文章编号: 1001-0920 (2016) 01-0130-07 DOI: 10.13195/j.kzyjc.2014.1555

信息观下基于不一致邻域矩阵的属性约简

续欣莹[a], 刘海涛[a], 谢 珺[a], 谢 刚[b]

(太原理工大学 a. 信息工程学院，b. 国际教育交流学院，太原 030024)

摘 要: 信息观下研究邻域决策系统的属性约简是一种新颖的思路. 通过分析论域下某样本邻域中其他样本与该样本决策属性值的异同, 定义不一致邻域矩阵. 在计算属性重要度时, 利用不一致邻域减少在原条件属性基础上增加一个属性后条件熵的计算时间. 分析得到邻域系统下条件熵与正域的关系, 提出一种信息观下基于不一致邻域矩阵的属性约简算法, 并分析该算法与其他算法的内在联系. 实验结果验证了所提出算法的有效性.

关键词: 粗糙集；邻域矩阵；条件熵；属性约简

中图分类号: TP186 **文献标志码:** A

Attribute reduction based on inconsistent neighborhood matrix under information view

XU Xin-ying[a], *LIU Hai-tao*[a], *XIE Jun*[a], *XIE Gang*[b]

(a. College of Information Engineering，b. College of International Education and Exchange，Taiyuan University of Technology，Taiyuan 030024，China．Correspondent：XU Xin-ying，E-mail：xuxinying@tyut.edu.cn)

Abstract: It is a new approach for attribute reduction in the neighborhood decision system from the viewpoint of information theory. By analyzing the decision attribute values of samples in neighborhood, the inconsistent neighborhood matrix is defined. The inconsistent neighborhood matrix can be used to narrow the research range while adding more attributes with respect to existing condition attributes. Therefore, it is faster to calculate the significance of attributes by means of condition entropy. The relationship between the conditional entropy and the positive region is found under the neighborhood systems. An attribute reduction algorithm based on inconsistent neighborhood matrix is proposed, and the internal connection between this algorithm and other algorithms is analyzed. The experiment verifies the effectiveness of the proposed algorithm.

Keywords: rough set；neighborhood matrix；conditional entropy；attribute reduction

0 引 言

粗糙集理论[1]自 1982 年被提出以来, 已经广泛应用于属性约简[2]、规则挖掘[3]、知识分类[4]、辅助决策[5]、图像分割[6]等领域. 其中, 属性约简是在删除冗余信息后, 不改变信息系统分类或者决策能力, 从而简化了数据库结构的复杂度. 因此, 属性约简已经是粗糙集的主要研究内容之一, 也是辅助决策、数据挖掘和知识发现的关键步骤.

现在主要有 3 种属性约简的方法: 一是基于可辨识矩阵的属性约简算法[7]; 二是代数观, 即基于正域的属性约简算法[8]; 三是信息观, 即基于启发式信息的属性约简算法[9].

信息熵是一种被广泛应用于不确定性度量的度量工具, 文献 [10] 应用 shannon 熵的变形研究了粗糙集和粗糙关系数据库的不确定性信息度量; 王国胤等[11]给出了基于 shannon 熵的属性约简算法; 梁吉业等[12]提出了一种信息增益具有补特征的信息熵, 给出了其条件熵和互信息, 并指出这也是一种模糊熵, 可应用于度量粗糙集和粗糙分类的模糊性. 经典粗糙集只适合处理离散型数据, 而在现实生活中的实际数据大多是连续型数据, 因此在处理连续型数据时, 往往是先对其离散化, 再利用经典粗糙集处理. 但在离散化过程中势必造成信息的损失, 对处理结果产生影响, 其结果在很大程度上依赖于离散化的效果.

为了解决这一问题, Lin[13]提出了邻域关系模型, 该模型主要借助于拓扑学中内点和闭包的概念.

收稿日期: 2014-10-13；**修回日期:** 2015-01-07.

基金项目: 人社部留学回国人员科技活动择优资助项目(2013-68)；山西省自然科学基金项目(2014011018-2)；山西省回国留学人员科研项目(2013-033)；山西省留学回国人员科技活动择优资助项目(2013 年度).

作者简介: 续欣莹(1979−), 男, 副教授, 从事粒计算、大数据分析与智能控制的研究；刘海涛(1989−), 男, 硕士生, 从事粒计算、数据挖掘与智能信息处理的研究.

Yao[14]和 Wu 等[15]分别研究了 1-step 和 k-step 邻域信息系统. 在此基础上, 胡清华等[16]系统地分析了如何利用拓扑空间中球形邻域的概念, 并构造了代数观下基于邻域决策系统的数值数据特征选择算法, 该方法直观、易于理解, 能够直接处理连续型属性. 此后, Hu 等[17]又将信息熵引入邻域决策系统, 以互信息为启发条件构造了邻域决策系统信息观下的约简算法.

本文为了分析信息观下邻域决策系统的属性约简, 引入梁吉业等[12]在经典粗糙集下提出的信息熵, 定义了邻域决策系统的信息熵及其条件熵. 分析任一样本邻域中的其他各样本与该样本决策值的异同, 提出不一致邻域, 并发现通过不一致邻域可以加快增加属性后条件熵的计算速度. 此外, 研究发现了邻域系统下信息观 (条件熵) 与代数观 (正域) 之间的关系, 给出了其性质和定理; 结合条件熵和正域的这些定理可以加快算法的收敛速度, 提出了信息观下基于不一致邻域矩阵的属性约简算法; 分析了胡清华提出的基于相对正域和基于互信息约简算法与本文算法之间的内在联系, 最后得出这 3 种算法在一致邻域决策系统下是等价的, 并通过实验验证了所提出算法的有效性.

1　基本概念

1.1　邻域决策系统及其性质

定义 1[16]　给定实数空间上的非空有限集合 $U=\{u_1,u_2,\cdots,u_n\}$, $\delta>0$. 对于 U 上任意对象 u_i 定义其 δ 邻域为 $\delta(u_i)=\{u|\Delta(u,u_i)\leqslant\delta\}$, 其中距离函数 Δ 满足: 1) $\Delta(u_1,u_2)\geqslant 0$, 当且仅当 $u_1=u_2$, $\forall u_1,u_2\in R^N$ 时, $\Delta(u_1,u_2)=0$; 2) $\Delta(u_1,u_2)=\Delta(u_2,u_1)$, $\forall u_1,u_2\in R^N$; 3) $\Delta(u_1,u_2)\leqslant\Delta(u_1,u_2)+\Delta(u_2,u_3)$, $\forall u_1,u_2,u_3\in R^N$.

两个样本的相似程度可以利用距离函数来度量, 一般认为两个样本的距离越近越相似. 在二维实数空间中, 常用的距离有 1-范数 (曼哈顿距离)、2-范数 (欧氏距离) 和无穷范数 (切比雪夫距离). 在相同的邻域半径 δ 下, 无穷范数的邻域最大, 因此本文均采用无穷范数. 无穷范数距离定义为

$$\Delta_\infty(u_1,u_2)=\max_{i=1}^{n}(|f(u_1,a_i)-f(u_1,a_j)|),$$

其中 n 是条件属性的个数.

定义 2[16]　给定一个邻域决策系统 $\mathrm{NDT}=\langle U,C,D\rangle$. 其中: C 是条件属性集, D 是决策属性, $\forall B\subseteq C$. $\delta_A(u)$ 表示样本 u 在属性 $A\subseteq R$ 下的邻域, 则决策 D 关于 B 的下近似和上近似定义为

$$\underline{N_B}D=\{u_i|\delta_B(u_i)\subseteq\delta_D(u_i),u_i\in U\},$$

$$\overline{N_B}D=\{u_i|\delta_B(u_i)\bigcap\delta_D(u_i)\neq\varnothing,u_i\in U\}.$$

且决策正域 $\mathrm{POS}_B(D)$ 就是下近似 $\underline{N_B}(D)$.

胡清华定义的邻域没有考虑决策属性, 本文在加入决策属性分析后给出如下定义.

定义 3[16]　给定一个邻域决策系统 $\mathrm{NDT}=\langle U,C,D\rangle$. 其中: C 是条件属性集, D 是决策属性. 根据邻域中样本决策值的差异, 将邻域分为一致邻域和不一致邻域. 样本 u 的一致邻域定义为样本 u 的邻域中决策值相同的邻域: $\forall u_i\in\delta_C(u)$, 且 $u_i\in\delta_D(u)$, 即 $\delta_C(u)\bigcap\delta_D(u)$.

反之, 样本 u 的不一致邻域定义为样本 u 的邻域中决策值不同的邻域: $\forall u_i\in\delta_C(u)$, 且 $u_i\notin\delta_D(u)$, 即 $\delta_C(u)-\delta_D(u)$.

图 1 给出了样本 u 一致邻域和不一致邻域的集合解释.

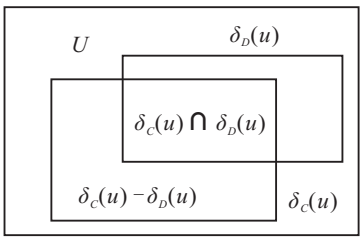

图 1　一致邻域和不一致邻域

性质 1　邻域决策系统 $\mathrm{NDT}=\langle U,C,D\rangle$, 给定邻域大小 δ, 对于决策表中任意 $u\subseteq U$, 有 $\delta_{B\cup A}(u)\subseteq\delta_B(u)\bigcap\delta_A(u)$, 且其在无穷范数距离下相等.

证明　因为 $A\subseteq A\bigcup B, B\subseteq A\bigcup B$, 所以

$$\delta_{B\cup A}(u)\subseteq\delta_A(u),\delta_{B\cup A}(u)\subseteq\delta_B(u)\Rightarrow$$

$$\delta_{B\cup A}(u)\subseteq\delta_B(u)\bigcap\delta_A(u).$$

由无穷范数距离的定义可知: 设 $\forall u_i\in\delta_B(u)\bigcap\delta_A(u)$, 在属性 B 下得到最大 Δ 的属性为 a_i, 在属性 A 下得到最大 Δ 的属性为 a_j, 所以其在属性 $B\bigcup A$ 下得到的最大 Δ 必小于给定的 δ, 由此可得 $\delta_B(u)\bigcap\delta_A(u)\subseteq\delta_{B\cup A}(u)$. 因此, 有

$$\delta_{B\cup A}(u)=\delta_B(u)\bigcap\delta_A(u).\qquad\square$$

性质 2　邻域决策系统 $\mathrm{NDT}=\langle U,C,D\rangle$, $\forall B\subseteq C$, 给定邻域大小 δ, 决策正域是不一致邻域为空集的样本的集合, 即

$$\mathrm{POS}_B(D)=\{u_i|\delta_B(u_i)-\delta_D(u_i)=\varnothing,\ u_i\in U\}.$$

证明　假设 $\exists u_j\in\mathrm{POS}_B(D)$ 使得 $\delta_B(u_j)-\delta_D(u_j)\neq\varnothing$. 由于 $u_j\in\mathrm{POS}_B(D)$, 根据定义 2 可得 $\delta_B(u_j)\subseteq\delta_D(u_j)$, 推导出 $\delta_B(u_j)-\delta_D(u_j)=\varnothing$, 与假设矛盾. $\square$

由性质 2 可以得到, 非正域就是不一致邻域不为空的样本的集合.

1.2　信息熵及其性质

梁吉业等[12]给出了信息系统在经典粗糙集下信息熵的统一表示.

定义 4[12] 设$S=(U,A)$是一个信息系统, 信息可统一表示为$K(A)=(S_A(u_1),S_A(u_2),\cdots,S_A(u_{|U|}))$, 其中$S_A(u_i)$为样本$u_i$在属性$A$下的类. 则$A$的信息熵定义为

$$E(A)=\sum_{i=1}^{|U|}\frac{1}{|U|}\Big[1-\frac{|S_A(u_i)|}{|U|}\Big].$$

在梁吉业给出的经典粗糙集下信息熵的基础上, 定义了邻域决策系统的信息熵.

定义 5 给定一个邻域决策系统$\mathrm{NDT}=\langle U,C,D\rangle$. 其中: C是条件属性集, D是决策属性. 任一样本u_i在条件属性C下的邻域为$\delta_C(u_i)$, 则邻域决策系统条件属性的信息熵定义为

$$E(A)=\sum_{i=1}^{|U|}\frac{1}{|U|}\Big[1-\frac{|\delta_C(u_i)|}{|U|}\Big].$$

通过该信息熵可以推导出其条件熵为

$$\begin{aligned}&E(D|C)=E(C,D)-E(C)=\\&\frac{1}{|U|}\sum_{i=1}^{|U|}\Big[1-\frac{|\delta_C(u_i)\bigcap\delta_D(u_i)|}{|U|}\Big]-\\&\frac{1}{|U|}\sum_{i=1}^{|U|}\Big[1-\frac{|\delta_C(u_i)|}{|U|}\Big]=\\&\frac{1}{|U|}\sum_{i=1}^{|U|}\Big[\frac{|\delta_C(u_i)-\delta_C(u_i)\bigcap\delta_D(u_i)|}{|U|}\Big]=\\&\frac{1}{|U|}\sum_{i=1}^{|U|}\Big[\frac{|\delta_C(u_i)-\delta_D(u_i)|}{|U|}\Big].\end{aligned}$$

定义 6 给定一个邻域决策系统$\mathrm{NDT}=\langle U,C,D\rangle$. 其中: C是条件属性集, D是决策属性. 任一样本u_i在条件属性C下的邻域为$\delta_C(u_i)$, 在决策属性D下的决策类为$\delta_D(u_i)$, 为样本的不一致邻域. 则邻域决策系统中决策D关于属性C的条件熵定义为

$$E(D|C)=\sum_{i=1}^{|U|}\frac{|\delta_C(u_i)-\delta_D(u_i)|}{|U|^2},$$

$$0\leqslant E(D|C)\leqslant 1-\frac{1}{|U|}.$$

当且仅当所有样本都是正域时, 条件熵$E(D|C)=0$; 当$\forall u_i\in U,\delta_C(u_i)=U,\delta_D(u_i)=\{u_i\}$时, 条件熵$E(D|C)$取得最大值$1-\dfrac{1}{|U|}$.

定理 1 给定一个邻域决策系统$\mathrm{NDT}=\langle U,C,D\rangle$, $B\subseteq C$. 对于任意$u_i\in U-\mathrm{POS}_B(D)$, $u_j\in\mathrm{POS}_B(D)$, 必有$u_j\notin(\delta_B(u_i)-\delta_D(u_i))$, 且剔除决策正域后, 其条件熵为

$$E_{U-\mathrm{POS}_B(D)}(D|C)=\frac{|U|^2}{|U-\mathrm{POS}_B(D)|^2}E_U(D|C).$$

证明 由定理 1 可知, $\forall u_i\in U-\mathrm{POS}_B(D)$, $u_j\in\mathrm{POS}_B(D)$, 分两种情况讨论.

1) $u_j\in\delta_B(u_i)\Rightarrow u_i\in\delta_B(u_j)\Rightarrow u_i\in\delta_D(u_j)\Rightarrow u_j\in\delta_D(u_i)\Rightarrow u_j\notin\delta_B(u_i)-\delta_D(u_i)$;

2) $u_j\notin\delta_B(u_i)\Rightarrow u_j\notin\delta_B(u_i)-\delta_B(u_j)$.

因此, 对于任意$u_i\in U-\mathrm{POS}_B(D)$, $u_j\in\mathrm{POS}_B(D)$, 必有$u_j\notin(\delta_B(u_i)-\delta_D(u_i))$.

由此可得, 对于任意$u_i\in U-\mathrm{POS}_B(D)$, 有$(\delta_B(u_i)-\delta_D(u_i))\bigcap\mathrm{POS}_B(D)=\varnothing$. 可得$u_i$在$U-\mathrm{POS}_B(D)$下的不一致邻域$\delta'_B(u_i)-\delta'_D(u_i)=\delta_B(u_i)-\delta_D(u_i)-\mathrm{POS}_B(D)=\delta_B(u_i)-\delta_D(u_i)$.

由性质 2 可得, 对于任意$u_j\in\mathrm{POS}_B(D)$, 有$\delta_B(u_j)-\delta_D(u_j)=\varnothing$, 所以$\displaystyle\sum_{j=1}^{\mathrm{POS}_B(D)}(\delta_B(u_j)-\delta_D(u_j))=\varnothing$. 可得

$$\begin{aligned}&E_U(D|B)=\frac{1}{|U|^2}\sum_{i=1}^{|U|}(\delta_B(u_i)-\delta_D(u_i))=\\&\frac{1}{|U|^2}\sum_{i}^{|U-\mathrm{POS}_B(D)|}(\delta_B(u_i)-\delta_D(u_i))=\\&\frac{1}{|U|}\sum_{i=1}^{|U-\mathrm{POS}_B(D)|}(\delta'_B(u_i)-\delta'_D(u_i)).\end{aligned}$$

因此,

$$E_{U-\mathrm{POS}_B(D)}(D|C)=\frac{|U|^2}{|U-\mathrm{POS}_B(D)|^2}E_U(D|C),$$

其中$\delta'(u_i)$表示在$U-\mathrm{POS}_B(D)$下的邻域. □

由定理 1 可知, 邻域条件熵的实质是非正域样本的不一致性度量.

定理 2[12] 设$B\subseteq C$, 若$E(D|B)=E(D|C)$, 且不存在属性$A\subset B$使$E(D|A)=E(D|B)$, 则称属性B为属性C相对于决策D的一个约简.

定义 7[12] 给定一个邻域决策系统$\mathrm{NDT}=\langle U,C,D\rangle$, $B\subseteq C$. 对于任意$a\in C-B$, a相对于B的重要度定义为$\mathrm{sig}(a,B,D)=E(D|B)-E(D|B\bigcup\{A\})$.

2 属性约简算法

2.1 基于条件熵的邻域决策系统的属性约简算法

结合文献[11]中经典粗糙集下基于条件熵的属性约简算法, 给出一种基于条件熵的邻域决策系统的属性约简算法.

算法将信息观下的属性重要度作为指标, 以空集作为起点, 每次对全部剩余属性$a_i\in C-\mathrm{red}$计算$E(D|B\bigcup\{a_i\})$, 选择$E(D|B\bigcup\{a_i\})$值最小的属性(也就是属性重要度值最大的属性)加入约简集red中, 直到所有剩余属性的条件熵等于决策系统的条件熵$E(D|B\bigcup\{a_i\})=E(D|C)$.

算法 1 基于条件熵的邻域决策系统属性约简算法.

输入: 邻域决策系统$\mathrm{NDT}=\langle U,C,D\rangle$;

输出: 邻域决策系统的约简集red.

Step 1: 计算决策系统$\mathrm{NDT}=\langle U,C,D\rangle$的条件

熵 $E(D|C)$;

Step 2: 令 $\mathrm{red}=\varnothing$;

Step 3: 对于任意 $a_i\in C-\mathrm{red}$ 计算 $E(D|\mathrm{red}\bigcup\{a_i\})$;

Step 4: 选择使 $E(D|\mathrm{red}\bigcup\{a_i\})$ 最小的 a_k, 将 a_k 加入到约简 red 中;

Step 5: 如果 $E(D|\mathrm{red}\bigcup\{a_k\})\neq E(D|C)$, 跳转到 Step 3, 否则输出约简 red.

2.2 基于不一致邻域矩阵的邻域决策系统属性约简算法

2.2.1 基于不一致邻域矩阵求解条件熵算法

定理 3　邻域决策系统 $\mathrm{NDT}=\langle U,C,D\rangle$, 给定邻域大小 δ, 如果 $B_1\subseteq B_2\subseteq C$, 在计算样本 u 在属性 B_2 下的不一致邻域时, 只需要考虑样本 u 在属性 B_1 下的不一致邻域, 即 $\delta_{B_2}(u)-\delta_D(u)\subseteq\delta_{B_1}(u)-\delta_D(u)$.

证明　因为 $B_1\subseteq B_2\subseteq C$, 所以

$$\delta_{B_2}(u)\subseteq\delta_{B_1}(u)\Rightarrow$$

$$\delta_{B_2}(u)-\delta_D(u)\subseteq\delta_{B_1}(u)-\delta_D(u).\qquad\square$$

定义 8　根据定义 3 给出不一致邻域矩阵定义为

$$\mathrm{NM}(i,j)=\begin{cases}1,\ u_j\in\delta_C,\ (u_i)u_j\notin\delta_D(u_i);\\ 0,\ \text{else}.\end{cases}$$

定理 3 可以大大降低在寻找增加条件属性后样本的不一致邻域矩阵时搜索范围, 定义 8 给出了算法实现的基础, 因此结合定理 3 和定义 8 设计一种根据现在条件属性下的不一致邻域来计算增加一个条件属性后条件熵的算法.

算法 2　基于不一致邻域矩阵求解条件熵算法.

输入: B 属性下的不一致邻域 NM, 新增条件属性 $\{a\}$;

输出: $E(D|B\bigcup\{a\})$.

Step 1: 对于任意 $u_i\in U$, 找到其不一致邻域 $\delta_B(u_i)-\delta_D(u_i)=\{u_j|\mathrm{NM}(i,j)=1,\forall u_j\in U\}$;

Step 2: 对于任意 $u_j\in(\delta_B(u_i)-\delta_D(u_i))$ 作如下判别, 并得到 $B\bigcup\{a\}$ 属性下的不一致邻域矩阵:

$$\text{If }\Delta_{B\cup\{a\}}(u_i,u_j)>\delta,$$

$$\mathrm{NM}(i,j)=0;$$

Step 3: 根据 $B\bigcup\{a\}$ 属性下的不一致邻域矩阵得到每一个样本的不一致邻域 $|\delta_B(u_i)-\delta_D(u_i)|$;

Step 4: 根据条件熵求得 $E(D|B\bigcup\{a\})$.

2.2.2 信息观下基于不一致邻域矩阵的邻域决策系统信息观下的属性约简算法

定理 4　邻域决策系统 $\mathrm{NDT}=\langle U,C,D\rangle$, $B_1\subseteq B_2\subseteq C$. 如果属性 a 在属性 B_1 下是不必要的, 即 $E(D|B_1)=E(D|B_1\bigcup\{a\})$, 则属性 a 在属性 B_2 下也是不必要的.

证明　由题可得

$$E(D|B_1)=$$

$$\sum_{i=1}^{|U|}\frac{|\delta_{B_1}(u_i)-\delta_D(u_i)|}{|U|^2}=\sum_{i=1}^{|U|}\frac{|\delta_{B_1}(u_i)-\delta_D(u_i)|}{|U|}=$$

$$E(D|B_1\bigcup\{a\})\Rightarrow\sum_{i=1}^{|U|}|\delta_{B_1}(u_i)-\delta_D(u_i)|=$$

$$\sum_{i=1}^{|U|}|\delta_{B_1\cup\{a\}}(u_i)-\delta_D(u_i)|.$$

可推导出, 对于任意 u_i, 有

$$\delta_{B_1}(u_i)=\delta_{B_1\cup\{a\}}(u_i)=$$

$$\delta_{B_1}(u_i)\bigcap\delta_{\{a\}}(u_i)\,(\text{性质 2})\Rightarrow\delta_{B_1}(u_i)\subseteq\delta_{\{a\}}(u_i).$$

设 $B_2=B_1\bigcup B$, 则对于任意 u_i, 有

$$\delta_{B_2}(u_i)=\delta_{B_2\cup B}(u_i),$$

$$\delta_{B_2\cup\{a\}}(u_i)=\delta_{B_2\cup B_1\cup\{a\}}(u_i),$$

$$\delta_{B_1\cup\{a\}\cup B}(u_i)=$$

$$\delta_{B_1}(u_i)\bigcap\delta_B(u_i)\bigcap\delta_{\{a\}}(u_i)=$$

$$\delta_{B_1}(u_i)\bigcap\delta_B(u_i)=\delta_{B_2}(u_i).$$

根据定理 4 在算法迭代中提前找到不必要的属性并剔除. 此外, 由定理 1 可知, 剔除正域后的不一致邻域并没有发生变化, 此时可以计算整个论域的条件熵. 可以通过剔除不必要的属性和对条件熵计算没有影响的样本加快算法的运行速度. 在此基础上结合算法 1 和算法 2 给出信息观下基于不一致邻域矩阵的约简算法.

算法 3　信息观下基于不一致邻域矩阵的属性约简 (INM).

输入: 邻域决策系统 $\mathrm{NDT}=\langle U,C,D\rangle$;

输出: 邻域决策系统的约简集 red.

Step 1: 计算决策系统 $\mathrm{NDT}=\langle U,C,D\rangle$ 的条件熵 $E(D|C)$.

Step 2: 令 $\mathrm{red}=\varnothing$, $B=\varnothing$.

Step 3: 如果 $\mathrm{red}=\varnothing$, 则对于任意 $a_i\in C-B$, 计算 $E(D|\mathrm{red}\bigcup\{a_i\})$; 否则, 对任意 $a_i\in C-B$, 利用不一致邻域矩阵 NM 计算 $E(D|\mathrm{red}\bigcup\{a_i\})$. 如果 $E(D|\mathrm{red}\bigcup\{a_i\})=E(D|\mathrm{red})$, 则 $B\bigcup\{a_i\}\to B$.

Step 4: 选择使 $E(D|\mathrm{red}\bigcup\{a_i\})$ 最小的 a_k, 将 a_k 加入到约简 red 中, $B\bigcup\{a_k\}\to B$, 并得到相应的不一致邻域矩阵 NM.

Step 5: 如果 $E(D|\mathrm{red}\bigcup\{a_k\})\neq E(D|C)$, 那么对于任意 $u_i\in U$, 若 $\mathrm{NM}(i)=[0]$, 则 $U=U-\{u_i\}$, 转入 Step 3; 否则输出约简 red.

算法时间复杂度分析: 设 $|U|$ 和 $|C|$ 分别表示邻

域决策系统的样本数和条件属性个数. 首先分析计算条件熵的时间复杂度, 假设正域为空集, 对每个样本求不一致邻域时, 在最坏的情况下, 算法需要循环$|U|(|U|-1)(|C|-1)/2$次; 然后分析属性约简算法的时间复杂度, 假设所有的属性都是必要的, 即约简以后还是属性全集, 此时的循环次数为$|C|(|C|+1)/2$. 因此, 该算法总的时间复杂度为$O(|U|^2\times|C|^3)$.

2.3 分析几种约简算法之间的联系

目前, 粗糙集属性约简的理论观点主要有两种: 代数观和信息观. 胡清华等[16-17]将代数观和信息观约简算法推广到邻域系统. 这些约简算法都是启发式约简算法, 只是启发条件不同. 约简算法之间的联系可以简化为启发条件之间的关系.

文献[16]根据Pawlak定义的传统粗糙集理论约简思想(代数观), 提出了以决策属性对条件属性的依赖度作为启发条件的属性约简算法(NRS), 即依据决策表的相对正域是否发生变化; 文献[17]将信息熵(用H表示)引入邻域系统, 并提出了以互信息($H(C;D)$)作为启发条件的约简算法(NMI). 本文将梁吉业等[12]定义的新信息熵(用E表示)引入邻域系统, 以条件熵($E(D|C)$)作为启发条件. 这些启发条件的表达式如下所示:

依赖度

$$r_B(D)=\frac{|\mathrm{POS}_B(D)|}{|U|};$$

互信息

$$H(B;D)=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_B(u_i)|\times|\delta_D(u_i)|}{|U|\times|\delta_B(u_i)\bigcap\delta_D(u_i)|};$$

条件熵

$$E(D|B)=\sum_{i=1}^{|U|}\frac{|\delta_B(u_i)-\delta_D(u_i)|}{|U|^2}.$$

这几种启发式算法的约简条件都类似: 对于任意$a\in B$, 若有$\varphi_{B-\{a\}}=\varphi_B$, 则$a$是多余的, 其中$\varphi$表示依赖度、互信息和条件熵.

依赖度$r_B(D)$是相对正域中样本在整个论域中所占的比率, 即论域中确定信息所占的比重. 也就是说, 以依赖度为启发条件的约简算法中, 判断一个属性是否为冗余属性, 是依据删除该属性是否影响决策表中的确定信息决定的. 互信息($H(B;D)$)是反映属性B与决策D之间的依赖程度. 以互信息为启发条件的约简算法中, 判断一个属性是否为冗余属性, 是依据删除该属性是否影响不确定信息和确定信息的比重决定的.

$$H(B;D)=$$
$$-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{|\delta_D(u_i)|}{|U|}\Big(1-\frac{|\delta_B(u_i)-\delta_D(u_i)|}{|\delta_B(u_i)\bigcap\delta_D(u_i)|}\Big).$$

其中$\delta_B(u_i)=(\delta_B(u_i)-\delta_D(u_i))+(\delta_B(u_i)\bigcap\delta_D(u_i))$.

本文定义的条件熵($E(D|B)$)是每一个样本的不一致邻域的平均样本在整个论域中所占的比率, 即不确定信息所占的平均比重. 由定理1可知, 确定部分的条件熵为零, 条件熵由不确定部分产生. 因此约简算法以条件熵为启发条件时, 若删除某属性后, 既不影响不确定部分的概率分布, 也不影响确定部分的分类, 则该属性可以约简.

当邻域决策表在属性B下为一致决策表(即$\delta_B(u_i)\subseteq\delta_C(u_i)$)时, 依赖度$r_B(D)=1$达到最大值, 信息熵$E(D|B)=0$取得最小值, 而互信息$E(C,D)=\dfrac{1}{|U|}\displaystyle\sum_{i=1}^{|U|}\log_2\frac{|\delta_D(u_i)|}{|U|}$, 此时的互信息只与决策类有关, 而决策类是不变的. 也就是说, 一个决策表的互信息在所有的一致属性下是固定的.

当决策表在属性B下为完全不一致决策表(即$\delta_B(u_i)=\{u_i\},\delta_D(u_i)=U$)时, 依赖度$r_B(D)=0$取得最小值, 而此时信息熵$E(D|B)$达到最大值$1-\dfrac{1}{|U|}$. 可以看出, 依赖度与条件熵成反比, 互信息$H(C;D)=0$.

推论1 当邻域决策系统$\mathrm{NDT}=\langle U,C,D\rangle$是一致决策时, 如果$r_B(D)=r_C(D)$, $B\subseteq C$, 则$E(D|B)=E(D|C)$.

证明 因为邻域决策表在属性C下一致, 所以$r_B(D)=1$, 即$\mathrm{POS}_C(D)=|U|$. 根据$r_B(D)=r_C(D)$, 有$\mathrm{POS}_B(D)=\mathrm{POS}_C(D)=|U|$.

由性质2可知, 所有样本在属性B和C下的不一致邻域均为空集. 因此, $E(D|B)=E(D|C)=0$. □

文献[17]已证明, 对于一致邻域系统, 若$r_B(D)=r_C(D)$, 则$H(B;D)=H(C;D)$必成立. 由此可知依赖度、互信息、条件熵在一致邻域系统中是等价的.

3 实验及分析

本文从信息观出发, 结合正域和条件熵的关系, 提出了信息观下基于不一致邻域的属性约简算法(INM). 为了验证算法的有效性, 针对邻域阈值对分类精度的影响较大, 首先通过实验设置合适的邻域阈值, 然后对INM算法与NRS和NMI约简算法的约简结果进行对比实验. 本次实验选取5组UCI数据集进行测试, 所选用的数据集如表1所示. 本次实验环境为一台A4 1.9 GH处理器2G内存的PC机.

表1 数据描述

	数据集	样本数	条件属性个数	决策类别数
1	wine	178	13	3
2	wpbc	198	33	2
3	iono	351	34	2
4	wdbc	569	31	2
5	mushroom	8 124	22	2

3.1　邻域阈值 δ 的选取

邻域决策表的一致性是随 δ 的变化而变化的, 因此当 δ 发生变化时, 其属性约简的结果也随之变化, 邻域阈值 δ 的选取直接影响其属性约简结果. 本次实验选取 wine、wpbc 和 iono 三组数据集, 通过改变 δ 的大小, 验证 δ 对属性约简的结果及约简后分类精度的影响, 实验结果如图 2～图 5 所示. 本次实验 δ 的取值以步长 0.01 从 0 到 1 变化.

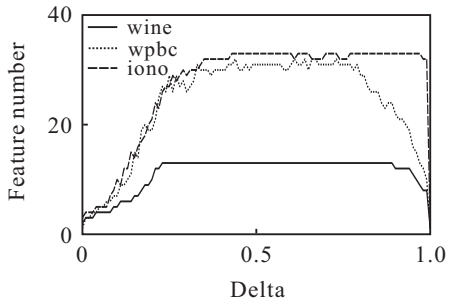

图 2　特征数量随 δ 的变化情况

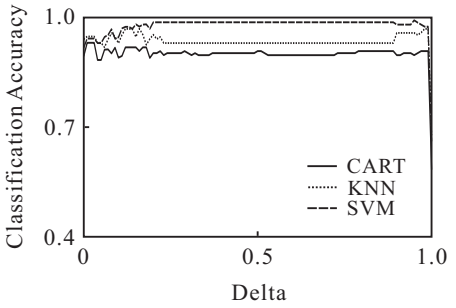

图 3　wine 数据集

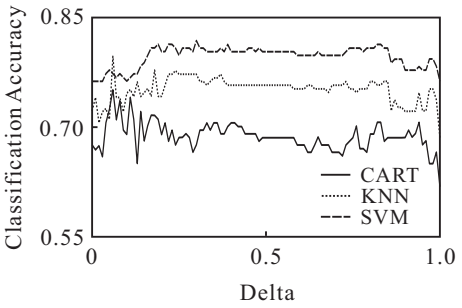

图 4　wpbc 数据集

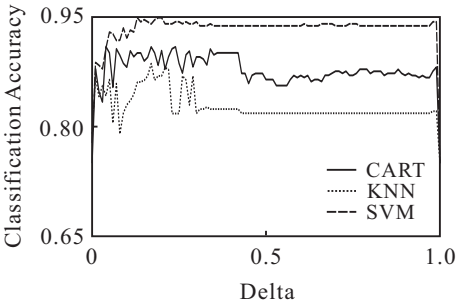

图 5　iono 数据集

图 2 展示了 3 组数据集的约简中特征数量随 δ 的变化情况. 从图 2 可以看出, 当 δ 在 0～0.1 之间以及接近 1 时, 约简集中特征数量较少, 在 0.1～0.2 之间适中, 在 0.2～0.3 达到最大并保持.

图 3～图 5 分别展示了 3 组数据集约简后, 在 CART、SVM 和 KNN ($k = 5$) 三种分类算法下的分类精度随 δ 的变化情况. 从图 3～图 5 可以看出, δ 在 0 和 1 附近时, 其分类精度普遍较低; 在 0.05～0.2 分类精度较高; 但是在 0.05～0.15 之间, 数据集 wpbc 在 CART 算法下的分类精度波动较大, 而数据集 iono 在 KNN 算法下的分类精度也有较大波动. 综合约简集的特征数量和约简后分类精度的情况, δ 选择在 0.15～0.2 之间比较合适. 为了确保实验的一致性, 邻域阈值 δ 选择为 0.17.

3.2　算法有效性验证

本次实验对约简算法的约简结果进行比较. 为了更好地验证算法的有效性, 实验引入经典的 CART 和 KNN ($k = 5$) 两种分类器, 并以十折交叉验证的分类精度来评价所选属性的质量. 此外, 为减少各属性量纲不一致对结果的影响, 将所有数值型属性标准化到 [0,1] 之间. 具体的实验结果如表 2～表 4 所示.

表 2　约简前后特征数量比较

data	特征数量			
	原始	NRS	NMI	INM
wine	13	7	5	8
wpbc	33	19	6	19
iono	34	17	8	17
wdbc	31	21	6	22
mushroom	22	5	3	4

表 3　CART 分类器下的分类精度　%

data	原始	NRS	NMI	INM
wine	89.9±6.4	92.1±4.8	91.0±6.0	92.0±6.3
wpbc	69.6±6.8	71.2±8.1	66.6±10.6	71.7±7.4
iono	85.9±7.2	87.6±7.0	93.2±3.7	90.9±5.1
wdbc	90.5±4.6	92.1±3.0	91.8±3.4	91.9±2.6
mushroom	96.4±9.9	96.4±9.9	96.0±9.8	96.9±9.9
平均值	86.5	87.9	87.7	88.6

表 4　KNN 分类器下的分类精度　%

data	原始	NRS	NMI	INM
wine	93.2±5.9	95.0±3.2	98.3±2.7	95.5±3.5
wpbc	75.8±5.6	74.3±7.9	73.7±6.5	74.3±7.9
iono	81.8±6.7	81.2±7.7	82.7±6.2	88.6±9.2
wdbc	95.6±2.0	95.3±2.5	96.1±2.3	95.4±2.2
mushroom	90.8±12.9	95.2±10.7	93.2±14.0	93.3±14.0
平均值	87.4	88.2	88.8	89.4

表 2 展示了原始特征数量和在 3 种算法约简后的特征数量. 从实验的结果可以看出, 代数观和信息观都可以约简掉不必要属性, NRS 和 INM 算法所得到的约简结果大致一致, 而 NMI 所得的结果比较少. 从邻域阈值选择实验的图 2 可以看出, INM 在 δ 取值 0.1 附近也可以找到较少的约简集, 但是此时的分类精度的波动比较大.

表 3 展示了 5 组数据集在 CART 分类器下的原始分类精度和几种约简算法约简后的分类精度; 表 4

展示了5组数据集在KNN分类器下的原始分类精度和几种约简算法约简后的分类精度. 从实验结果可以看出, 几种算法的分类精度与原始分类精度相比较, 都有一定的提高, 这说明约简后的特征可以表征整个数据集, 几种约简算法都是有效的. INM算法与NRS和NMI算法相比, 分类精度略有增幅. 但是, 具体到iono数据集, NMI算法在CART分类器下的分类精度比之其他两种算法要高, NRS算法在KNN分类器下的分类精度又相对较高, 这说明3种算法对不同类型的数据集敏感程度不同, 这也是今后需要深入研究的方向之一. 总体而言, 这3种算法是从不同角度对邻域决策系统进行属性约简, 实验表明本文提出的INM算法是有效的.

4 结 论

本文通过分析信息观下的邻域系统, 引入梁吉业定义的信息熵, 给出了邻域系统的信息熵和条件熵, 定义了不一致邻域矩阵, 得到了其在邻域决策系统下的性质和定理. 根据这些性质和定理提出了INM算法, 然后分析了INM算法与胡清华提出的代数观下NRS算法和信息观下NMI算法之间的联系. 其中, NRS算法是保持确定部分不发生变化, INM算法是保持不确定部分不发生变化, NMI算法是保持不确定部分和确定部分的比重不发生变化来进行属性约简, 论证后得到这3种算法在一致决策系统下是等价的. 本文通过实验分析邻域阈值对INM算法约简结果的影响来选取邻域阈值, 最后通过实验验证了所提出算法的有效性.

参考文献(References)

[1] Pawlak Z. Rough sets-theoretical aspects of reasoning about data[M]. Dordrecht: Kluwer Academic, 1991: 9-11.

[2] Xu X Y, Liu H F, Shen X F, et al. The research of attribute reduction algorithm based on extension neighborhood relation[J]. J of Computational Information Systems, 2013, 9(16): 6613-6620.

[3] Huang C C, Tzu-Liang(Bill) Tseng, Jiang F H, et al. Rough set theory: A novel approach for extraction of robust decision rules based on incremental attributes[J]. Annals of Operations Research, 2014, 216(1): 163-189.

[4] Eman Mohamed Fadl Ahmed, Sameh Ebrahim Rehan, Ahmed Atwan Mohamed. Rough set analysis and cloud model algorithm to automated knowledge acquisition forclassification iris to chieve high security[C]. 11th Int Conf on Hybrid Intelligent Systems(HIS). Malacca: IEEE, 2011: 55-60.

[5] 张明, 唐振民, 杨习贝. 不完备信息系统中的否定决策规则和知识约简[J]. 控制与决策, 2011, 26(6): 851-856.
(Zhang M, Tang Z M, Yang X B. Negative decision rules and knowledge reduction in incomplete information system[J]. Control and Decision, 2011, 26(6): 851-856.)

[6] Amiya Halder, Avijit Dasgupta. Image segmentation using rough set based K-means algorithm[J]. Association for Computing Machinery, 2012, 9(11): 53-58.

[7] 杨明, 吕静. 一种基于C-Tree的属性约简增量式更新算法[J]. 控制与决策, 2012, 27(12): 1769-1775.
(Yang M, Lv J. An incremental updating algorithm for attribute reduction based on C-Tree[J]. Control and Decision, 2012, 27(12): 1769-1775.)

[8] Xie J, Shen X F, Liu H F, et al. Research on an incremental attribute reduction based on relative positive region[J]. J of Computational Information Systems, 2013, 9(16): 6621-6628.

[9] Wang C R, Ou F F. An Attribute Reduction algorithm in rough set theory based on information entropy[J]. Int Symposium on Computational Intelligence and Design, 2008, 8(11): 3-6.

[10] Beaubouef T, Petry F E. Fuzzy rough set techniques for uncertainty processing in a relational database[J]. Int J of Intel System, 2000, 15(5): 389-424.

[11] 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): 760-766.
(Wang G Y, Yu H, Yang D C. Decision table reduction based on conditional information entropy[J]. Chinese J of Computers, 2002, 25(7): 760-766.)

[12] 梁吉业, 李德玉. 信息系统中的不确定性与知识获取[M]. 北京: 科学出版社, 2005: 13-113.
(Liang J Y, Li D Y. Uncertainty measurement and knowledge acquiring in information system[M]. Beijing: Science Press, 2005: 13-113.)

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

[14] Yao Y Y. Relational interpretation of neighborhood operators and rough set approximation operators[J]. Information Sciences, 1998, 111(1): 239-259.

[15] Wu W Z, Zhang W X. Neighborhood operator systems and approximations[J]. Information Sciences, 2002, 144(1): 201-217.

[16] 胡清华, 于达仁, 谢宗霞, 等. 基于邻域粒化和粗糙逼近的数值属性约简[J]. 软件学报, 2008, 19(3): 640-649.
(Hu Q H, Yu D R, Xie Z X, et al. Numerical attribute reduction based on neighborhood granulation and rough approximation[J]. J of Software, 2008, 19(3): 640-649.)

[17] Hu Q H, Zhang L, Zhang D, et al. Measuring relevance between discrete and continuous features based on neighborhood mutual information[J]. Expert Systems with Applications, 2011, 38(9): 10737-10750.

(责任编辑: 齐 霁)