第 56 卷　第 4 期
2020 年 7 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 56, No. 4
July, 2020

DOI:10.13232/j.cnki.jnju.2020.04.004

# 信息表中约简补集对及其一般定义

王宝丽[1*],姚一豫[2]

(1.运城学院数学与信息技术学院,运城,044000;2.加拿大里贾纳大学计算机系,里贾纳,S4S 0A2,加拿大)

**摘　要:**约简是粗糙集理论的核心研究内容,也是粗糙集区分于其他数据挖掘算法最精彩的部分.现有约简以信息表中满足某种性质的最小属性子集为主要特征,不考虑属性集之间的相互作用和相互补充关系.从一对满足某种性质的互补属性子集出发,寻找其最小互补属性子集对,从而保障所求约简属性集对满足一定的互补性,可对实际应用中的限制条件约简进行建模.其次,从约简需保持的性质与特定划分不确定度量之间的关系出发,提出几类基本的约简补集对的一般化定义.此外,结合经典启发式约简算法,提出约简补集对的通用求解算法.最后运用一个中西医结合诊断肺炎的约简补集对求解实例说明所提算法的实用性及有效性.

**关键词:**粗糙集,约简,属性补集对,约简补集对,不确定性度量

**中图分类号:**TP391　　**文献标识码:**A

# A definition of complementary attribute reduct pairs in an information table

Wang Baoli[1*], Yao Yiyu[2]

(1. School of Mathematics & Information Technology, Yuncheng University, Yuncheng, 044000, China;
2. Department of Computer Science, University of Regina, Regina, S4S 0A2, Canada)

**Abstract:** Attribute reducts are an essential notion in rough set theory and one of the most distinguishing features of the rough set approaches to data mining. A reduct construction algorithm is to find a minimal set of attributes that has some property. However, it does not consider the complementary relationship between attributes. This paper proposes a general definition of complementary attribute pairs and complementary attribute reduct pairs to model a reciprocal relationship between subsets of attributes in real applications. Three primary definitions are adapted for different situations. A reduct construction algorithm is proposed for finding a complementary pair of subsets of attributes in an information table. A diagnosis and treatment example that combines traditional Chinese medicine and Western medicine is used to illustrate the practicality and usefulness of the proposed concepts at the end of this paper.

**Key words:** rough set, attribute reduct, complementary pair of attributes, complementary reduct pair, uncertainty measure

粗糙集理论的基本思想是基于一定的近似空间对概念及分类对象进行近似描述.作为一种处理不精确、不一致和不确定知识的粒计算方法,其理论和方法都已取得了巨大的进展,是当前人工智能理论及其应用领域中的研究热点之一[1-2].

属性约简是粗糙集理论研究的一个重要问题,也是其区别于其他数据挖掘算法最核心的特征.属性约简在保持信息表某种性质不变的前提下,删除不重要的或者不相关的属性.无论是保持哪种性质的约简,其约简通常不是唯一的.目

基金项目:国家自然科学基金(61703363),山西省自然科学基金(201901D211462,201801D121148),山西省高等学校科技创新项目(2019L0864),山西省重点实验室开放课题(CICIP2018008)

收稿日期:2020－06－24

*通讯联系人,E-mail:pollycomputer@163.com

前求解约简主要有两种方式:区分矩阵法[3-6]和启发式算法[7-16]. 区分矩阵法直观、易理解,它通过构造属性区分矩阵建立差别逻辑函数并计算其最小析取范式来找到信息表的完全约简,但如果当数据集比较大时,求解完全约简会比较困难,现已证明完全约简的求解是NP-hard问题. 启发式算法将系统保持的某种性质运用一种度量进行表征,以属性的度量表现特性为启发式信息求取信息表的一个约简. 目前启发式约简算法应用最为广泛,人们针对不同的应用场景,设计了不同的改进启发式算法,提高了约简的求解效率[7-16].

苗夺谦和胡桂荣[8]提出一种以互信息为启发式信息具有多项式时间复杂度的决策表约简算法. 王国胤等[9]比较代数观及信息观下决策表约简得到二者的等价性质及不同特征,提出基于条件熵的约简算法. 胡峰和王国胤[10]提出给定属性序,采用分治策略计算定属性序下的唯一约简,降低算法的时空复杂度,可用来高效计算海量数据的属性约简. Qian et al[11]通过正向近似的策略不断缩小粗糙集的边界域,减少当前计算启发式信息的数据量,从而获取与原信息表相同的属性约简,极大地提高了大数据的属性约简效率. Liang et al[12]从多粒度视角采用统计抽样计算各个对象子集的约简后进行约简集成,从而高效获取决策表的约简. Sun et al[13]以模糊邻域熵为启发式信息求取邻域多粒度粗糙集框架下的属性约简. 这些算法虽然采用不同的启发式信息求取信息表或者决策表的约简,但它们均有相同的结构,其本质均为保持信息表或决策表的某种性质不变为基础,求取能够代表所有属性集的约简属性. Zhao et al[14]分析了各种约简的特征与模式,给出了属性约简的通用定义,将各类属性约简进行统一描述,形成通用框架. Jia et al[15]分析现有约简特征,提出了结合数据及用户需求的通用属性约简定义并讨论现有约简在其一般定义框架下的特殊形态.

以上的约简是在信息表中属性互不关联、相互独立的基础上进行的. 然而在实际应用中,属性之间可能具有不同的关联特性,其中互补便是一种最简单的关联特征. 例如在中西医结合医疗诊断中,中医使用望闻问切等手段收集特征对病人进行刻画,西医利用精密仪器检测对病人进行数值化描述,因此,从疾病的治疗来说,中西医结合的诊疗方式一般而言优于任何单一的治疗技术. 在已知两个视角相结合可以保持性质的前提下,从给定的诊断信息表中求取既有中医诊疗属性又有西医精确描述的属性约简,具有重要的现实意义. 而目前从数据集出发不考虑属性集之间关联信息的约简算法没有考虑这方面的特征,这也造成最后求取的约简属性集虽然满足了性质要求但属性分布却不符合用户需求的缺点,因此需要对此类的约简特征进行建模. 本文考虑初始属性集描述性质之间的互补特征,提出属性补集对以及约简补集对的概念,给出约简补集对的一般求取框架,最后通过一个实例说明本文所提约简补集对及算法在实践中的可行性及有效性.

## 1 基础知识

本节介绍划分信息表、粗糙集以及划分知识度量等基本概念.

**1.1 信息表和粗糙集** 给定一个四元组 $T=\langle U, At, V, f\rangle$,其中,非空有限集 $U$ 为论域;$At$ 为非空属性集;值域 $V=\bigcup_{a\in At}V_a$,$V_a$ 为 $a$ 的值域;$f: U\times At\to V$ 为信息函数,表示论域中对象在属性下的取值. 通常将 $T=\langle U, At, V, f\rangle$ 称为信息表,简记为 $T=\langle U, At\rangle$,$f(a,u)$ 简记为 $a(u)$. 若属性集 $At=C\cup\{d\}$,$C$ 为条件属性集,$d$ 为决策属性,则称 $DT=\langle U, C\cup\{d\}, V, f\rangle$ 为决策表.

信息表中任意属性子集 $P\subseteq At$ 决定等价关系:

$$ind(P)=\{(u,v)|a(u)=a(v), a\in P\}$$

商集 $U/ind(P)=\left\{[u]_P\middle|u\in U\right\}$ 构成属性集上的划分,简记为 $U/P$,并称为划分知识 $P$ 中的等价类满足 $[u]_P\cap[v]_P=\varnothing\left([u]_P\neq[v]_P\right)$ 且 $\bigcup_{u\in U}[u]_P=U$.

粗糙集是一种处理不精确不确定知识的工具. 约简是粗糙集的核心内容. 约简的本质是保持系统某种性质不变的最小属性子集,而粗糙集便利用属性子集对应的划分知识建立的近似空间

对概念或者知识进行近似描述，识别新概念，获取新认知.

**定义 1**[1]　给定信息表 $T=\langle U, At, V, f\rangle$，属性子集 $P$ 对概念 $X\subseteq U$ 进行近似刻画，$X$ 在知识 $P$ 下的下近似和上近似分别定义为：

$$\underline{P}(X)=\left\{u\middle|[u]_P\subseteq X\right\}$$

和

$$\overline{P}(X)=\left\{u\middle|[u]_P\cap X\neq\varnothing\right\}$$

若 $\underline{P}(X)=\overline{P}(X)$，则知识 $P$ 可精确描述概念 $X$；否则称 $X$ 被知识 $P$ 粗糙描述.

Zhao et al[14]总结各类约简的基本形态，给出了信息表约简的一般定义.

**定义 2**[14]　给定一个信息表 $T=\langle U, At, V, f\rangle$，考虑信息表上的某种性质 $P$ 且可被一个评价函数 $e: 2^{At}\to(L, \preceq)$ 所表达. 若属性子集 $R\subseteq A\subseteq At$ 满足条件：

(1) $e(A)\preceq e(R)$；

(2) $\forall a\in R, \neg\left(e(A)\preceq e\left(R-\{a\}\right)\right)$

则称 $R$ 是 $A$ 的一个约简，若 $A=At$，则称 $R$ 是信息表的一个约简.

评价函数 $e$ 是从属性子集到一个偏序集 $L$ 的映射，一般要求偏序集 $L$ 上的偏序与属性子集上的集合包含关系的偏序是单调的，即：若 $A\subseteq B$，有 $e(A)\preceq e(B)$.

验证属性集是否满足性质 $P$ 是一个繁琐复杂的过程，因此运用评价函数 $e$ 将性质数量化表达，以便于知识获取的计算.

**1.2　划分知识及度量**　信息表 $T$ 中的两个任意属性子集 $P, Q\subseteq At$，它们对应的知识分别为 $U/P=\left\{[u]_P\middle|u\in U\right\}$ 和 $U/Q=\left\{[u]_Q\middle|u\in U\right\}$. 若对任意的 $u\in U$ 均有 $[u]_P\subseteq[u]_Q$，则称知识 $P$ 是知识 $Q$ 的细化，记为 $U/P\preceq U/Q$. 若 $U/P\preceq U/Q$ 且 $U/P\neq U/Q$，则称知识 $P$ 是知识 $Q$ 的严格细化.

在信息表中，$\mathrm{K}=\left\{U/P\middle|P\subseteq At\right\}$ 中存在两个特殊的划分：最细划分记为 $U/At$，最粗划分记为 $U/\varnothing=\{U\}$. 记 $\varepsilon=U/At$ 和 $\delta=U/\varnothing$.

信息表中的属性决定对应的划分知识，属性具有的某种性质或者属性之间的一些关联可以数量化为划分知识的不确定性度量. 近似精度、熵、知识粒度、条件熵和知识距离是最常用的四种度量[1,7−18]，前两个仅刻画单个划分知识的性质，而后两个可用以刻画两个属性或属性集之间的关联性.

划分知识 $P$ 下粗糙集的近似精度[1]：

$$\rho_P(X)=\frac{\underline{P}(X)}{\overline{P}(X)}\tag{1}$$

划分知识 $U/ind(P)=\left\{X_1, X_2, \cdots, X_M\right\}$ 的熵与知识粒度[8]分别为：

$$H(P)=-\sum_{i=1}^{M}\frac{\left|X_i\right|}{|U|}\lg\frac{\left|X_i\right|}{|U|}\tag{2}$$

和

$$G(P)=-\sum_{i=1}^{M}\frac{\left|X_i\right|}{|U|}\lg\frac{1}{\left|X_i\right|}\tag{3}$$

设两个划分知识 $U/ind(P)=\left\{X_1, X_2, \cdots, X_M\right\}$ 和 $U/ind(Q)=\left\{Y_1, Y_2, \cdots, Y_L\right\}$，则条件熵[11−13]为：

$$H(Q|P)=-\sum_{i=1}^{M}\frac{\left|X_i\right|}{|U|}\sum_{j=1}^{L}\frac{\left|X_i\cap Y_j\right|}{\left|X_i\right|}\lg\frac{\left|X_i\cap Y_j\right|}{\left|X_i\right|}\tag{4}$$

知识距离[18]为：

$$d(P,Q)=\sum_{i=1}^{M}\frac{\left|X_i\right|}{|U|}\sum_{j=1}^{L}\frac{\left|X_i\oplus Y_j\right|}{|U|}\tag{5}$$

以上几种划分度量对于刻画信息表中知识获取的性质具有非常重要的意义，在经典的粗糙集约简中常常作为启发式约简的信息，用以求取整个信息表或决策表的约简.

## 2　补集对及约简补集对

本节给出信息表中补集对以及约简补集对的定义.

**定义 3**　给定信息表 $T=\langle U, At, V, f\rangle$ 和两个非空子集 $A, B\subseteq At$ 且 $A\cap B=\varnothing$. 若属性子集 $A$ 和 $B$ 均不满足性质 $P$，$A\cup B$ 满足性质 $P$，则称属性子集 $A$ 和 $B$ 是信息表中关于性质 $P$ 的一组补集对，记为 $A\complement_P B$.

信息表中存在许多组满足性质 $P$ 的补集对. 当属性集 $At$ 满足性质 $P$ 时，则可能有许多个属性

子集对是性质 $P$ 的补集对. 同时,这些补集对中的属性有的是非必要的,运用这些非必要属性描述对象或者知识获取比较耗时而且会使模型变得复杂. 由此,提出以下约简补集对的概念.

**定义 4** 给定一个信息表 $T=\langle U, At, V, f\rangle$. 考虑信息表上的某种性质 $P$ 可被一个评价函数 $e$: $2^{At}\to(L,\preceq)$ 表达. 设属性子集对 $A$ 和 $B$ 是一组性质 $P$ 的补集对,若属性子集 $A'\subseteq A\subseteq At$, $B'\subseteq B\subseteq At$ 满足以下条件:

(1) $e(A\cup B)\preceq e(A'\cup B')$;

(2) $\forall a\in A', b\in B'$,

$$\neg\left(\begin{array}{l}\left(e(A\cup B)\preceq e\left((A'-\{a\})\cup B'\right)\right)\vee\\ \quad\left(e(A\cup B)\preceq e\left(A'\cup(B'-\{b\})\right)\right)\end{array}\right)$$

则称 $A'$ 和 $B'$ 是性质 P 的一个约简补集对. 记为 $A'\tilde{C}_P B'$.

约简补集对与经典的约简集合之间有一定的联系和区别. 二者都是保持某种性质的极小集合,但对于经典约简来讲,只要其与对应超集之间具有相同性质 $P$,其中任何属性都是使得其保持性质必要的属性. 而约简补集对限制两个集合分属不同的属性集补集对,可以保证从不同的补集对中选择极小属性集对,使之满足性质 $P$.

当性质 $P$ 被赋予特殊的意义并选择合适的评价函数或者度量时,便可以对信息表进行约简,为进一步的知识获取做准备.

## 3 几类基本的约简补集对的定义

本节对上节中的性质 $P$ 赋予实际含义,同时选择合适的度量对性质进行描述,得到三类基本的约简补集对.

**3.1 划分粒结构约简补集对** 信息表 $T=\langle U, At, V, f\rangle$ 的任意属性子集 $R\subseteq At$ 下的可分辨关系诱导论域上的划分 $U/R$. $U/R$ 划分块中的任意两个对象在属性 $R$ 下不可区分.

设 $\Pi(U)$ 是 $U$ 上的所有划分构成的集合, $\pi_1=\{X_1, X_2, \cdots, X_M\}$, $\pi_2=\{Y_1, Y_2, \cdots, Y_L\}$ 是 $\Pi(U)$ 上的两个划分,若 $\forall X_i\in\pi_1, \exists Y_j\in\pi_2$, 使得 $X_i\subseteq Y_j$ 成立,则称 $\pi_1$ 细于 $\pi_2$,记为 $\pi_1\preceq\pi_2$. 显然 $\Pi(U)$ 上的细化关系 $\preceq$ 满足自反性、反对称性和传递性. 因此 $(\Pi(U),\preceq)$ 是一个偏序集.

划分知识的熵和粒度度量在 $(\Pi(U),\preceq)$ 上满足单调性的两个度量. 令性质 $P$ 对应于当前信息表能表示的最细的划分粒结构 $\varepsilon$,这样给出划分粒结构约简补集对的定义如下:

**定义 5** 给定一个信息表 $T=\langle U, At, V, f\rangle$, 考虑信息表上最细划分粒结构和属性集对应的信息熵度量 $E$: $2^{At}\to(\Re,\leqslant)$. 设属性子集对 $A$ 和 $B$ 的划分知识 $\varepsilon\prec U/A$, $\varepsilon\prec U/B$ 且 $U/(A\cup B)=\varepsilon$,若属性子集 $A'\subseteq A$, $B'\subseteq B$ 满足以下条件:

(1) $E(A\cup B)\leqslant E(A'\cup B')$;

(2) $\forall a\in A', b\in B'$,

$$\neg\left(\begin{array}{l}\left(E(A\cup B)\leqslant E\left((A'-\{a\})\cup B'\right)\right)\vee\\ \quad\left(E(A\cup B)\leqslant E\left(A'\cup(B'-\{b\})\right)\right)\end{array}\right)$$

则称 $A'$ 和 $B'$ 是粒结构的一组约简补集对. 记为 $A'\tilde{C}_S B'$.

因属性集的包含关系与粒结构的细化之间具有单调性质,以上定义条件中的"$\leqslant$"也可为"$=$",并不影响所求结果. 由于划分知识上的信息熵与知识粒度具有互补关系,因此在以上定义中的"$\leqslant$"换成"$=$"亦可得到粒结构约简补集对在粒度度量下的定义,二者的意义相同.

**3.2 概念描述约简补集对** 给定信息表 $T=\langle U, At, V, f\rangle$,设 $X\subseteq U$ 是论域上的一个概念,根据粗糙集定义知,任意属性集 $P\subseteq At$ 可以粗糙描述 $X$ 为 $(\underline{P}(X),\overline{P}(X))$,由粗糙集的性质可知,概念 $X$ 在 $At$ 下可获得最大近似精度 $\rho_{At}(X)=\dfrac{\underline{At}(X)}{\overline{At}(X)}$. 且近似精度与属性集的包含关系具有单调性,即 $A\subseteq B\subseteq At$, $\rho_A(X)\leqslant\rho_B(X)$.

考虑性质 $P$ 为获得信息表示下最大近似精度 $\rho_{At}(X)$. 给出概念描述约简对的定义.

**定义 6** 给定一个信息表 $T=\langle U, At, V, f\rangle$ 和一个概念子集 $X\subseteq U$. 考虑概念 $X$ 获得最细描述精度的性质和属性集对应的精度度量 $\rho$: $2^{At}\to(\Re,\leqslant)$. 设属性子集对 $A$ 和 $B$ 描述概念 $X$ 的近似精度 $\rho_A(X)<\rho_{At}(X)$, $\rho_B(X)<\rho_{At}(X)$ 且 $\rho_{A\cup B}(X)\geqslant\rho_{At}(X)$. 若属性子集 $A'\subseteq A, B'\subseteq B$

满足以下条件：

(1) $\rho_{A\cup B}(X)\leqslant\rho_{A'\cup B'}(X)$；

(2) $\forall a\in A', b\in B'$，

$$\neg\left(\begin{array}{l}\left(\rho_{A\cup B}(X)\leqslant\rho_{A'-\{a\}\cup B}(X)\right)\\ \quad\vee\left(\rho_{A\cup B}(X)\leqslant\rho_{A'\cup B'-\{b\}}(X)\right)\end{array}\right)$$

则称$A'$和$B'$是关于概念$X$描述的一组约简补集对，记为$A'\tilde{\mathcal{C}}^{X}_{Acc}B'$.

当精度度量与属性子集的包含关系单调时，上述定义中的“$\leqslant$”可为“$=$”，并不影响所求结果. 在决策理论粗糙集等其他带容忍机制的粗糙集下，由于精度度量不再具有单调性，因此为保证模型的广泛性，这里仍然采用“$\leqslant$”.

**3.3　分类描述约简补集对**　给定决策表$DT=\langle U, C\cup\{d\}, V, f\rangle, U/\{d\}=\{Y_1, Y_2, \cdots, Y_s\}$是一个决策分类. 条件属性子集$R\subseteq At$可为决策属性分类提供一定的信息，由条件熵的性质可知，决策知识$U/\{d\}$在$At$条件下可获得最小条件熵，即运用所有属性可以达到对决策分类信息一致描述的最小值，而且条件信息熵与属性集的包含关系具有单调性，即：

$$A\subseteq B\subseteq At,$$
$$H(d|A)\geqslant H(d|B)\geqslant H(d|At)$$

根据信息熵所描述的分类一致性描述的特点及度量，可以给出关于分类描述约简对的定义.

**定义7**　给定决策表$X\subseteq U$. 考虑决策分类$d$获得最为一致描述的性质和属性集对应的条件熵$H_c: 2^{At}\to(\mathfrak{R},\leqslant)$. 设属性子集对$A$和$B$条件下的决策分类$d$的描述条件熵$H(d|A)>H(d|At)$，$H(d|B)>H(d|At)$且$H(d|A\cup B)\leqslant H(d|At)$. 若属性子集$A'\subseteq A, B'\subseteq B$满足以下条件：

(1) $H(d|A'\cup B')\leqslant H(d|A\cup B)$；

(2) $\forall a\in A', b\in B'$，

$$\neg\left(\begin{array}{l}\left(H(d|(A'-\{a\})\cup B')\leqslant H(A\cup B)\right)\vee\\ \quad\left(H(d|A'\cup(B'-\{b\}))\leqslant H(A\cup B)\right)\end{array}\right)$$

则称$A'$和$B'$是关于决策$d$的一组分类描述约简补集对，记为$A'\tilde{\mathcal{C}}^{d}_{\text{infor}}B'$.

条件熵可用来表达条件属性集对决策类中各个概念描述的一致程度. 约简补集对既能保证决策描述最大一致性，又可满足其来自不同属性子集的特性. 从粒计算的角度来讲，约简子集对中的子集分别来自不同的视角，这种限制性约简为多视角学习提供了一个有效的分析工具.

本节提供了三种基本的约简补集对的定义. 由于信息表中具有多种多样的性质，而这些性质可以由不同的度量进行分析与刻画，这样性质与度量的组合便可以产生多个约简补集对的定义. 在实际应用中，可根据实际问题的需要，选择与问题求解相关的性质$P$和适合描述性质的度量获得相应的约简补集对.

## 4　约简补集对的通用求解算法

当约简补集对满足不同的性质以及选择不同的度量进行分析时，便会产生不同的约简补集对. 本节给出约简补集对的通用求解算法. 该算法以求取分类描述约简补集对为例，其他类似的算法进行相应的性质与度量的改变亦可得到相应的结果. 如涉及高效的启发式约简方法需依据相关度量设定属性的内部、外部重要度，以属性重要度为启发式信息进行属性的选择和去除，得到最终约简[9,12−13]. 这里暂且不考虑算法的高效性，仅根据定义给出一般的求解方法。

**算法1　决策表的约简补集对求解算法**

输入：决策表$DT=\langle U, C\cup\{d\}, V, f\rangle$，属性补集对$A, B\subseteq At$，$e: 2^{At}\to(\mathfrak{R},\leqslant)$关于属性包含单调.

输出：从$A$到$B$的一个约简子集对$(A', B')$.

初始化：$A'=A, B'=\varnothing$

Step 1：

While $|B|\neq\varnothing$ $e(A'\cup B')>e(A\cup B)$

计算$\overline{b}=\arg\min\limits_{b\in B}(A'\cup\{b\}), b\in B$；

% 选取属性为当前最重要属性，可根据属性重要度计算

If $e(A'\cup B'\cup\{\overline{b}\})<e(A'\cup B')$

$B'=B'\cup\{\overline{b}\}$；

$B=B-\{\overline{b}\}$；

Endif

End

% 选择属性集$B$中的元素使之与$A'$互补满足度量极值条件

Step 2：

For each $b\in B'$

If $e(A'\cup B'-\{b\})\leqslant e(A\cup B)$

$B'=B'-\{b\}$

Endif

Endfor

% 删去上一步中$B'$中非必要属性

Step 3:

For each $a\in A'$

If $e(A'-\{a\}\cup B')\leqslant e(A\cup B)$

$A'=A'-\{a\}$;

Endif

Endfor

% 删去$A'$中非必要属性.

上述算法将输入条件改为一般信息表,便可获得信息表的粒结构约简补集对和概念约简补集对.由于以上算法中度量及性质均未定,可根据实际需要确定性质选择度量进行计算.此外,由于满足约简补集对定义的属性集对不唯一,上述算法是从属性集$A$出发求取约简补集对,若从属性集$B$出发则可以获得不同的约简补集对;同时,在算法求解过程中可根据属性的重要度或者其他外在评价标准,以不同次序选择属性,相应地也会获得不同的约简补集对结果.在实际应用中,将出发属性集视为主视角求取以其为主的约简补集对,可以使约简补集对具有更丰富的含义.

在删除属性的过程中,属性删除顺序不同也会导致约简补集对的结果不同,所以可根据属性对应的测试复杂性或者代价不同,根据实际需求按照其属性的代价排序进行删除,从而得到代价最小或者操作简便的约简补集.这里的通用算法仅给出一种求解方案,属原型算法,应用中可结合实际需求对属性选择及去除进行特殊限定,个性化求解.

## 5 一个中西医结合医疗实例

众所周知,中西医结合医疗技术无论对非典型性肺炎还是对新型冠状病毒导致的肺炎都有很好的医疗效果,而单独中医或者西医对这类疾病的治疗均不能达到二者结合的效果,因此在特殊肺炎的诊治中,中医与西医的完美互补使这类疾病在我国可以得到快速有效的控制.表1给出了一个中医和西医对患者的判断信息表,从该信息表出发求取中西医结合约简补集对,能使医疗人员可以从极小属性集对的特征出发,对疾病进行准确把握.本例来自临床诊断的几个病例,仅用来对本文所提概念以及算法进行说明.

**表1 一个发热病人诊断表达信息系统**

**Table 1 An information system for fever patient diagnosis system**

| | 脉象 | 听诊 | 体质 | 咳嗽 | X光像 | 血沉 | 诊断 |
|---|---|---|---|---|---|---|---|
| $u_1$ | 沉 | 水泡 | 阳虚 | 剧烈 | 片状 | 正常 | 肺炎 |
| $u_2$ | 迟 | 水泡 | 阳虚 | 剧烈 | 片状 | 正常 | 肺炎 |
| $u_3$ | 浮 | 干鸣 | 阴虚 | 轻微 | 片状 | 快 | 肺炎 |
| $u_4$ | 沉 | 干鸣 | 阳虚 | 中度 | 片状 | 正常 | 肺炎 |
| $u_5$ | 沉 | 啰音 | 阳虚 | 轻微 | 片状 | 正常 | 肺炎 |
| $u_6$ | 浮 | 正常 | 痰湿 | 轻微 | 索条状 | 正常 | 肺炎 |
| $u_7$ | 浮 | 啰音 | 阴虚 | 剧烈 | 空洞 | 快 | 肺结核 |
| $u_8$ | 浮 | 啰音 | 阴虚 | 轻微 | 索条状 | 正常 | 肺结核 |
| $u_9$ | 浮 | 干鸣 | 阴虚 | 轻微 | 点状 | 快 | 肺结核 |
| $u_{10}$ | 迟 | 干鸣 | 痰湿 | 中度 | 片状 | 快 | 肺结核 |
| $u_{11}$ | 浮 | 干鸣 | 阴虚 | 轻微 | 点状 | 正常 | 肺炎 |
| $u_{12}$ | 沉 | 啰音 | 阴虚 | 剧烈 | 空洞 | 快 | 肺结核 |

表1中属性集$\{$脉象,听诊,体质,咳嗽$\}$为中医诊疗属性,$\{$X光像,血沉$\}$为西医精密仪器检测属性.从给定决策表中求从中医诊疗出发的中西医结合约简补集对.

表1是一个决策表,这里令$e(A)=H(d|A)$,通过计算可得$e(A)\neq e(At)$,$e(B)\neq e(At)$,而$e(A\cup B)=e(C)$.通过上一节给出的算法求取约简补集对.初始化$A'=A$,$B'=\varnothing$.根据算法1,其求解步骤如下:

Step 1. 顺序选择中的属性进行度量运算,根据算法规则求得$B'=\{$X光像,血沉$\}$.

Step 2. 由于$e(A'\cup(B'-\{$X光像$\}))>e(At)$且$e(A'\cup(B'-\{$血沉$\}))>e(At)$,因此$B'=\{$X光像,血沉$\}$中任意属性均不可去除.

Step 3. 为简便起见,这里以属性在表中出现的顺序对属性进行一一测试计算.经计算可知:

$e((A'-\{$脉象$\})\cup B')\leqslant e(A'\cup B')$,脉象可去除$A'=A'-\{$脉象$\}$;

$e((A'-\{$听诊$\})\cup B')\leqslant e(A'\cup B')$,听诊可去除$A'=A'-\{$听诊$\}$;

$e\left(\left(A'-\{\text{体质}\}\right)\cup B'\right)>e\left(A'\cup B'\right)$，体质不可去除$A'$不更新；

$e\left(\left(A'-\{\text{咳嗽}\}\right)\cup B'\right)\leqslant e\left(A'\cup B'\right)$，咳嗽可去除$A'=A'-\{\text{咳嗽}\}$.

因此可知$A'=\{\text{体质}\}$.

这样便可得到$A'=\{\text{体质}\}$，$B'=\{\text{X成像，血沉}\}$是约简补集对．这个结果表明运用中医体质再查血沉对于区分肺炎和肺结核具有重要的意义．这里约简补集对根据给定数据计算，当数据分布变化或者数据量增多，其约简补集对可能不同．

本文中约简补集对限定约简属性集的来源使约简对中的属性分属不同的属性集合．经典的约简算法中，约简求取完全根据数据分布特点，不考虑属性及属性集之间的内在关系，所求约简中属性分布较为随意．值得注意的是，在满足相同性质的条件下，约简属性对的并集一定是一个经典约简，但并非所有的经典约简都满足来自限定补集对的要求．因此，本文所提出的约简补集对是对经典约简集进行满足某种条件的筛选．

## 6　结　论

约简是粗糙集理论区分于其他数据挖掘的最重要的内容，传统约简针对整个信息表依据数据分布进行求解，不考虑属性及属性集之间的互补特性．本文从互补视角出发求取信息表的约简补集对，使不同视角均有属性参与数据挖掘过程，能平衡不同视角属性的参与度．本文提取约简补集对的一般共同特征，从给定的集合出发，结合特定性质$P$及相关特征度量给出了约简补集对的一般定义和三种基本的约简补集对的定义，最后通过一个中西医结合诊疗实例说明约简补集对在多视角求解问题中的重要作用．本文提出的约简补集对定义可为信息表带限制条件的属性约简提供新的研究思路．

### 参 考 文 献


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

[2] Yao Y Y. Two views of the theory of rough sets in finite universes. International Journal of Approximate Reasoning, 1996, 15(4): 291－317.

[3] Skowron A, Rauszer C. The discernibility matrices and functions in information systems//Intelligent Decision Support. Springer Berlin Heidelberg, 1992: 331－362.

[4] Yao Y Y, Zhao Y. Discernibility matrix simplification for constructing attribute reducts. Information Sciences, 2009, 179(7): 867－882.

[5] 杨明，杨萍．差别矩阵浓缩及其属性约简求解方法．计算机科学，2006，33(9)：181－183，269.（Yang M, Yang P. Discernibility matrix enriching and computation for attributes reduction. Computer Science, 2006, 33(9): 181－183, 269.）

[6] 冯琴荣，胡兢丹．利用浓缩布尔矩阵重排技术求所有约简．控制与决策，https://kns.cnki.net/KCMS/detail/21.1124. tp. 20200114.1555.013. html, 2020－01－15.

[7] Miao D Q, Zhao Y, Yao Y Y, et al. Relative reducts in consistent and inconsistent decision tables of the Pawlak rough set model. Information Sciences, 2009, 179(24): 4140－4150.

[8] 苗夺谦，胡桂荣．知识约简的一种启发式算法．计算机研究与发展，1999，36(6)：681－684.（Miao D Q, Hu G R. A heuristic algorithm for reduction of knowledge. Journal of Computer Research and Development, 1999, 36(6): 681－684.）

[9] 王国胤，于洪，杨大春．基于条件信息熵的决策表约简．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. Chinese Journal of Computers, 2002, 25(7): 759－766.）

[10] 胡峰，王国胤．属性序下的快速约简算法．计算机学报，2007，30(8)：1429－1435.（Hu F, Wang G Y. Quick reduction algorithm based on attribute order. Chinese Journal of Computers, 2007, 30(8): 1429－1435.）

[11] Qian Y H, Liang J Y, Pedrycz W, et al. Positive approximation: an accelerator for attribute reduction in rough set theory. Artificial Intelligence, 2010, 174(9－10): 597－618.

[12] Liang J Y, Wang F, Dang C Y, et al. An efficient rough feature selection algorithm with a multi-

granulation view. International Journal of Approximate Reasoning, 2012, 53(6): 912－926.

[13] Sun L, Wang L Y, Ding W P, et al. Feature selection using fuzzy neighborhood entropy-based uncertainty measures for fuzzy neighborhood multigranulation rough sets. IEEE Transactions on Fuzzy Systems, 2020, doi: 10.1109/TFUZZ.2020.2989098.

[14] Zhao Y, Luo F, Wong S K M, et al. A general definition of an attribute reduct//International Conference on Rough Sets and Knowledge Technology. Springer Berlin Heidelberg, 2007: 101－108.

[15] Jia X Y, Shang L, Zhou B, et al. Generalized attribute reduct in rough set theory. Knowledge-Based Systems, 2016, 91: 204－218.

[16] Yao Y Y, Zhao Y. Attribute reduction in decision-theoretic rough set models. Information Sciences, 2008, 178(17): 3356－3373.

[17] Yao Y Y, Zhao L Q. A measurement theory view on the granularity of partitions. Information Sciences, 2012, 213: 1－13.

[18] 王宝丽, 梁吉业. 信息系统中的知识距离与知识粗糙熵. 计算机科学, 2007, 34(3): 151－154. (Wang B L, Liang J Y. Knowledge distance and rough entropy in information systems. Computer Science, 2007, 34(3): 151－154.)

(责任编辑　杨可盛)