

$\Leftrightarrow\forall g\in G,N_D(g)=\{G\}$

$\Leftrightarrow Ext(\mathbb{K}_D)=\{\varnothing,G\}$

(4)若$|N_D(g)|=2$，则$\log_2|N_D(g)|=1$，由g的任意性可得：

$$E_R(D)=\frac{1}{|G|}\sum_{g\in G}\log_2|N_D(g)|=E_R(D)=1$$

推论 1 当形式背景$\mathbb{K}=(G,M,I)$正则时，$E_R(M)\geqslant 1$。

证明：当形式背景正则时，$\forall g\in G,Em\in M$，使得$(g,m)\in I$，即$g\in m^*$。又因为$Eh\in G$，使得$(h,m)\notin I$，即$h\notin m^*$，这说明$m^*\neq G$，所以$\{G,m^*\}\subseteq N_M(g)$，即$|N_M(g)|\geqslant 2$，因此：

$$E_R(M)=\frac{1}{|G|}\sum_{g\in G}\log_2|N_M(g)|$$

$$\geqslant\frac{1}{|G|}\sum_{g\in G}\log_2 2=1$$

定理 2 设$\mathbb{K}=(G,M,I)$是一个形式背景。若$D\subseteq F\subseteq M$，则$E_R(D)\leqslant E_R(F)$。

证明：任意$X\in N_D(g)$，$X\in Ext(\mathbb{K}_D)$且$g\in X$成立。因为$D\subseteq F\subseteq A$，所以根据定理 1 可得$Ext(\mathbb{K}_D)\subseteq Ext(\mathbb{K}_F)$，则$X\in Ext(\mathbb{K}_F)$，又因为$g\in X$，所以$X\in N_F(g)$。由$X$的任意性即得$N_D(g)\subseteq N_F(g)$，$|N_D(g)|\leqslant|N_F(g)|$。由$g$的任意性可得$E_R(D)\leqslant E_R(F)$。

定理 2 说明子形式背景的粗糙熵小于原形式背景的粗糙熵，在此意义下，形式背景的粗糙熵具有单调性。

定义 6 设$\mathbb{K}=(G,M,I)$是一个形式背景，$D\subseteq M$。若$E_R(D)=E_R(M)$，则称D为形式背景$\mathbb{K}$的一个粗糙熵协调集；若$E_R(D)=E_R(M)$，且任意$d\in D$，$E_R(D-\{d\})<E_R(M)$，则称D是形式背景$\mathbb{K}$的一个粗糙熵约简。

定理 3 任意形式背景的粗糙熵约简集一定存在。

证明：显然M是一个粗糙熵协调集。如果任意$m\in M$，都有$E_R(M-\{m\})<E_R(M)$，那么M是一个粗糙熵约简。否则，假设存在$m_0\in M$，使得$E_R(M-\{m_0\})=E_R(M)$，那么$M-\{m_0\}$是一个粗糙熵协调集。重复上述过程，由于M是有限集合，因此必然存在粗糙熵约简。

形式背景$\mathbb{K}$的粗糙熵约简可能不止一个。记$\{D_t\mid t\in\Omega\}$（Ω是一个指标集）为$\mathbb{K}$的所有粗糙熵约简的集合。根据属性与粗糙熵约简的关系，可以将其分为 3 类：

(1)若属性$m\in\bigcap_{t\in\Omega}D_t$，则称$m$为核心属性；

(2)若属性$m\in\bigcup_{t\in\Omega}D_t-\bigcap_{t\in\Omega}D_t$，则称$m$为相对必要属性；

(3)若属性$m\in M-\bigcup_{t\in\Omega}D_t$，则称$m$为绝对不必要属性。

相对必要属性和绝对不必要属性统称为不必要属性。

定理 4 设$\mathbb{K}=(G,M,I)$是一个形式背景，下列命题成立：

(1)m是一个核心属性当且仅当$M-\{m\}$不是粗糙熵协调集；

(2)m是不必要属性当且仅当$M-\{m\}$是粗糙熵协调集。

证明：(1)“⇒”假设m是一个核心属性，显然m在每个粗糙熵约简中，因此也在每个粗糙熵协调集中。这说明$M-\{m\}$不是粗糙熵协调集。

“⇐”如果$M-\{m\}$不是粗糙熵协调集，那么$M-\{m\}$的任何真子集都不是粗糙熵协调集，因此也不是粗糙熵约简，这说明任何粗糙熵约简都必须包含m，即m是一个核心属性。

(2)“⇒”若m是不必要属性，则存在一个粗糙熵约简D，$m\notin D$。因为$D\subseteq M-\{m\}\subseteq M$，$E_R(D)\leqslant E_R(M-\{m\})\leqslant E_R(M)$（定理 2），又因为$D$是粗糙熵约简，所以$E_R(D)=E_R(M)$，综上可得$E_R(D)=E_R(M-\{m\})$，这说明$M-\{m\}$是一个粗糙熵协调集。

“⇐”如果$M-\{m\}$是一个粗糙熵协调集，那么至少存在一个粗糙熵约简$D\subseteq M-\{m\}$，即$m\notin D$，这说明m是一个不必要属性。

2.2 形式背景的粗糙熵

形式背景$\mathbb{K}$的粗糙熵约简就是保持形式背景粗糙熵不变的最小属性子集，而粗糙熵从横向和纵向综合度量了概念格的层次结构。另一方面，概念格约简是保持概念格结构不变的最小属性子集，那么这两种约简之间有什么关系呢？下文将给出这个问题的答案。

定理 5 设$\mathbb{K}=(G,M,I)$是一个形式背景，$D\subseteq M$。下列说法是等价的：

(1)D 是$\mathbb{K}$ 的一个粗糙熵协调集；

(2)D 是$\mathbb{K}$ 的一个概念格协调集。

证明：(1)⇒(2)。因为 $D\subseteq M$，所以 $Ext(\mathbb{K})\supseteq Ext(\mathbb{K}_D)$，由定义 5 可知 $\forall g\in G, N_D(g)\subseteq N_M(g)$，所以 $|N_D(g)|\leqslant|N_M(g)|$。若存在一个 $g_0\in G$，使得 $|N_D(g)|<|N_M(g)|$ 成立，则 $E_R(D)<E_R(M)$。这与 D 是$\mathbb{K}$ 的一个粗糙熵协调集矛盾。因此 $\forall g\in G, |N_D(g)|=|N_M(g)|$，又因为 $N_D(g)\subseteq N_M(g)$，所以 $N_D(g)=N_M(g)$ 成立。

$$Ext(\mathbb{K})=(\bigcup_{g\in G}N_M(g))\cup\{\varnothing\}$$
$$=(\bigcup_{g\in G}N_D(g))\cup\{\varnothing\}=Ext(\mathbb{K}_D)$$

即 D 是概念格协调集。

(1)⇐(2)。假设 D 是$\mathbb{K}$ 的一个概念格协调集，则 $Ext(\mathbb{K})=Ext(\mathbb{K}_D)$，那么由定义 5 可得 $\forall g\in G, N_D(g)=N_M(g)$，进而可得：

$$\frac{1}{|G|}\sum_{g\in G}\log_2|N_D(g)|=\frac{1}{|G|}\sum_{g\in G}\log_2|N_M(g)|$$

因此 $E_R(D)=E_R(M)$，即证 D 是$\mathbb{K}$ 的一个粗糙熵协调集。

推论 2　$\mathbb{K}=(G,M,I)$ 是一个形式背景，$D\subseteq M$。下列说法是等价的：

(1)D 是$\mathbb{K}$ 的一个粗糙熵约简；

(2)D 是$\mathbb{K}$ 的一个概念格约简。

3　形式背景中计算粗糙熵约简的方法

本节首先提出了两种基于粗糙熵的属性重要度，并引入了外延矩阵对其进行计算，最后利用这两种属性重要度设计了属性约简算法。

3.1　基于粗糙熵的属性重要度

定义 7　设$\mathbb{K}=(G,M,I)$ 是一个形式背景，$D\subseteq M$。$\forall m\in M-D$，m 相对于 D 的重要度定义为：

$$SIG_{out}(m,D)=E_R(D\cup\{m\})-E_R(D)$$

定义 8　设$\mathbb{K}=(G,M,I)$ 是一个形式背景，$D\subseteq M$。$\forall m\in D$，m 在 D 的重要度定义为：

$$SIG_{in}(m,D)=E_R(D)-E_R(D-\{m\})$$

定理 6　设$\mathbb{K}=(G,M,I)$ 是一个形式背景，$m\in M$。

(1)m 是一个核心属性当且仅当 $SIG_{out}(m,M-\{m\})>0$；

(2)m 是不必要属性当且仅当 $SIG_{out}(m,M-\{m\})=0$。

证明：(1)m 是一个核心属性⇔$M-\{m\}$ 不是粗糙熵协调集(定理 5)⇔$E(M)>E(M-\{m\})$(定义 6 和性质 2)⇔$SIG_{out}(m,M-\{m\})>0$。

(2)m 是不必要属性⇔$M-\{m\}$ 是粗糙熵协调集(定理 5)⇔$E(M)=E(M-\{m\})$(定义 6 和性质 2)。

定理 6 刻画了核心属性和不必要属性的数字特征。

根据定义 7 和定义 8 可知，计算属性重要度的核心步骤是计算粗糙熵。接下来引入外延矩阵，并将其用于计算形式背景的粗糙熵。

定义 9　设$\mathbb{K}=(G,M,I)$ 是一个形式背景，$D\subseteq M$，k_D 表示 $Ext(\mathbb{K}_D)-\{\varnothing\}$ 的大小，记 $G=\{g_1,g_2,\cdots,g_{|G|}\}$。任意的 $X_i\in(Ext(\mathbb{K}_D)-\{\varnothing\})$，$a_i$ 是一个 $|G|$ 维列向量，其中：

$$a_i(j)=\begin{cases}1, & g_j\in X_i\\ 0, & g_j\notin X_i\end{cases}$$

称 $Mat(\mathbb{K}_D)=(a_1,a_2,\cdots,a_{kD})$ 为$\mathbb{K}_D$ 的外延矩阵。

定理 7　设 $Mat(\mathbb{K}_D)=(a_1,a_2,\cdots,a_{kD})$ 是$\mathbb{K}_D$ 的外延矩阵，则：

$$b_D=(|N_D(g_1)|,|N_D(g_2)|,\cdots,|N_D(g_{|G|})|)^{\mathrm{T}}=\sum_{i=1}^{kD}a_i$$

证明：$X_i\in N_D(g_j)\Leftrightarrow g_j\in X_i\Leftrightarrow a_i(j)=1$，得证。

算法 1　计算形式背景的粗糙熵

输入：形式背景$\mathbb{K}=(G,M,I)$

输出：$E_R(M)$

Step1　计算$\mathbb{K}$ 的外延矩阵 $Mat(\mathbb{K})=(a_1,a_2,\cdots,a_k)$；

Step2　对 $Mat(\mathbb{K})$ 的每行求和，得 $b_M=(|N_M(g_1)|,|N_M(g_2)|,\cdots,|N_M(g_{|G|})|)^{\mathrm{T}}$；

Step3　根据定义 5 和定理 7 计算得 $E_R(M)$。

算法 1 描述了利用外延矩阵计算形式背景的粗糙熵的过程，接下来通过一个例子具体说明该算法的计算过程。

例 2　设$\mathbb{K}$ 为例 1 中给出的形式背景，计算$\mathbb{K}$ 的粗糙熵。

(1)计算$\mathbb{K}$ 的外延矩阵，得 $Ext(\mathbb{K})=\{G,\{1,3\},\{1,2,4\},\{1\},\{2,4\},\varnothing\}$，因此：

$$Mat(\mathbb{K})=\begin{pmatrix}1&1&1&1&0\\1&0&1&0&1\\1&1&0&0&0\\1&0&1&0&1\end{pmatrix}$$

(2)对 $Mat(\mathbb{K})$ 的每行求和，得 $b_M=(4,3,2,3)^{\mathrm{T}}$。

(3)根据定义 5 计算 $E_R(M)$：

$$E_R(M)=\frac{1}{4}(\log_24+\log_23+\log_22+\log_23)$$
$$=\frac{3}{4}+\frac{\log_23}{2}$$

3.2　基于粗糙熵的属性重要度

本节给出了计算粗糙熵约简的启发式算法(见算法 2)。算法 2 主要分为 3 个阶段：1)计算核心属性(Step1 和 Step2)；2)以核心属性集合为出发点，逐渐增加外部重要度最大的属性直到获得一个粗糙熵协调集(Step3－Step5)；3)从粗糙熵协调集中删除冗余的属性(Step6)。

算法 2　计算粗糙熵约简的启发式算法

输入：形式背景$\mathbb{K}=(G,M,I)$

输出：$\mathbb{K}$ 的一个粗糙熵约简 D

Step1　初始化 CORE=∅。

Step2　对于每个 $m\in M$，计算 m 相对于 $M-\{m\}$ 的重要度 $SIG_{out}(m,M-\{m\})$。如果 $SIG_{out}(m,M-\{m\})>0$，那么将 m 加入 CORE。

Step3　令 D=CORE，CAND=M－D。

Step4　判断 $E_R(M)>E_R(D)$ 是否成立：若不成立，则转 Step6，否则执行 Step5。

Step5　任意 $m\in CAND$，计算 $SIG_{out}(m,D)$；如果 $SIG_{out}(n,D)=\max_{m\in CAND}SIG_{out}(m,D)$，那么 $D=D\cup\{n\}$，$CAND=CAND-\{n\}$；返回 Step4。

Step6　对于任意的 $m\in(D-CORE)$，如果 $SIG_{in}(m,D)=0$，$D=D-\{m\}$。

Step7　返回 D。

接下来通过一个例子具体说明算法 2 的计算过程。

例 3　设$\mathbb{K}$ 为例 1 中给出的形式背景。下面通过算法 2

计算$\mathbb{K}$的粗糙熵约简。

(1)初始化 $CORE=\emptyset$。

(2)$SIG_{out}(a,M-\{a\})=\frac{\log_2 3}{2}$, $SIG_{out}(b,M-\{b\})=\frac{\log_2 3-1}{2}$, $SIG_{out}(c,M-\{c\})=0$, $SIG_{out}(d,M-\{d\})=0$, $SIG_{out}(e,M-\{e\})=0$。因此,$CORE=\{a,b\}$, $D=CORE=\{a,b\}$, $CAND=\{c,d,e\}$。

(3)根据算法 1 计算 $E_R(D)=\frac{1}{4}+\frac{\log_2 3}{2}$,因为 $E_R(D)<E_R(M)$,所以 D 不是约简。

(4)$SIG_{out}(c,D)=\frac{1}{2}$, $SIG_{out}(d,D)=SIG_{out}(e,D)=\frac{\log_2 3-1}{4}$,因为$\frac{1}{2}>\frac{\log_2 3-1}{4}$,所以将 c 并入 D 中,$D=\{a,b,c\}$。

(5)计算 $E_R(D)=\frac{3}{4}+\frac{\log_2 3}{2}$, $E_R(D)=E_R(M)$ 成立,且任何一个属性都不能被删除,因此 $D=\{a,b,c\}$是一个粗糙熵约简。

4 实验分析

为了说明本文方法的有效性,将算法 2 与基于辨识矩阵求属性约简的算法[27]进行对比。首先计算概念格,在此基础上构造辨识矩阵,并调用文献[27]中的方法得到一个概念格约简,具体过程如算法 3 所示。

算法 3 基于辨识矩阵计算形式背景的一个属性约简

输入:形式背景$\mathbb{K}=(G,M,I)$

输出:$\mathbb{K}$的一个概念格约简 D

Step1 调用现有概念格构造算法计算 $\mathcal{L}(\mathbb{K})$;

Step2 根据文献[17]计算辨识矩阵;

Step3 调用文献[27]中的方法计算$\mathbb{K}$的一个概念格约简 D;

Step4 返回 D。

实验数据集 Zoo, Soybean 来源于 UCI[28]。原始数据转化为形式背景后的信息如表 2 所列。

表 2 实验数据集

Table 2 Experimental datasets

数据集	$\lvert G\rvert$	$\lvert M\rvert$	$\lvert L\rvert$
Zoo	67	27	247
Soybean	47	49	1048

实验环境如下:操作系统为 Win7 系统 64 位,处理器为 Inter(R) Core(TM) i5-4590 CPU @3.30GHz 3.30GHz,内存为 8GB,编程语言为 MATLAB R2016a。

对于数据集 Zoo,算法 2 得到的约简是{a1,a3,a5-a16,a18-a21,a27},算法 3 得到的约简是{a1,a3-a16,a18-a20,a27},其差别在于前者包含属性 a21,后者包含属性 a4。经检验,$a4^*=a21^*$,这两个属性约简是等价的。

对于数据集 Soybean,算法 2 得到的约简是{a1-a30,a32,a35,a36,a40,a41,a44,a45,a46},算法 3 得到的约简是{a1-a30,a32,a34,a35,a36,a40,a41,a44,a45},两者的差别在于前者包含 a9 和 a34,后者包含 a46 和 a47,经检验,$a9^*=a47^*$, $a34^*=a46^*$,因此这两个属性约简是等价的。

两个算法的运行时间如表 3 所列。从表 3 可以看出,在数据集 Zoo 上,算法 2(计算粗糙熵约简的启发式算法)的运行时间与算法 3(基于辨识矩阵计算形式背景的一个属性约简)大致相当,略优于算法 3。在数据集 Soybean 上,算法 2 的运行时间远少于算法 3。

表 3 运行时间/s

Table 3 Runtime/s

数据集	算法 2	算法 3
Zoo	23.03	23.05
Soybean	140.9991	595.6889

结束语 本文将粗糙熵引入概念格理论中,利用所有的概念外延定义了形式背景的粗糙熵,并利用粗糙熵衡量属性重要度,在此基础上提出了一种基于粗糙熵的启发式属性约简方法,并通过实验验证了该方法的有效性。下一步工作将定义决策形式背景的粗糙熵约简,并设计启发式的属性约简算法。

参 考 文 献

[1] WILLE R. Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts [M]// Ordered Sets. Springer Netherlands, 1982: 445-470.

[2] PRISS U. Formal Concept Analysis In Information Science [J]. Arist, 2006, 40(1): 521-543.

[3] POELMANS J, IGNATOV D I, KUZNETSOV S O, et al. Formal Concept Analysis in Knowledge Processing: A Survey on Applications [J]. Expert Systems with Applications, 2013, 40(16): 6538-6560.

[4] TILLEY T, EKLUND P. Citation Analysis Using Formal Concept Analysis: A Case Study in Software Engineering [C]// Proceedings of the 18th International Workshop on Database and Expert Systems Applications. Washington, USA: IEEE, 2007: 545-550.

[5] POELMANS J, IGNATOV D I, VIAENE S, et al. Text Mining Scientific Papers: A Survey on FCA-Based Information Retrieval Research [C]// Proceeding of the 12th Industrial Conference on Advances in Data Mining: Applications and Theoretical Aspects. Berlin, Germany: Springer, 2012: 273-287.

[6] LAKHAL L, STUMME G. Efficient Mining of Association Rules Based on Formal Concept Analysis [C]// Formal Concept Analysis. Berlin, Germany: Springer, 2005: 180-195.

[7] LIANG J Y, WANG J H. An Algorithm for Extracting Rule-Generating Sets Based on Concept Lattice[J]. Journal of Computer Research nd Development, 2004, 41(8): 1339-1344. (in Chinese)
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.

[8] SHAO M W, LI K W. Attribute reduction in generalized one-sided formal contexts[J]. Information Sciences, 2016, 378: 317-327.

[9] CHEUNG K S, VOGEL D. Complexity reduction in lattice-based information retrieval [J]. Information Retrieval, 2005, 8(2): 285-299.

[10] BĚLOHLÁVEK R, SKLENÁŘV, ZACPAL J. Crisply generated fuzzy concepts [C]// International Conference on Formal Concept Analysis. Berlin Heidelberg: Springer, 2005: 269-284.

[11] KUMAR C A, SRINIVAS S. Concept lattice reduction using fuzzy K-means clustering[J]. Expert systems with applications, 2010, 37(3): 2696-2704.

[12] 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.

[13] LI J H, MEI C L, LV Y J. Knowledge Reduction in Decision Formal Contexts [J]. Knowledge-Based Systems, 2011, 24(5): 709-715.

[14] 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.

[15] SHAO M W, LEUNG Y, WU W Z. Rule Acquisition and Complexity Reduction in Formal Decision Contexts[J]. International Journal of Approximate Reasoning, 2014, 55(1): 259-274.

[16] LI L J, MI J S, XIE B. Attribute Reduction Based on Maximal Rules in Decision Formal Context[J]. International Journal of Computational Intelligence Systems, 2014, 7(6): 1044-1053.

[17] ZHANG W X, WEI L, QI J J. Attribute Reduction Theory and Approach to Concept Lattice[J]. Science in China Series E, 2005, 35(6): 628-639. (in Chinese)
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.

[18] MI J S, LEUNG Y, WU W Z. Approaches to Attribute Reduction in Concept Lattices Induced by Axialities[J]. Knowledge-Based Systems, 2010, 23(6): 504-511.

[19] SHAO M W, LEUNG Y, WANG X Z, et al. Granular Reducts of Formal Fuzzy Contexts [J]. Knowledge-Based Systems, 2016, 114: 156-166.

[20] LI M Z, WANG G Y. Approximate Concept Construction With Three-Way Decisions and Attribute Reduction in Incomplete Contexts [J]. Knowledge-Based Systems, 2016, 91: 165-178.

[21] WANG J H, LIANG J Y, QIAN Y H. A Heuristic Method to Attribute Reduction for Concept Lattice [C]// International Conference on Machine Learning and Cybernetics. New York: IEEE Press, 2010, 1: 483-487.

[22] LV Y J, LI J H. Heuristic Algorithms for Attribute Reduction on Concept Lattice[J]. Computer Engineering and Applications, 2009, 45(2): 154-157. (in Chinese)
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.

[23] LIANG J Y, XU Z B. Uncertainty Measures of Roughness of Knowledge and Rough Sets in Incomplete Information Systems [C]// Proceedings of the 3rd World Congress on Intelligent Control and Automation, 2000. New York: IEEE Press, 2000: 2526-2529.

[24] LIANG J Y, XU Z B. The Algorithm on Knowledge Reduction in Incomplete Information Systems [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(1): 95-103.

[25] HUANG B, ZHOU X Z, SHI Y C. Entropy of Knowledge and Rough Set Based on General Binary Relation [J]. Systems Engineering-Theory & Practice, 2004, 24(1): 93-96. (in Chinese)
黄兵，周献中，史迎春. 基于一般二元关系的知识粗糙熵与粗集粗糙熵[J]. 系统工程理论与实践，2004，24(1)：93-96.

[26] GANTER B, WILLE R. Formal Concept Analysis: Mathematical Foundations [M]. Berlin, Germany: Springer, 1999.

[27] YAO Y Y, ZHAO Y. Discernibility matrix simplification for constructing attribute reducts[J]. Information Sciences, 2009, 179(7): 867-882.

[28] LICHMAN M. UCI machine learning repository [EB/OL]. http://archive.ics.uci.edu/ml.

(上接第 66 页)

[18] YU H, LIU Z G, WANG G Y. An automatic method to determine the number of clusters using decision-theoretic rough set [J]. International Journal of Approximate Reasoning, 2014, 55(1): 101-115.

[19] YU H, ZHANG C, WANG G Y. A tree-based incremental overlapping clustering method using the three-way decision theory [J]. Knowledge-Based Systems, 2016, 91(C): 189-203.

[20] YU H, JIAO P, YAO Y Y, et al. Detecting and refining overlapping regions in complex networks with three-way decisions [J]. Information Sciences, 2016, 373(1): 21-41.

[21] MACQUEEN J B. Some methods for classification and analysis of multivariate observations [C]// Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967: 281-297.

[22] PERONA P, FREEMAN W T. A Factorization Approach to Grouping [C]// European Conference on Computer Vision. Berlin: Springer, 1998: 655-670.

[23] SHI J, MALIK J. Normalized cuts and image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905.

[24] SCOTT G L, LONGUET-HIGGINS H C. Feature grouping by relocalisation of eigenvectors of proximity matrix [C]// Proceedings of British Machine Vision Conference. Oxford: BMVA Press, 1990: 103-108.

[25] NG A, JORDAN M, WEISS Y. On spectral clustering: analysis and an algorithm[C]// International Conference on Neural Information Processing Systems: Natural and Synthetic. Shanghai: MIT Press, 2001: 849-856.

[26] UCI machine Learning Repository [OL]. http://www.ics.uci.edu/mlearn/MLRepository.html.