ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 4,Jul. 2020,pp. 630－641
DOI:10. 16337/j. 1004-9037. 2020. 04. 004

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

邻域互补信息度量及其启发式属性约简

陈　帅[1,2]，张贤勇[1,2]，唐玲玉[1,2]，姚岳松[1,2]

(1. 四川师范大学数学科学学院,成都,610066;2. 四川师范大学智能信息与量子信息研究所,成都,610066)

摘　要：信息熵体系是进行不确定刻画与近似推理的重要理论,已经被引入粗糙集进行数据分析与智能处理。经典的互补熵、互补条件熵和互补互信息能够刻画粗糙性与模糊性,该信息体系的拓展具有应用意义。本文基于邻域粗糙集,扩张构建邻域互补信息度量并研究其启发式属性约简。通过解析式模拟与信息粒替换,定义邻域互补熵、邻域互补条件熵和邻域互补互信息,得到系统方程、双界刻画和粒化非单调性;基于邻域互补互信息,提出非单调属性约简并设计启发式约简算法;采用决策表实例与UCI数据实验有效验证性质与算法。基于邻域扩张,相关信息度量与属性约简具有应用前景。

关键词：邻域粗糙集;信息论;邻域互补信息度量;不确定性;属性约简;粒计算

中图分类号：TP18　　**文献标志码：**A

Neighborhood Complementary Information Measures and Heuristic Attribute Reduction

CHEN Shuai[1,2], ZHANG Xianyong[1,2], TANG Lingyu[1,2], YAO Yuesong[1,2]

(1. School of Mathematical Sciences, Sichuan Normal University, Chengdu, 610066, China;2. Institute of Intelligent Information and Quantum Information, Sichuan Normal University, Chengdu, 610066, China)

Abstract: The information entropy system serves as a fundamental theory of uncertainty description and approximate reasoning, and it has been introduced into rough sets to implement data analyses and intelligence processing. Classical complementary entropy, conditional-entropy and mutual-information can effectively describe roughness and fuzziness, and their system expansion has application significance. In terms of neighborhood rough sets, neighborhood complementary information measures are extendedly constructed, and their heuristic attribute reduction is investigated. According to analytical simulation and granular replacement, neighborhood complementary entropy, conditional-entropy and mutual-information are defined, and their system equation, double bounds and granulation non-monotonicity are achieved. Based on the neighborhood complementary mutual-information, non-monotonic attribute reduction and its heuristic reduction algorithm are proposed. The validity of property and algorithm is verified by decision tables and data experiments. By virtue of neighborhood expansion, relevant information measures and attribute reduction have application prospects.

Key words: neighborhood rough set; information theory; neighborhood complementary information measure; uncertainty; attribute reduction; granular computing

---

基金项目：国家自然科学基金(61673258)资助项目;四川省科技基金(19YYJC2845)资助项目;四川省青年科技基金(2017JQ0046)资助项目。
收稿日期：2019-06-24;**修订日期：**2019-09-15

引　言

信息理论能够有效实施不确定表示与应用[1],已经被系统地引入粗糙集进行数据分析与智能处理。例如,文献[2-5]采用香农熵、粗糙熵和加权熵等体系进行不确定性刻画与约简算法启发。特别地,文献[6]基于互补机制提出互补信息体系,相关不确定性度量包括互补熵、互补条件熵和互补互信息,它们刻画了粗糙性与模糊性。继而,文献[7]以互补条件熵为启发式信息,构建基于正域的属性约简算法;文献[8]基于三层粒度结构,构建三支加权互补熵体系。互补信息度量采用的互补刻画$p(1-p)$区别于香农信息度量的对数刻画$-p\log_2 p$,但两者具有一些共同特征(如函数上凸非单调),进而也存在一定相似性(如度量系统关系)。对比于香农信息体系,互补信息系统能够有效刻画粗糙集的模糊性[6]。可见,互补信息度量具有独特的不确定性刻画优势,但其研究还相对较少,相关的深入与拓展具有创新价值。

粗糙集具有双向逼近认知,能够进行知识约简与特征选择,广泛应用于数据挖掘、机器学习和人工智能等领域。传统粗糙集主要采用等价关系与知识剖分,相关数值型数据处理往往需要离散化。邻域粗糙集引入邻域关系与覆盖结构,具有理论扩张性与应用鲁棒性,能够有效处理符号型数据、数值型数据乃至混合型数据。邻域粗糙集的不确定性度量与属性约简得到了广泛研究[9-15]。特别地,文献[16]建立基于对数函数$-\log_2 p$的信息体系(包括邻域熵、邻域条件熵和邻域互信息);文献[17]用邻域互信息进行特征选择与多标签学习;文献[18]提出邻域精度、邻域熵、信息粒度及相关粒化单调性;文献[19]建立一种模糊熵来度量邻域粗糙集的不确定性;文献[20]采用邻域熵及其发展度量进行离群点检测。总之,邻域粗糙集尚未涉及互补信息度量体系,相关的构建具有应用意义。

基于上述背景,本文拟将经典互补信息度量推广到邻域粗糙集,并研究相关的启发式属性约简。基于邻域扩张,具体构建邻域互补熵、邻域互补条件熵和邻域互补互信息,研究基本性质;基于邻域互补互信息及其粒化非单调性,设计属性约简及其启发算法。研究结果被决策表实例与UCI数据实验有效验证,并有利于基于邻域粗糙集的不确定性信息处理。

1　经典邻域互补信息度量与邻域系统

本节利用文献[6,8]与文献[16,21]分别回顾经典互补信息度量与邻域系统。

决策表为四元组$DT=(U,AT=C\cup D,V,f)$,U为有限论域,C,D分别为条件、决策属性集,$V=\bigcup_{a\in AT}V_a$(V_a是属性$a\in AT$的值域),信息函数$f:U\times AT\to V$具有$\forall x\in U,\forall a\in AT,f(x,a)\in V_a$。设条件属性子集$A\subseteq C$诱导知识划分$U/IND(A)=\{[x]_A^i|i=1,\cdots,n\}$,决策属性集$D$诱导决策分类};设$*^C=U-*$表示补集,例如$[x]_A^{iC}$,$D_j^C$。

定义 1[6,8]　$A\subseteq C$的互补熵、D相对于A的互补条件熵、A与D之间的互补互信息分别为

$$\begin{aligned}
H(A)&=\sum_{i=1}^{n}\frac{|[x]_A^i|}{|U|}\frac{|[x]_A^{iC}|}{|U|}=\sum_{i=1}^{n}\frac{|[x]_A^i|}{|U|}\left(1-\frac{|[x]_A^i|}{|U|}\right)\\
HC(D|A)&=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|[x]_A^i\cap D_j|}{|U|}\frac{|D_j^C-[x]_A^{iC}|}{|U|}=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|[x]_A^i\cap D_j|}{|U|}\frac{|[x]_A^i-D_j|}{|U|}\\
MI(D;A)&=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|[x]_A^i\cap D_j|}{|U|}\frac{|D_j^C\cap[x]_A^{iC}|}{|U|}=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|[x]_A^i\cap D_j|}{|U|}\left(1-\frac{|[x]_A^i\cup D_j|}{|U|}\right)
\end{aligned}\tag{1}$$

定理 1[6,8]　互补熵、互补条件熵、互补互信息满足如下系统方程

$$MI(D;A)=H(A)-HC(A|D)=H(D)-H(D|A) \tag{2}$$

式中

$$\begin{aligned} H(D)&=\sum_{j=1}^{m}\frac{|D_j|}{|U|}\frac{|D_j{}^C|}{|U|}=\sum_{j=1}^{m}\frac{|D_j|}{|U|}\left(1-\frac{|D_j|}{|U|}\right)\\ HC(A|D)&=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|[x]_A^i\cap D_j|}{|U|}\frac{|[x]_A^{i\,C}-D_j{}^C|}{|U|}=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|[x]_A^i\cap D_j|}{|U|}\frac{|D_j-[x]_A^i|}{|U|}\end{aligned} \tag{3}$$

定理 2[6,8] 若 $U/IND(A)\geqslant U/IND(B)$(即 $\forall[x]_B\in U/IND(B)$, $\exists[x]_A\in U/IND(A)$, s.t.$[x]_B\subseteq[x]_A$),则

$$H(A)\leqslant H(B),HC(D|A)\geqslant HC(D|B),MI(D;A)\leqslant MI(D;B) \tag{4}$$

定义1提供了经典互补信息体系,其涉及补集描述与概率形式(如 $p([x]_A^i)=\frac{|[x]_A^i|}{|U|}$, $p([x]_A^i\cap D_j)=\frac{|[x]_A^i\cap D_j|}{|U|}$),即其采用对称二次函数 $p(1-p)$ 替换经典的非对称对数函数 $-p\log_2 p$ 进行信息集成,从而成为一种新型不确定性度量,互补熵刻画条件知识结构的不确定性,互补条件熵与互补互信息描述条件结构与决策结构之间的信息关系,相关体系可以度量粗糙性与模糊性等不确定性[6]。事实上,$U/IND(A)$ 与 $U/IND(D)$ 具有关于等价划分的平等性,故定理1中出现的 $H(D),HC(A|D)$ 分别对称于定义1中的 $H(A),HC(D|A)$;进而,定理1提供5种互补信息度量的系统方程。定理2则表明互补信息系统具有粒化单调性,其中的粒化关系 $U/IND(A)\geqslant U/IND(B)$ 可由 $A\subseteq B$ 实现。

邻域粗糙集是经典粗糙集的拓展,其基础为邻域系统[16,21]。对决策表 DT,设 $A=\{c_1,\cdots,c_n\}\subseteq C$,则其对应距离函数:$d_A(x,y)=\left(\sum_{h=1}^{g}|f(x,c_h)-f(y,c_h)|^p\right)^{\frac{1}{p}}$。本文采用 $p=1$ 的Manhattan距离函数。加上半径参数 δ,则 x 的领域 $n_A^\delta(x)=\{y\in U|d_A(x,y)\leqslant\delta\}$ 可以诱导邻域关系 $NR_A=\{(x,y)\in U\times U|y\in n_A^\delta(x)\}$ 与邻域覆盖 $U/NR_A=\{n_A^\delta(x)|x\in U\}=\{n_A^\delta(x)_1,\cdots,n_A^\delta(x)_n\}$,其中 $n_A^\delta(x)_i(i=1,\cdots,n)$ 表示 n 个邻域。

定理 3[16,21] (1)若 $\delta_1\leqslant\delta_2$,则 $\forall x\in U$ 有 $n_A^{\delta_1}(x)\subseteq n_A^{\delta_2}(x)$。若 $\delta=0$,则 $n_A^\delta(x)=[x]_A$。(2)若 $A\subseteq B$,则 $\forall x\in U$ 有 $n_A^\delta(x)\supseteq n_B^\delta(x)$(此时也记 $U/NR_A\geqslant U/NR_B$)。

邻域依托其覆盖结构为邻域粗糙集奠定了粒计算基础。基于定理3,邻域具有参数单调性,$\delta=0$ 导致退化,即邻域粒退化到等价类且邻域粗糙集退化到经典粗糙集;此外,邻域还具有关于属性子集关系的单调性,这为相关粒化单调性奠定了基础。

2 邻域互补信息度量及其性质

经典互补信息度量[6,8]适用于经典粗糙集。针对扩张的邻域粗糙集,本节自然定义邻域互补信息度量,以实施度量扩张与拓展应用。下面主要针对决策表 DT 及其邻域覆盖 $U/NR_A=\{n_A^\delta(x)_i|i\in1,\cdots,n\}$ 与决策分类 $U/IND(D)=\{D_j|j\in1,\cdots,m\}$。

2.1 邻域互补信息度量

定义 2 $A\subseteq C$ 的邻域互补熵、D 关于 A 的邻域互补条件熵、A 与 D 之间的邻域互补互信息分别为

$$NH_\delta(A)=\sum_{i=1}^{n}\frac{|n_A^\delta(x)_i|}{|U|}\frac{|n_A^\delta(x)_i{}^C|}{|U|}=\sum_{i=1}^{n}\frac{|n_A^\delta(x)_i|}{|U|}\left(1-\frac{|n_A^\delta(x)_i|}{|U|}\right)$$
$$NHC_\delta(D|A)=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|D_j{}^C-n_A^\delta(x)_i{}^C|}{|U|}=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\left(\frac{|n_A^\delta(x)_i-D_j|}{|U|}\right)\tag{5}$$
$$NMI_\delta(D;A)=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|D_j{}^C\cap n_A^\delta(x)_i{}^C|}{|U|}=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\left(1-\frac{|n_A^\delta(x)_i\cup D_j|}{|U|}\right)$$

定理4　若$\delta=0$,则$NH_\delta(A)=H(A)$,$NHC_\delta(D|A)=HC(D|A)$,$NMI_\delta(D;A)=MI(D;A)$。

定义2的邻域互补信息度量模拟了定义1的经典互补信息度量,主要将知识划分$U/IND(A)$拓展替换为邻域覆盖U/NR_A(即等价类$[x]_A^i$换为邻域$n_A^\delta(x)_i$。这种基于解析式的拓展方案,比较自然也更为稳妥。由此,邻域互补信息度量具有扩张性(定理4)。3种邻域度量具有类似于经典互补熵的不确定性语义,但代替地使用邻域覆盖结构。为了有利于邻域覆盖结构的近似推理,它们具体采用邻域粒不重复计数机制,这区别于元素诱导的邻域可重复机制[16]。此外,定义2也提供了补集描述与等价本质两种形式。下面模拟确定$NH_\delta(A)$,$NHC_\delta(D|A)$的对称度量$NH_\delta(D)_A$,$NHC_\delta(A|D)$,并发展系统关系。

命题1　邻域互补熵具有等价"双和形式"

$$NH_\delta(A)=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|n_A^\delta(x)_i{}^C|}{|U|}=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\left(1-\frac{|n_A^\delta(x)_i|}{|U|}\right)\tag{6}$$

证明　由式(5)与$U/IND(D)$的剖分性有

$$NH_\delta(A)=\sum_{i=1}^{n}\frac{|n_A^\delta(x)_i|}{|U|}\frac{|n_A^\delta(x)_i{}^C|}{|U|}=\sum_{i=1}^{n}\frac{|n_A^\delta(x)_i\cap(D_1\cup\cdots\cup D_m)|}{|U|}\frac{|n_A^\delta(x)_i{}^C|}{|U|}=$$
$$\sum_{i=1}^{n}\frac{|(n_A^\delta(x)_i\cap D_1)\cup\cdots\cup(n_A^\delta(x)_i\cap D_m)|}{|U|}\frac{|n_A^\delta(x)_i{}^C|}{|U|}=\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|n_A^\delta(x)_i{}^C|}{|U|}$$

命题2　$H(D)\leqslant\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|D_j{}^C|}{|U|}$,若$\delta=0$时等号成立。

证明　由式(3)与U/NR_A的覆盖性,类似于命题1的证明过程有

$$H(D)=\sum_{j=1}^{m}\frac{|D_j|}{|U|}\frac{|D_j{}^C|}{|U|}\leqslant\sum_{j=1}^{m}\left(\sum_{i=1}^{n}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\right)\frac{|D_j{}^C|}{|U|}=$$
$$\sum_{i=1}^{n}\left(\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|n_A^\delta(x)_i{}^C|}{|U|}\right)=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|n_A^\delta(x)_i{}^C|}{|U|}$$

其中覆盖U/NR_A退化为划分$U/IND(A)$时(此时$\delta=0$),上述等号成立。

定理5　采用类似于$NH_\delta(A)$(式(6))与$NHC_\delta(D|A)$(式(5))的"双和形式",设置符号

$$NH_\delta(D)_A=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|D_j{}^C|}{|U|}=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\left(1-\frac{|D_j|}{|U|}\right)$$
$$NHC_\delta(A|D)=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|n_A^\delta(x)_i{}^C-D_j{}^C|}{|U|}=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|D_j-n_A^\delta(x)_i|}{|U|}\tag{7}$$

则邻域互补熵、邻域互补条件熵和邻域互补互信息满足如下系统方程

$$NMI_\delta(D;A)=NH_\delta(A)-NHC_\delta(A|D)=NH_\delta(D)_A-NHC_\delta(D|A)\tag{8}$$

证明　(1)注意到$n_A^\delta(x)_i{}^C$具有两剖分部分:$n_A^\delta(x)_i\cap D_j$,$n_A^\delta(x)_i{}^C-D_j{}^C$,由式(6,7)有

$$NH_\delta(A)=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|n_A^\delta(x)_i{}^C|}{|U|}=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|(n_A^\delta(x)_i{}^C\cap D_j{}^C)\cup(n_A^\delta(x)_i{}^C-D_j{}^C)|}{|U|}=$$
$$\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|n_A^\delta(x)_i{}^C\cap D_j{}^C|}{|U|}+\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|n_A^\delta(x)_i{}^C-D_j{}^C|}{|U|}\right)=$$
$$NMI_\delta(D;A)+NHC_\delta(A|D)$$

(2) 由式(6,7)有

$$NHC_\delta(D|A)=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|D_j{}^C-n_A^\delta(x)_i{}^C|}{|U|}=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|D_j{}^C-(n_A^\delta(x)_i{}^C\cap D_j{}^C)|}{|U|}=$$
$$\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|D_j{}^C|}{|U|}-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|n_A^\delta(x)_i{}^C\cap D_j{}^C|}{|U|}=$$
$$NH_\delta(D)_A-NMI_\delta(D;A)$$

命题1提供邻域互补熵$NH_\delta(A)$的“双和形式”。命题2涉及的“双和形式”类似且对称于$NH_\delta(A)$的,但其不同于且不小于$H(D)$,这是因为覆盖U/NR_A替换了划分$U/IND(A)$。由此,定理5提取命题2“双和形式”形成新符号$NH_\delta(D)_A$,其依赖于A从而区别于只依赖于D的$H(D)$;此外,定理5还提供了$NHC_\delta(A|D)$。可见,$NH_\delta(D)_A$与$NH_\delta(A)$具有在“双和层面"的粒交换性,而$NHC_\delta(A|D)$与$NHC_\delta(D|A)$具有集差换序。从而,定理5证明并表现了邻域互补信息度量的系统关系,其对应经典互补信息度量的系统关系(式(2))。由此,下述推论1补充了$NH_\delta(D)_A$,$NHC_\delta(A|D)$对于$H(D)$,$HC(A|D)$的扩张性,还揭示了$NMI_\delta(D;A)\neq H(D)-NHC_\delta(D|A)$的扩张特异性。由此定理所述0值条件容易检验。

推论1 (1)若$\delta=0$,则$NH_\delta(D)_A=H(D)$,$NHC_\delta(A|D)=HC(A|D)$。

(2) $NH_\delta(D)_A\geqslant H(D)$,$NMI(D;A)\geqslant H(D)-NHC_\delta(D|A)$,若$\delta=0$时等号成立。

定理6 $U/NR_A\preccurlyeq U/IND(D)$(即$\forall n_A^\delta(x)_i\in U/NR_A,\exists D_j\in U/IND(D),\text{s.t.},n_A^\delta(x)_i\subseteq D_j$)的充分必要条件是$NHC_\delta(D|A)=0$。

证明 (1)若$U/NR_A\preccurlyeq U/IND(D)$,则$\forall n_A^\delta(x)_i\in U/NR_A,\exists D_j\in U/IND(D)$有$n_A^\delta(x)_i\cap D_j=\varnothing$,$|n_A^\delta(x)_i\cap D_j|=0$或$n_A^\delta(x)_i\subseteq D_j$,$|n_A^\delta(x)_i-D_j|=0$,即有$|n_A^\delta(x)_i\cap D_j||n_A^\delta(x)_i-D_j|=0$。因此由式(5)有$NHC_\delta(D|A)=0$。(2)反设$NHC_\delta(D|A)=0$,但$U/NR_A\not\preccurlyeq U/IND(D)$。此时,$\exists i^*\in\{1,\cdots,n\},j^*\in\{1,\cdots,m\},\text{s.t.}\ \varnothing\subset n_A^\delta(x)_{i^*}\cap D_{j^*}\subset n_A^\delta(x)_{i^*}$,$|n_A^\delta(x)_{i^*}\cap D_{j^*}||n_A^\delta(x)_{i^*}-D_{j^*}|>0$。因此,由式(5)有$NHC_\delta(D|A)\geqslant\frac{|n_A^\delta(x)_{i^*}\cap D_{j^*}|}{|U|}\frac{|n_A^\delta(x)_{i^*}-D_{j^*}|}{|U|}>0$。该矛盾意味着充分性成立。

推论2 $U/NR_A\preccurlyeq U/IND(D)$等价于$NMI_\delta(D;A)=NH_\delta(D)_A$,$NH_\delta(A)=NHC_\delta(A|D)+NH_\delta(D)_A$。

定理7 (1) $NH_\delta(A)\in[0,|U|/4]$,且$U/NR_A=\{U\}\Rightarrow NH_\delta(A)=0$。

(2) $NHC_\delta(D|A)\in[0,|U|]$,且$U/NR_A\preccurlyeq U/IND(D)\Rightarrow NHC_\delta(D|A)=0$。

(3) $NMI_\delta(D;A)\in[0,|U|]$,且$U/NR_A=\{U\}\Rightarrow NMI_\delta(D;A)=0$。

证明 (1)$\forall i\in\{1,\cdots,n\}$,$\frac{|n_A^\delta(x)_i|}{|U|}(1-\frac{|n_A^\delta(x)_i|}{|U|})\leqslant\frac{1}{4}$,由式(5)有$NH_\delta(A)\leqslant\frac{n}{4}\leqslant\frac{|U|}{4}$。$NH_\delta(A)\geqslant0$是显然的,且可由$U/NR_A=\{U\}$取得。

(2) 由式(5)有

$$NHC_\delta(D|A)=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\frac{|n_A^\delta(x)_i-D_j|}{|U|}\leqslant\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|D_j|}{|U|}\frac{|n_A^\delta(x)_i|}{|U|}=$$
$$\sum_{i=1}^{n}\left(\frac{|n_A^\delta(x)_i|}{|U|}\sum_{j=1}^{m}\frac{|D_j|}{|U|}\right)=\sum_{i=1}^{n}\frac{|n_A^\delta(x)_i|}{|U|}\leqslant n\times 1\leqslant|U|$$
$$NMI_\delta(D;A)=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cap D_j|}{|U|}\left(1-\frac{|n_A^\delta(x)_i\cup D_j|}{|U|}\right)\leqslant\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|n_A^\delta(x)_i\cup D_j|}{|U|}=$$
$$\sum_{i=1}^{n}\frac{|n_A^\delta(x)_i|}{|U|}\leqslant n\times 1\leqslant|U| \tag{9}$$

由此定理所述0值条件容易检验。

推论3　$H(A),HC(D|A),MI(D;A)$都具有双界范围$[0,1]$。

定理6(及推论2)提供了粒化条件$U/NR_A\leqslant U/IND(D)$的信息描述,这有利于覆盖粒化的依赖推理。定理7提供了$NH_\delta(A),NHC_\delta(D|A),NMI_\delta(D;A)$的双界及其下确界0取得情形。基于证明式(9),$NHC_\delta(D|A),NMI_\delta(D;A)$通过"双和形式"放缩,从而获得上界$|U|$;类似地,$NH_\delta(A)$也可以采用"双和形式"(式(6))得到相同上界$|U|$。若覆盖$U/NR_A$退化为划分$U/IND(A)$时(此时$\delta=0$),式(9)中$\sum_{i=1}^{n}\frac{|n_A^\delta(x)_i|}{|U|}=1$,则上界$|U|$皆可以降低到1,故推论3描述了退化的经典互补度量情形。此外,$NH_\delta(A)$采用"单和形式"及抛物函数$p(1-p)$最大值1/4来提供上界$|U|/4$,其通常小于上述上界$|U|$。进而,相关的更小上界乃至上确界值得探讨。

2.2　邻域互补信息度量的粒化非单调性

不确定性度量的粒化单调性或非单调性是信息应用的一个重要特性[2,3,15],决定着属性约简的定义构造与算法启发。定理2表明,经典互补信息度量具有粒化单调性。本小节阐述邻域互补信息度量基于扩张变异的粒化非单调性。下面首先通过一个实例来计算信息值并提供非单调事实。

表1　实例决策表

Table 1　Instance decision table

U	c_1	c_2	c_3	c_4	c_5	D
x_1	0.8	1.0	0.7	0.5	0.4	1
x_2	0.9	0.8	0.6	0.5	0.9	1
x_3	0.6	1.0	0.4	0.8	1.0	2
x_4	1.0	0.4	0.8	0.7	0.5	1
x_5	0.7	0.6	0.3	0.2	0.4	1
x_6	0.5	0.2	0.1	0.9	0.3	2
x_7	0.9	0.3	0.7	0.5	0.6	2

例1　决策表DT如表1,其中$U/IND(D)=\{D_1=\{x_1,x_2,x_4,x_5\},D_2=\{x_3,x_6,x_7\}\}$。

基于Manhattan距离与半径$\delta=0.5$,可以构建邻域体系。为了研究粒化,下面聚焦自然属性增链:$\{c_1\}\subset\{c_1,c_2\}\subset\{c_1,c_2,c_3\}\subset\{c_1,c_2,c_3,c_4\}\subset\{c_1,c_2,c_3,c_4,c_5\}$(链元属性集用$A$统一表示)。表2提供相关邻域及覆盖。再由式(5,7),表3提供所有5种邻域互补信息度量值。作为例子,链元$\{c_1,c_2,c_3\}$的前3个度量计算过程为:

(1) $NH_\delta(\{c_1,c_2,c_3\})=\frac{3}{7}\times\frac{4}{7}+\frac{2}{7}\times\frac{5}{7}+\frac{2}{7}\times\frac{5}{7}+\frac{2}{7}\times\frac{5}{7}+\frac{1}{7}\times\frac{6}{7}+\frac{1}{7}\times\frac{6}{7}=1.1020$;

(2) $NHC_\delta(\{c_1,c_2,c_3\})=\frac{2}{7}\times\frac{1}{7}+\frac{1}{7}\times\frac{2}{7}+\frac{1}{7}\times\frac{1}{7}+\frac{1}{7}\times\frac{1}{7}+\frac{1}{7}\times\frac{1}{7}+\frac{1}{7}\times\frac{1}{7}=0.1633$;

(3) $NMI_\delta(\{c_1,c_2,c_3\})=\frac{2}{7}\times\frac{2}{7}+\frac{1}{7}\times\frac{2}{7}+\frac{2}{7}\times\frac{3}{7}+\frac{1}{7}\times\frac{2}{7}+\frac{1}{7}\times\frac{3}{7}+\frac{1}{7}\times\frac{2}{7}+\frac{1}{7}\times\frac{3}{7}+\frac{1}{7}\times\frac{3}{7}+\frac{1}{7}\times\frac{4}{7}=0.5918$。

表2 基于属性增链的邻域及覆盖

Table 2 Neighborhood and coverage based on attribute chaining

邻域	$\{c_1\}$	$\{c_1,c_2\}$	$\{c_1,c_2,c_3\}$	$\{c_1,c_2,c_3,c_4\}$	$\{c_1,c_2,c_3,c_4,c_5\}$
$n_A^\delta(x_1)$	$\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$	$\{x_1,x_2,x_3,x_5\}$	$\{x_1,x_2,x_3\}$	$\{x_1,x_2\}$	$\{x_1\}$
$n_A^\delta(x_2)$	$\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$	$\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$	$\{x_1,x_2\}$	$\{x_1,x_2\}$	$\{x_2\}$
$n_A^\delta(x_3)$	$\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$	$\{x_1,x_2,x_3,x_5\}$	$\{x_1,x_3\}$	$\{x_3\}$	$\{x_3\}$
$n_A^\delta(x_4)$	$\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$	$\{x_2,x_4,x_5,x_7\}$	$\{x_4,x_7\}$	$\{x_4,x_7\}$	$\{x_4\}$
$n_A^\delta(x_5)$	$\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$	$\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$	$\{x_5\}$	$\{x_5\}$	$\{x_5\}$
$n_A^\delta(x_6)$	$\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$	$\{x_6,x_7\}$	$\{x_6\}$	$\{x_6\}$	$\{x_6\}$
$n_A^\delta(x_7)$	$\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$	$\{x_2,x_4,x_5,x_6,x_7\}$	$\{x_4,x_7\}$	$\{x_4,x_7\}$	$\{x_7\}$
$n_A^\delta(x_i)\in U/NR_A$	$\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$	$\{x_1,x_2,x_3,x_5\}$	$\{x_1,x_2,x_3\}$	$\{x_1,x_2\}$	$\{x_1\}$
		$\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$	$\{x_1,x_2\}$	$\{x_3\}$	$\{x_2\}$
		$\{x_2,x_4,x_5,x_7\}$	$\{x_1,x_3\}$	$\{x_4,x_7\}$	$\{x_3\}$
		$\{x_6,x_7\}$	$\{x_4,x_7\}$	$\{x_5\}$	$\{x_4\}\{x_5\}$
		$\{x_2,x_4,x_5,x_6,x_7\}$	$\{x_5\}\{x_6\}$	$\{x_6\}$	$\{x_6\}\{x_7\}$

表3 基于属性增链的邻域互补信息度量

Table 3 Neighborhood complementary information metric based on attribute chaining

度量	$\{c_1\}$	$\{c_1,c_2\}$	$\{c_1,c_2,c_3\}$	$\{c_1,c_2,c_3,c_4\}$	$\{c_1,c_2,c_3,c_4,c_5\}$
(1)$NH_\delta(A)$	0.000 0	1.020 4	1.102 0	0.775 5	0.857 1
(2)$NHC_\delta(D\|A)$	0.489 8	0.816 3	0.163 3	0.040 8	0.000 0
(3)$NMI_\delta(D;A)$	0.000 0	0.632 7	0.591 8	0.449 0	0.489 8
(4)$NH_\delta(D)_A$	0.489 8	1.449 0	0.755 1	0.489 8	0.489 8
(5)$NHC_\delta(A\|D)$	0.000 0	0.387 7	0.510 2	0.326 5	0.367 3

基于表3结果，首先可以检验系统式(8)，即表3中第(3)个度量值等于第(4)与(2)的度量值的差，也等于第(1)与(5)的度量值的差。此外，$NH_\delta(D)_A \geqslant H(D)=\frac{4}{7}\times\frac{3}{7}+\frac{3}{7}\times\frac{4}{7}=0.489\,8$，当$A$取$\{c_1,c_2\}$或$\{c_1,c_2,c_3\}$时不等号为严格大于；可见，$NH_\delta(D)_A$依赖于$A$从而不同于只依赖于$D$的常值$H(D)$，进而$NMI_\delta(D;A)\neq H(D)-NHC_\delta(D|A)$(推论1)。最后聚焦粒化非单调性。伴随属性增链的覆盖细化，这5种度量都呈现“先增大后减少”趋势，该波动充分说明了所有度量的粒化非单调性。关于属性增链，虽然单元素具有邻域细化(即$\forall x\in U$有$n_{\{c_1\}}^\delta(x)\supseteq n_{\{c_1,c_2\}}^\delta(x)\supseteq\cdots\supseteq n_{\{C\}}^\delta(x)$，但覆盖在“脱离元素追踪”与“去除粒重复”后呈现对应变化的复杂性，比如覆盖粒数具有波动：$|U/NR_{\{c_1\}}|=1<|U/NR_{\{c_1,c_2\}}|=5<|U/NR_{\{c_1,c_2,c_3\}}|=6>|U/NR_{\{c_1,c_2,c_3,c_4\}}|=5<|U/NR_C|=7$。

定理8 设$U/NR_A\geqslant U/NR_B$，则同类型的邻域互补信息度量在A与B上的大小关系是不确定的，即以下5组度量无必然的大小关系：

(1) $NH_\delta(A)$与$NH_\delta(B)$；(2) $NHC_\delta(D|A)$与$NHC_\delta(D|B)$；(3) $NMI_\delta(D;A)$与$NMI_\delta(D;B)$；(4) $NH_\delta(D)_A$与$NH_\delta(D)_B$；(5) $NHC_\delta(A|D)$与$NHC_\delta(B|D)$。

基于例1的事实支撑，定理8自然提供5种邻域互补信息度量的粒化非单调性。主要分析前面3种重要度量的相关机制。事实上，度量的覆盖刻画具有对分类刻画的拓展性并利于后续近似推理，但直

接的邻域粒变化具有复杂性,这在很大程度上诱导了相关的粒化不确定性。

(1)基于式(5),邻域互补熵的"单和内部"涉及上凸抛物函数$p(1-p)$,其先增后减并在$p=0.5$处取得最大值0.25。当$U/NR_A \geqslant U/NR_B$时,有$\forall x\in U \Rightarrow n_A^\delta(x)\supseteq n_B^\delta(x)$(定理3),因此$\forall x\in U$有$p(n_A^\delta(x))=\frac{|n_A^\delta(x)|}{|U|}\geqslant\frac{|n_B^\delta(x)|}{|U|}=p(n_B^\delta(x))$,但无法确定$p(n_A^\delta(x))(1-p(n_A^\delta(x)))$与$p(n_B^\delta(x))(1-p(n_B^\delta(x)))$的大小关系。后续覆盖细化也具有不确定性,因此$NH_\delta(A)$与$NH_\delta(B)$无必然大小关系。

(2)类似地,当$U/NR_A \geqslant U/NR_B$时,领域互补条件熵"双和内部"具有信息减少的确定趋势:$\frac{|n_A^\delta(x)\cap D_j|}{|U|}\frac{|n_A^\delta(x)-D_j|}{|U|}\geqslant\frac{|n_B^\delta(x)\cap D_j|}{|U|}\frac{|n_B^\delta(x)-D_j|}{|U|}$。但在后续求和中信息具有增加的可能性,因为$B$的邻域粒数可以更多。最终,$NHC_\delta(D|B)$对比$NHC_\delta(D|A)$的大小关系还是无法确定。

(3)当$U/NR_A \geqslant U/NR_B$时,领域互补互信息的"双和内部"具有$\frac{|n_A^\delta(x)\cap D_j|}{|U|}\geqslant\frac{|n_B^\delta(x)\cap D_j|}{|U|}$与$(1-\frac{|n_A^\delta(x)\cup D_j|}{|U|})\leqslant(1-\frac{|n_B^\delta(x)\cup D_j|}{|U|})$,这两种相反方向的确定性导致两种因子乘积大小的不确定性。此外,"双求和"粒数目仍然是一个不确定问题,因此也无法最终确定$NMI_\delta(D;A)$与$NMI_\delta(D;B)$的大小关系。

3　基于邻域互补互信息的启发式属性约简

基于相关的不确定性语义,邻域互补熵和邻域互补条件熵、邻域互补互信息都可以被利用于属性约简构建。针对决策表,考虑到邻域互信息能够有效度量从条件属性到决策属性的信息量与依赖度,故本节主要采用该测度来构建启发式属性约简。基于粒化非单调性(定理8),这里采用文献[15]的约简策略与算法思路,其主要追求更高互信息量。

定义3　基于决策表DT,$R\subseteq C$称为C的一个约简,若(1)$NMI_\delta(D;R)\geqslant NMI_\delta(D;C)$;(2)$\forall r\in R$,$NMI_\delta(D;R-\{r\})<NMI_\delta(D;R)$。

定义4　属性$a\in A$,$a\in(C-A)$关于A的内部、外部重要度分别为$\mathrm{sig}_{\mathrm{in}}(a,A,D)=NMI_\delta(D;A)-NMI_\delta(D;A-\{a\})$;$\mathrm{sig}_{\mathrm{out}}(a,A,D)=NMI_\delta(D;A\cup\{a\})-NMI_\delta(D;A)$。

这里的约简追求更优的邻域互补互信息,定义3中的两条分别描述相关的"联合充分性"与"独立必要性"。内重要度$\mathrm{sig}_{\mathrm{in}}(a,A,D)$表示在$A$中删除属性$a$产生的关于邻域互补互信息的信息减量,而外重要度$\mathrm{sig}_{\mathrm{out}}(a,A,D)$表征在$A$上增加属性$a$产生的信息增量,两者提供了快速约简的属性选择机制。若$NMI_\delta(D;C-\{c\})<NMI_\delta(D;C)$,此时有$\mathrm{sig}_{\mathrm{in}}(c,C,D)>0$,即$c$关于$C$是重要的,因此可以构建$C$的重要属性子集。类似地,$\mathrm{sig}_{\mathrm{out}}(a,R,D)>0$说明$a$关于$R$是重要的,因此可以选择最大外重要度的对应属性加入$R$以实施快速更新。下面采用这两种重要度来设计一个启发式搜索算法,以快速得到一个约简。

算法1　基于邻域互补互信息的属性约简启发算法

输入　决策表DT、邻域半径δ;

输出　基于邻域互补互信息的属性约简R。

Step 1　设置$R=\varnothing$;

Step 2　$\forall c_i\in C$,计算$\mathrm{sig}_{\mathrm{in}}(c_i,C,D)$,若$\mathrm{sig}_{\mathrm{in}}(c_i,C,D)>0$,则实施更新$R\leftarrow R\cup\{c_i\}$;

Step 3　计算邻域互补互信息$NMI_\delta(D;R)$与$NMI_\delta(D;C)$,若$NMI_\delta(D;R)\geqslant NMI_\delta(D;C)$,则进入第5步,否则进入第4步;

Step 4 $\forall a\in(C-R)$,计算 $\mathrm{sig}_{\mathrm{out}}(a,R,D)$,并选择外部属性重要度最大的属性 a^*,进行更新 $R\leftarrow R\cup\{a^*\}$,并进入步骤3;

Step 5 $\forall r_i\in R$,若 $NMI_\delta(D;R-\{r_i\})\geqslant NMI_\delta(D;R)$,则进行更新 $R\leftarrow R-\{r_i\}$;

Step 6 返回 R。

算法1优化了文献[15]的非单调算法结构,并主要采用邻域互补互信息及其属性重要度进行启发式快速搜索。步骤1进行初始化。步骤2基于内重要度启发,搜索 C 中所有重要属性并循环加入 R。步骤3是1个评估过程,若 R 不满足约简第1条,则进入步骤4,选取最大外重要度的属性进行快速的循环更新;(最终)满足第1条,再利用步骤3进入步骤5。步骤5循环删除冗余属性。由此,步骤6输出结果。该算法能够快速得到一个基于邻域互补互信息的属性约简,相关时间复杂度是可行的[15]。

例2 基于例1的决策表及相关设置与计算,下面说明算法1及其有效性。具体地,步骤1赋值 $R=\varnothing$。步骤2计算 C 种属性的内重要度

$$\mathrm{sig}_{\mathrm{in}}(c_1,C,D)=\mathrm{sig}_{\mathrm{in}}(c_3,C,D)=\mathrm{sig}_{\mathrm{in}}(c_4,C,D)=\mathrm{sig}_{\mathrm{in}}(c_5,C,D)=0.040\,8>0,$$
$$\mathrm{sig}_{\mathrm{in}}(c_2,C,D)=-0.183\,7<0$$

由此,将 $\{c_1,c_3,c_4,c_5\}$4个重要属性循环添加到 R 中,此时 R 更新为 $R=\{c_1,c_3,c_4,c_5\}$。步骤3计算有

$$NMI_\delta(D;R)=0.673\,5>0.489\,8=NMI_\delta(D;C)$$

满足约简条件1,故进入步骤5。步骤5实施反向冗余剔除。由

$$NMI_\delta(D;R-\{c_1\})=0.591\,8<0.673\,5,NMI_\delta(D;R-\{c_3\})=0.795\,9>0.673\,5$$

故 R 剔除 c_3,保留 c_1,c_4,c_5。且 R 更新为 $R=\{c_1,c_4,c_5\}$。由

$$NMI_\delta(D;R-\{c_1\})=0.836\,7>0.795\,9$$

故剔除 c_1,保留 c_4,c_5。且 R 更新为 $R=\{c_4,c_5\}$。由

$$NMI_\delta(D;R-\{c_4\})=0.387\,8<0.836\,7,NMI_\delta(D;R-\{c_5\})=0.224\,5<0.836\,7$$

此时,步骤6输出最终约简结果 $R=\{c_4,c_5\}$。

4 UCI数据实验

本节实施数据实验来验证邻域互补熵、邻域互补条件熵和邻域互补互信息的粒化非单调性,以及基于邻域互信息的启发约简算法1。具体从UCI机器学习数据库(http://archive.ics.uci.edu/ml)选取5类数据集(如表4)。首先采用最大-最小标准化数据预处理,仍用Manhattan距离函数,邻域半径参见表4。

表4 5类UCI数据集描述

Table 4 Description of five categories of UCI data sets

标签	数据集	简写	样本数	条件属性数	决策分类数	邻域半径
(a)	Breast cancer coimbra	BCC	116	9	2	1.2
(b)	Breast cancer wisconsin	BCW	699	9	2	1.5
(c)	Indian liver patient dataset	ILPD	583	10	2	0.9
(d)	Wine	Wine	178	13	3	1.8
(e)	Ionosphere	Iono	351	33	2	8.5

为揭示信息变化,选取自然属性增链 $\{c_1\}\subset\{c_1,c_2\}\subset\cdots\subset C$(设链元 $\{c_1,\cdots,c_k\}=(A_k)$)。针对核心度量 $NH_\delta(A_k),NHC_\delta(D|A_k),NMI_\delta(D;A_k)$,表5提供了截断于 A_{13} 的主体信息值,图1则进行全部数值描绘(其横坐标对应 A_k 的 k,3种度量简记为 NH,NHC,NMI)。基于表5分析,结合图1趋势,3种度量的

表5　5类数据集关于属性增链的3种邻域互补信息值

Table 5　Three kinds of neighborhood complementary information values for attribute chaining with five categories of data sets

数据集	信息度量	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}
(a) BCC	$NH_\delta(A)$	0.000 0	1.534 3	6.540 8	13.460 5	17.137 2	23.215 1	23.692 8	21.960 3	17.435 9	—	—	—	—
	$NHC_\delta(D\|A_k)$	0.494 6	12.107 0	31.391 5	36.356 9	31.843 5	20.513 2	12.469 5	8.305 0	3.846 0	—	—	—	—
	$NMI_\delta(D;A_k)$	0.000 0	0.826 4	3.327 6	6.711 8	8.561 1	11.591 7	11.858 7	11.060 8	8.838 0	—	—	—	—
(b) BCW	$NH_\delta(A_k)$	0.000 0	1.435 5	26.122 0	53.343 7	58.154 8	49.673 6	54.493 3	45.297 2	45.654 1	—	—	—	—
	$NHC_\delta(D\|A_k)$	0.455 0	5.908 0	29.606 8	22.826 4	15.360 3	4.200 8	3.166 3	1.514 2	1.252 7	—	—	—	—
	$NMI_\delta(D;A_k)$	0.000 0	0.904 3	18.852 5	40.028 8	43.516 4	39.686 2	45.077 9	38.395 7	39.046 0	—	—	—	—
(c) ILPD	$NH_\delta(A_k)$	0.041 0	1.105 2	5.293 8	31.282 8	62.661 5	68.156 2	72.056 1	104.323 5	106.903 4	97.825 2	—	—	—
	$NHC_\delta(D\|A_k)$	2.805 5	0.736 6	5.038 6	25.609 5	50.815 6	52.082 3	54.081 1	60.158 2	38.666 1	25.596 9	—	—	—
	$NMI_\delta(D;A_k)$	0.023 5	0.475 8	2.260 3	13.025 0	25.822 4	27.873 1	29.349 8	42.718 3	44.357 3	40.953 8	—	—	—
(d) Wine	$NH_\delta(A_k)$	0.000 0	0.000 0	0.016 7	1.017 7	4.047 9	15.627 1	31.575 9	40.141 6	40.002 6	35.979 9	31.539 5	26.621 8	21.152 8
	$NHC_\delta(D\|A_k)$	0.658 3	0.658 3	1.951 8	18.887 3	50.147 7	90.152 3	59.780 7	31.383 5	18.921 6	8.357 3	4.293 0	2.086 0	0.394 8
	$NMI_\delta(D;A_k)$	0.000 0	0.000 0	0.011 6	0.722 5	2.726 6	10.921 4	23.749 8	31.182 6	31.289 2	28.958 7	25.496 6	21.266 3	16.821 3
(e) Iono	$NH_\delta(A_k)$	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.045 3	0.445 1	2.978 6	9.878 8	21.885 4	39.153 2	51.055 2	56.748 3
	$NHC_\delta(D\|A_k)$	0.460 2	0.460 2	0.460 2	0.460 2	0.460 2	5.004 0	21.997 8	51.135 9	81.805 6	99.672 8	99.497 7	79.361 6	62.754 5
	$NMI_\delta(D;A_k)$	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.029 2	0.275 7	1.740 6	5.674 1	12.655 2	21.506 9	27.071 3	29.610 1

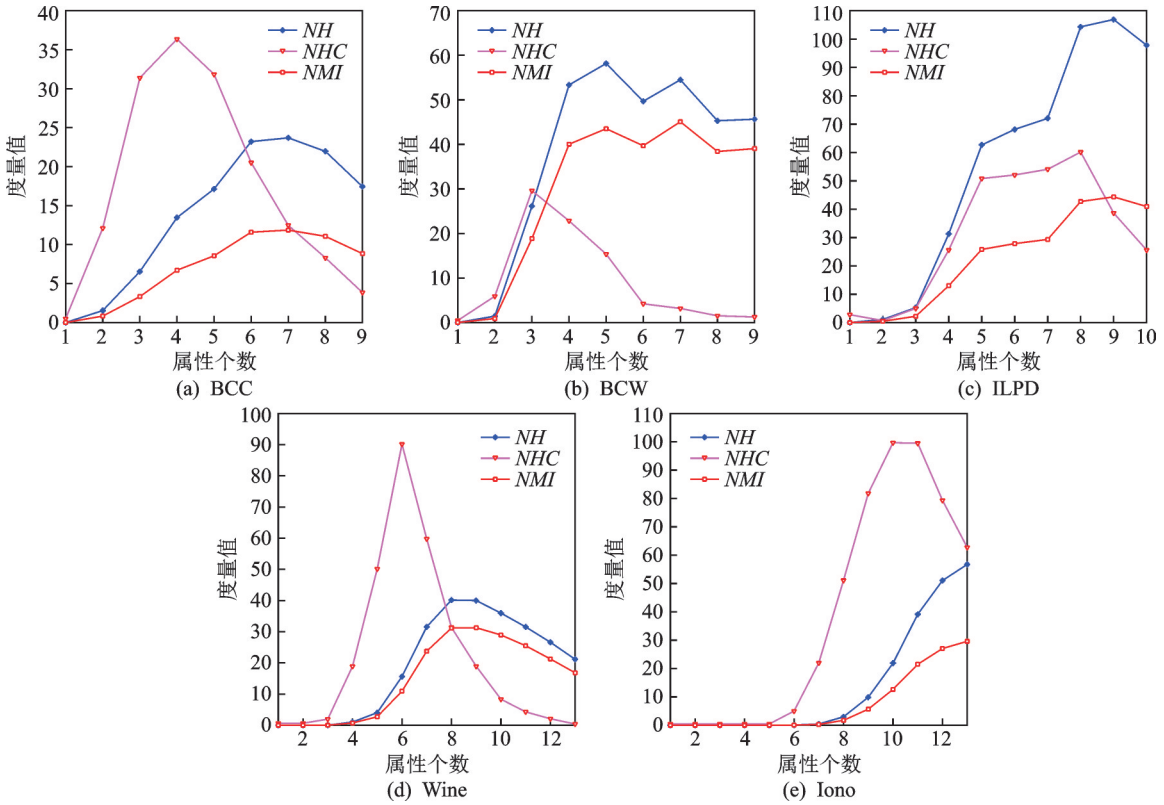

图1　5类UCI数据集关于属性增链的3种邻域互补信息的非单调变化

Fig.1　Non-monotonic changes in complementary information of three neighborhoods of attribute-added chain in five categories of UCI data sets

粒化非单调性均非常明显,对于基于邻域互补互信息的启发式属性约简,算法1提供如下有效约简结果(表6左栏)。

表6 两种算法下的约简结果

Table 6 Reduction results under two algorithms

标签	算法1	文献[15]算法
(a)	$\{c_1, c_2, c_6, c_9\}$	$\{c_1, c_2, c_3, c_9\}$
(b)	$\{c_1, c_3, c_4, c_5, c_9\}$	$\{c_1, c_2, c_3, c_4, c_6, c_7\}$
(c)	$\{c_1, c_2, c_8, c_9\}$	$\{c_1, c_2, c_9, c_{10}\}$
(d)	$\{c_1, c_7, c_{10}, c_{12}, c_{13}\}$	$\{c_1, c_2, c_6, c_7, c_{11}, c_{12}, c_{13}\}$
(e)	$\{c_1, c_2, c_3, c_4, c_5, c_{12}, c_{20}, c_{22}, c_{27}, c_{30}\}$	$\{c_1, c_2, c_3, c_4, c_5, c_{12}, c_{20}, c_{22}, c_{24}, c_{28}, c_{30}\}$

本文度量与算法最相关于文献[15],为了相关对比,补充了文献[15]的信息熵值及算法的数据实验,仍然基于表4的5类数据集。基于相关实验结果,文献[15]基于属性增链的3种度量值结果与表5的差距不大,对应的非单调图与图1也类似,故两者都省略。文献[15]算法所得约简结果放入表6右栏,其与本文算法1的结果(表6左栏)具有较明显的差异性;此外,文献[15]算法的实验处理比算法1需要更多的时间。综上实验结果对比可见,两套信息度量值具有一定的相似性,但启发的约简结果具有差异性,如表5(a)中$\{c_1,c_2,c_6,c_9\}\neq\{c_1,c_2,c_3,c_9\}$等,这种相关性与差异性来源于两者的度量机制。基于相关分析,两套度量具有相同的外部"叠加求和",在这种多信息融合情况下,内部的$p(1-p)$与$-\log_2^p$不能导致宏观显著性的值差异。但是这两种度量的微观差异在算法中会发生作用,从而致使启发属性的选择与顺序,即两种度量具有不同的约简启发性,故两者算法结果有所不同。此外,本文算法1对文献[15]算法的反向冗余剔除模块进行了结构改进,此实验中自然具有更高效率。

5 结束语

基于解析式模拟与粒替换,本文将经典粗糙集的经典互补信息度量推广到邻域粗糙集的邻域互补信息度量,得到了相似的系统体系(其中$H(D)$被$NH_\delta(D)_A$所替代),并将粒化单调性拓展为粒化非单调性,同时还得到了关于退化与双界等性质。基于邻域互补互信息及其粒化非单调性,提出属性约简及其启发式算法。最后,相关实例与实验都验证了研究结果的有效性。邻域互补信息度量及其属性约简还值得深入研究与应用,例如可以构建基于邻域互补(条件)熵的属性约简进行系统研究与应用。

参考文献:

[1] SHANNON C E. A mathematical theory of communication [J]. The Bell System Technical Journal, 1948, 27(4): 379-423.

[2] 苗夺谦. Rough Set理论及其在机器学习中的应用研究[D]. 北京: 中国科学院自动化研究所, 1997.
MIAO Duoqian. Rough Set theory and its application in machine learning [D]. Beijing: Institute of Automation, Chinese Academy of Sciences, 1997.

[3] 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 Guoyin, YU Hong, YANG Dachun. Decision table reduction based on conditional information entropy[J]. Journal of Computer Science, 2002, 25(7): 759-766.

[4] 江峰, 王凯邮, 于旭, 等. 基于粗糙熵的离群点检测方法及其在无监督入侵检测中的应用[J]. 控制与决策, 2019, 35(5): 1199-1204.
JIANG Feng, WANG Kaili, YU Xu, et al. Outlier detection method based on rough entropy and its application in unsupervised intrusion detection [J]. Control and Decision, 2019, 35(5): 1199-1204.

[5] ZHANG Xianyong, MIAO Duoqian. Three-layer granular structures and three-way informational measures of a decision table [J]. Information Sciences, 2017, 412/413: 67-86.

[6] LIANG Jiye, CHIN K S, DANG Chuangyin, et al. A new method for measuring uncertainty and fuzziness in rough set theory [J]. International Journal of General Systems, 2002, 31(4): 331-342.

[7] 魏巍, 陈红星, 王锋. 以互补条件熵为启发信息的正域属性约简[J]. 计算机工程与应用, 2013, 49(11): 96-100.
WEI Wei, CHEN Hongxing, WANG Feng. Positive domain attribute reduction based on complementary conditional entropy [J]. Computer Engineering and Applications, 2013, 49(11): 96-100.

[8] 唐玲玉. 基于三层粒结构的三支加权互补熵[D]. 成都: 四川师范大学, 2018.
TANG Lingyu. Three-weighted complementary entropy based on three-layer granular structure [D]. Chengdu: Sichuan Normal

University, 2018.
[9] YANG Xibei, LIANG Shaochen, YU Hualong, et al. Pseudo-label neighborhood rough set: Measures and attribute reductions [J]. International Journal of Approximate Reasoning, 2019, 105: 112-129.
[10] WANG Changzhong, HU Qinghua, WANG Xizhao, et al. Feature selection based on neighborhood discrimination index [J]. IEEE Transactions on Neural Networks Learning Systems, 2018, 29(7): 2986-2999.
[11] 邓志轩, 郑忠龙, 邓大勇. F-邻域粗糙集及其约简[J]. 自动化学报, 2019, 46(3): 1-11.
DENG Zhixuan, ZHENG Zhonglong, DENG Dayong. F-neighbor rough set and its reduction [J]. Chinese Journal of Automation, 2019. 46(3): 1-11.
[12] 周艳红, 张贤勇, 莫智文. 粒化单调的条件邻域熵及其相关属性约简[J]. 计算机研究与发展, 2018, 55(11): 2395-2405.
ZHOU Yanhong, ZHANG Xianyong, MO Zhiwen. Conditional neighborhood entropy of granular monotony and its related attribute reduction[J]. Computer Research and Development, 2018, 55(11): 2395-2405.
[13] 徐波, 张贤勇, 冯山. 邻域粗糙集的加权依赖度及其启发式约简算法[J]. 模式识别与人工智能, 2018, 31(3): 256-264.
XU Bo, ZHANG Xianyong, FENG Shan. Weighted dependency of neighborhood rough sets and its heuristic reduction algorithm[J]. Pattern Recognition and Artificial Intelligence, 2018, 31(3): 256-264.
[14] 姚晟, 汪杰, 徐风, 等. 不一致邻域粗糙集的不确定性度量和属性约简[J]. 小型微型计算机系统, 2018, 39(4): 700-706.
YAO Cheng, WANG Jie, XU Feng, et al. Uncertainty measure and attribute reduction of inconsistent neighborhood rough sets [J]. Small Microcomputer System, 2018, 39(4): 700-706.
[15] 姚晟, 徐风, 吴照玉, 等. 基于邻域粗糙互信息熵的非单调性属约简[J]. 控制与决策, 2019, 34(2): 353-361.
YAO Cheng, XU Feng, WU Zhaoyu, et al. Non-monotonicity reduction based on neighborhood rough mutual information entropy [J]. Control and Decision, 2019, 34(2): 353-361.
[16] HU Qinghua, ZHANG Lei, ZHANG David, et al. Measuring relevance between discrete and continuous features based on neighborhood mutual information [J]. Expert Systems with Applications, 2011, 38(9): 10377-10750.
[17] LIN Yaojin, HU Qinghua, LIU Jinghua, et al. Multi-label feature selection based on neighborhood mutual information [J]. Applied Soft Computing, 2016, 38: 244-256.
[18] CHEN Yumin, XUE Yu, MA Ying, et al. Measures of uncertainty for neighborhood rough sets [J]. Knowledge-Based Systems, 2017, 120: 226-235.
[19] ZHENG Tingting, ZHU Linjun. Uncertainty measures of neighborhood system-based rough sets [J]. Knowledge-Based Systems, 2015, 86: 57-65.
[20] YUAN Zhong, ZHANG Xianyong, FENG Shan. Hybrid data-driven outlier detection based on neighborhood information entroy and its development measures [J]. Expert Systems with Applications, 2018, 112: 243-257.
[21] 胡清华, 于达仁, 谢宗霞. 基于邻域粒化和粗基于邻域粒化和粗糙逼近是数值属性约简[J]. 软件学报, 2008, 19(3): 640-649.
HU Qinghua, YU Daren, XIE Zongxia. Attribute-based granulation and coarse-based neighborhood granulation and rough approximation are numerical attribute reductions[J]. Journal of Software, 2008, 19(3): 640-649.

作者简介:

陈帅(1992-),男,硕士研究生,研究方向:粗糙集、数据挖掘,E-mail:sshuaichen@163.com。

张贤勇(1978-),**通信作者**,男,教授,研究方向:粗糙集、粒计算、数据挖掘,E-mail:xianyongzh@sina.com。

唐玲玉(1984-),女,博士研究生,研究方向:粗糙集、粒计算,E-mail:278015337@qq.com。

姚岳松(1992-),男,硕士研究生,研究方向:粗糙集、数据挖掘,E-mail:yaoyuesongyuner@sina.com。

(编辑:刘彦东)