文章编号: 1001-0920 (2016) 07-1225-06 DOI: 10.13195/j.kzyjc.2015.0719

基于邻域组合测度的属性约简方法

何松华[1], 康婵娟[1,2], 鲁 敏[2], 滕书华[2]

(1. 湖南大学 信息科学与工程学院，长沙 410082；2. 国防科技大学 自动目标识别重点实验室，长沙 410073)

摘 要: 属性约简是机器学习和知识发现的研究热点, 而属性重要性度量则是构建属性约简算法的关键环节. 针对不完备的混合型信息系统, 在邻域关系下定义了一种新的属性集成重要性度量——邻域组合测度, 并据此提出一种基于邻域组合测度的属性约简 (NCMAR) 算法. 通过多个 UCI 数据集上的实验表明, NCMAR 算法不仅能够直接处理符号和数值属性共存的混合信息系统, 而且适用于不完备信息系统, 在获得较小约简结果的同时, 能够保证较高的分类精度.

关键词: 粗糙集；属性约简；不确定性度量；不完备信息系统；混合数据

中图分类号: TP18 **文献标志码:** A

Attribute reduction method based on neighborhood combination measure

HE Song-hua[1], *KANG Chan-juan*[1,2], *LU Min*[2], *TENG Shu-hua*[2]

(1. Institute of Information Science and Engineering，Hunan University，Changsha 410082，China；2. ATR Laboratory, National University of Defense Technology，Changsha 410073，China. Correspondent：TENG Shu-hua，E-mail: tengshuhua1979@sohu.com)

Abstract: Atttribute reduction is a hot point in the machine learning and knowledge discover research, while the attribute importance measurement is the key link in the structure of the attribute reduction algorithm. For the imcomplete of the mixed information system, a new measurement method of attribute integration importance, named neighborhood combination measure, is defined under the neighborhood relation, and a neighborhood combination measure based attribute reduction(NCMAR) algorithm is also proposed. Some experiments are carried out on UCI data sets. And the experiments results show that the NCMAR algorithm can not only deal with mixed decision system with symbol data and numerical data, but is suitable for the imcomplete information system. What's more, it can obtain smaller reducts and better classification accuracy than current algorithms.

Keywords: rough sets；attribute reduct；uncertainty measure；incomplete information system；mixed data

0 引 言

粗糙集理论[1]作为一种处理不精确、不一致和不完备数据的智能信息处理技术, 已广泛应用于模式识别、机器学习、人工智能、知识获取以及数据挖掘[2]等方面.

传统粗糙集理论利用等价关系(自反性、对称性、传递性)进行信息粒化, 将论域划分成等价类, 作为描述论域中任意概念的基本信息粒子. 它只适用于处理符号型数据, 对于现实应用中(如科学研究、工程应用、金融、医疗等领域)广泛存在的数值型数据不能直接处理[3]. 通过扩展传统粗糙集理论中的等价关系, Lin[4]首次将邻域关系引入粗糙集理论, Yao[5]对邻域近似空间的性能进行了分析. Hu 等[6]使用邻域粗糙集模型, 为邻域分类器构造了一种统一理论结构, 针对混合型数据提出了一种基于邻域依赖度函数的特征选择算法(NDFS). 邻域依赖度函数判定一个对象对决策分类的贡献度太严格, 若对象的邻域不一致, 则将该对象对决策分类的贡献度视为零. Jing 等[7]放宽了这个限制, 把满足某一条件的不一致邻域也视为一致邻域, 提出了基于邻域变精度的特征选择算法

收稿日期: 2015-06-06；**修回日期:** 2015-09-10.

基金项目: 国家自然科学基金项目(61471371)；湖南省自然科学基金项目(2015jj3022)；中国博士后科学基金项目(2012M512168).

作者简介: 何松华(1964–), 男, 教授, 从事雷达信号与信息处理、精确制导、数据密集型企业的数据挖掘与商业智能等研究；康婵娟(1991–), 女, 硕士生, 从事智能信息处理、粗糙集理论的研究.

(VPT-NRS). Hu 等[8]进一步放松了这个限制, 把满足相对多数决策原则的不一致邻域也视为一致邻域, 提出了基于邻域辨识率的特征选择 (NRFS) 算法.

以上基于邻域的约简算法通常是基于集合的不确定性 (粗糙集合边界的大小) 来定义属性的重要性, 而在经典粗糙集理论中, 不确定性分为知识不确定性和集合不确定性, 两者分别通过考虑属性对论域中不确定分类子集和确定分类子集的影响来衡量属性的重要性[9], 两者具有较强的互补性. 近年来, 研究者们开始综合考虑两方面的不确定性来定义属性的重要性. 江峰等[10-11]提出了一种基于近似决策熵的属性约简算法 (ADEAR) 和基于相对决策熵的属性约简算法 (RDEAR), ADEAR 利用近似精度和条件熵来定义属性的重要性, RDEAR 利用分类精度和近似粗糙度定义的熵来定义属性的重要性, 比现有的属性重要性定义更加全面, 信息更丰富, 但它们只能处理完备信息系统, 且不适合处理数值型数据. 于是, 在 2011 年, Hu 等[12]首次提出了邻域信息熵的概念. Chen 等[13]在 2014 年提出了一种基于邻域熵的决策表约简 (NEFS), 可以用于完备信息系统中的数值型属性, 但对于不完备信息系统或具有混合属性的信息系统则无能为力. 在实际应用中, 由于对数据理解、数据测量或数据获取等方面的限制, 使得获取到的知识通常是不完备的信息系统[14], 因而直接对不完备信息系统进行属性约简便成为粗糙集理论向实用化方向发展的关键.

基于此, 本文在现有的两种互补的不确定性度量基础上, 针对邻域信息系统定义了一种更加全面的属性集成不确定性度量 —— 邻域组合测度, 并提出了一种基于邻域组合测度的属性约简算法. 该算法不仅能够处理符号型和数值型属性共存的混合信息系统, 也可以用于不完备信息系统.

1 邻域系统的相关概念

定义 1[15] $I=(U,A,V,f,\delta)$ 是一个邻域信息系统. 其中: U 是对象 (样本或实体) 的非空有限集合, 称为论域; A 是非空有限的属性集合; V 是所有属性值域, 即 $V=\bigcup\limits_{a\in A}V_a$, V_a 表示属性 a 所有可能取值的集合; $f:U\times A\to V$ 是一个信息函数; $\delta\,(0\leqslant\delta\leqslant 1)$ 是邻域阈值, 用来确定邻域的大小. 若 $A=C\bigcup D$ 且 $C\bigcap D=\varnothing$, 即 $I=(U,C,D,V,f,\delta)$, 则称其为邻域决策信息系统, C 表示条件属性集, D 表示决策属性集. 本文中, 邻域决策信息系统简写为 $I=(U,C,D,\delta)$.

定义 2[6] 给定邻域决策信息系统 $I=(U,C,D,\delta)$, $\forall B\subseteq C$, 则 B 的 δ 邻域关系为

$$\mathrm{NR}_\delta(B)=\{(x,y)\in U\times U|D_B(x,y)\leqslant\delta\},\quad(1)$$

其中 $D_B(x,y)$ 表示对象 x 与 y 之间的距离, 即距离度量函数. 本文中用 $\dfrac{U}{\mathrm{NR}_\delta(B)}$ 表示 U 上基于 B 的邻域分类.

考虑到实际数据集的复杂性, 即可能同时存在数值型属性和符号型属性, 也可能存在缺失属性值, 本文使用文献 [6] 中的距离函数, 即

$$D_B(x,y)=\sqrt{\sum_{l=1}^{N}d_{a_l}^2(x,y)}.\quad(2)$$

其中: $C=\{a_1,a_2,\cdots,a_N\}$; $1\leqslant l\leqslant N$;

$$d_{a_l}=\begin{cases}0,\ f(x,a_l)=*\text{或}f(y,a_l)=*\text{或}f(x,a_l)=f(y,a_l);\\ 1,\ a_l\text{为符号型且}f(x,a_l)\neq f(y,a_l);\\ \dfrac{|f(x,a_l)-f(y,a_l)|}{\max_{a_l}-\min_{a_l}},\ \text{otherwise}.\end{cases}\quad(3)$$

$f(x,a_l)=*$ 表示对象 x 在属性 a_l 上的值未知, 即由该数据集组成的系统是不完备信息系统. 邻域关系仅满足自反性和对称性, 是一种相似关系. 本文在处理不完备信息系统时, 如果某个对象属性值未知, 则认为该对象与其他对象的距离值为 0.

定义 3[6] 给定邻域信息系统 $I=(U,A,V,f,\delta)$, $\forall x\in U,B\subseteq A$, 则 x 在 B 上的 δ 邻域为

$$N_B^\delta(x)=\{y|x,y\in U,D_B(x,y)\leqslant\delta\}.\quad(4)$$

定理 1[6] 给定邻域信息系统 $I=(U,A,V,f,\delta)$, $\forall x\in U,B\subseteq A$, x 关于 B 的 δ 邻域 $N_B^\delta(x)$ 满足

1) $x\in N_B^\delta(x)$;

2) 若 $y\in N_B^\delta(x)$, 则 $x\in N_B^\delta(y)$;

3) $\bigcup\limits_{x\in U}N_B^\delta(x)=U$.

定理 2[6] 给定邻域信息系统 $I=(U,A,V,f,\delta)$, $\forall x\in U,P,Q\subseteq A$, 则有

1) 若 $P\subseteq Q$, 则 $N_P^\delta(x)\supseteq N_Q^\delta(x)$;

2) 若 $0\leqslant\delta_1\leqslant\delta_2\leqslant 1$, 则 $N_P^{\delta_1}(x)\subseteq N_P^{\delta_2}(x)$.

定义 4[6] 给定邻域信息系统 $I=(U,A,V,f,\delta)$, $\forall X\subseteq U,B\subseteq A$, 则 X 关于 B 的 δ 上近似和下近似定义为

$$\overline{N}_B^\delta(X)=\{x|x\in U,N_B^\delta(x)\bigcap X\neq\varnothing\},\quad(5)$$

$$\underline{N}_B^\delta(X)=\{x|x\in U,N_B^\delta(x)\subseteq X\}.\quad(6)$$

定义 5[15] 给定邻域决策信息系统 $I=(U,C,D,\delta)$, 决策属性 D 对于论域 U 的划分为 $U/D=\{D_1,D_2,\cdots,D_m\}$, $\forall B\subseteq C$, U/D 相对于 B 的邻域近似精度和邻域近似粗糙度为

$$\mathrm{Acapp}_B^{\delta}(D)=\frac{\sum\limits_{D_i\in U/D}|\underline{N}_B^{\delta}(D_i)|}{\sum\limits_{D_i\in U/D}|\overline{N}_B^{\delta}(D_i)|},\tag{7}$$

$$\mathrm{Rapp}_B^{\delta}(D)=1-\mathrm{Acapp}_B^{\delta}(D),\tag{8}$$

其中 $|X|$ 表示集合 X 中元素的个数.

邻域近似精度 $\mathrm{Acapp}_B^{\delta}(D)$ 刻画了信息系统中包含有效知识的多少, $\mathrm{Acapp}_B^{\delta}(D)$ 越大, 表明系统包含的有效知识越多, 系统分类能力越强; $\mathrm{Acapp}_B^{\delta}(D)$ 越小, 则系统分类能力越弱. 而邻域近似粗糙度的性质刚好与邻域近似精度相反[15].

定理 3[15]　给定邻域决策信息系统 $I=(U,C,D,\delta)$, $P\subseteq Q\subseteq C$, 则有 $\mathrm{Acapp}_P^{\delta}(D)\leqslant\mathrm{Acapp}_Q^{\delta}(D)$, $\mathrm{Rapp}_P^{\delta}(D)\geqslant\mathrm{Rapp}_Q^{\delta}(D)$.

2　邻域组合测度

为了更加精确地度量粗糙集的不确定性, 文献[9]在完备信息系统中提出了一种更全面的属性重要性度量——集成不确定性度量, 但它并不适用于不完备信息系统[9]. 本节首先在邻域关系下给出一种新的能适用于完备和不完备信息系统的知识不确定性度量的定义, 然后结合现有的集合不确定性度量, 构建一个新的不确定性度量函数——邻域组合测度. 下面先给出知识不确定性度量, 即邻域粒度测度的定义.

定义 6　给定邻域决策信息系统 $I=(U,C,D,\delta)$, $\forall B\subseteq C$, $N_B^{\delta}(x)$ 为 x 关于 B 的 δ 邻域, 决策属性 $D=\{d\}$, $x_i\in U$, 则 B 的邻域粒度测度定义为

$$\mathrm{NG}(B)=1-\frac{2}{|U|}\Big(1-\sum_{i=1}^{|U|}P^2(x_i)\Big).\tag{9}$$

其中 $P(x_i)=|N_B^{\delta}(x_i)|/|U|$, $0\leqslant P(x_i)\leqslant 1$, 表示对象 x_i 的 δ 邻域内对象数目与论域内对象数目的比值. 邻域粒度测度 $\mathrm{NG}(B)$ 反映了属性 B 的分类能力, $\mathrm{NG}(B)$ 值越小, 属性的区分能力就越好.

由定义 5 和定义 6 可知, 邻域近似精度刻画了粗糙集边界域的大小, 而邻域粒度测度则度量了知识对论域划分粒度的大小. 下面将知识不确定性度量和集合不确定性度量组合, 定义新的邻域组合测度如下.

定义 7　给定邻域决策信息系统 $I=(U,C,D,\delta)$, $B\subseteq C$, $N_B^{\delta}(x)$ 为 x 关于 B 的 δ 邻域, 决策属性 $D=\{d\}$, d 的值域为 $V_D=\{d_1,d_2,\cdots,d_l\}$, $U=\{x_1,x_2,\cdots,x_{|U|}\}$, 则 B 的邻域组合测度定义为

$$\mathrm{NCM}_B(D)=\frac{\mathrm{Acapp}_B^{\delta}(D)}{\mathrm{NG}(B)}.\tag{10}$$

由定义 7 可知, 邻域组合测度同时考虑了集合的不确定性和知识的不确定性, 相比于现有文献中单一的不确定性测度, 包含的信息量更丰富, 是一种更加全面的集成不确定性度量. 由式 (1) ∼ (3) 可知, 邻域关系下的邻域组合测度不仅适用于不完备信息系统, 而且对于包含符号型和数值型数据的混合信息系统也适用. 因此, 邻域组合测度的适用性更为广泛.

定理 4　给定邻域决策信息系统 $I=(U,C,D,\delta)$, $\forall B\subseteq C$, $a\in C-B$, 有

$$\mathrm{NCM}_B(D)\leqslant\mathrm{NCM}_{B\bigcup\{a\}}(D).$$

证明　由定理 2 可得

$$1\leqslant|N_{B\bigcup\{a\}}^{\delta}(x)|\leqslant|N_B^{\delta}(x)|\leqslant|U|,$$

即

$$\frac{1}{|U|^2}\leqslant\Big(\frac{|N_{B\bigcup\{a\}}^{\delta}(x_i)|}{|U|}\Big)^2\leqslant\Big(\frac{|N_B^{\delta}(x_i)|}{|U|}\Big)^2\leqslant 1,$$

所以有

$$\frac{1}{|U|}\leqslant\sum_{i=1}^{|U|}\Big(\frac{|N_{B\bigcup\{a\}}^{\delta}(x_i)|}{|U|}\Big)^2\leqslant\sum_{i=1}^{|U|}\Big(\frac{|N_B^{\delta}(x_i)|}{|U|}\Big)^2\leqslant|U|.$$

因此

$$1-|U|\leqslant 1-\sum_{i=1}^{|U|}\Big(\frac{|N_B^{\delta}(x_i)|}{|U|}\Big)^2\leqslant 1-\sum_{i=1}^{|U|}\Big(\frac{|N_{B\bigcup\{a\}}^{\delta}(x_i)|}{|U|}\Big)^2\leqslant 1-\frac{1}{|U|}.$$

由定义 6 有

$$1-\frac{2}{|U|}+\frac{2}{|U|^2}\leqslant\mathrm{NG}(B\bigcup\{a\})\leqslant\mathrm{NG}(B)\leqslant 3-\frac{2}{|U|},$$

所以有

$$0<\mathrm{NG}(B\bigcup\{a\})\leqslant\mathrm{NG}(B).$$

又由定理 3 有

$$0\leqslant\mathrm{Acapp}_B^{\delta}(D)\leqslant\mathrm{Acapp}_{B\bigcup\{a\}}^{\delta}(D),$$

由定义 7 可知

$$0\leqslant\frac{\mathrm{Acapp}_B^{\delta}(D)}{\mathrm{NG}(B)}\leqslant\frac{\mathrm{Acapp}_{B\bigcup\{a\}}^{\delta}(D)}{\mathrm{NG}(B\bigcup\{a\})},$$

因此

$$0\leqslant\mathrm{NCM}_B(D)\leqslant\mathrm{NCM}_{B\bigcup\{a\}}(D).$$

定理 4 得证. □

定理 4 表明, 邻域组合测度的大小随着条件属性集 B 中元素个数的增加而单调增加, 这对于构建基于前向添加搜索策略的约简算法很重要[9]. 下面利用邻域组合测度分别给出属性重要性度量和属性约简的定义.

定义 8 给定邻域决策信息系统 $I=(U,C,D,\delta)$, $\forall B\subseteq C$, $a\in C-B$, 则属性 a 在决策表 I 中相对于 B 的重要性定义为

$$\mathrm{Sig}(a,B,D)=\mathrm{NCM}_{B\bigcup\{a\}}(D)-\mathrm{NCM}_B(D). \quad (11)$$

由定义 7 和定义 8 可知, $\mathrm{Sig}(a,B,D)$ 表示增加属性 a 后对于条件属性集 B 重要性的提高程度, $\mathrm{Sig}(a,B,D)$ 越大, a 对 B 越重要.

定义 9 给定邻域决策信息系统 $I=(U,C,D,\delta)$, $\forall B\subseteq C$, 若满足: 1) $\mathrm{NCM}_B(D)=\mathrm{NCM}_C(D)$; 2) $\forall b\in B$, 都有 $\mathrm{NCM}_{B-\{b\}}(D)<\mathrm{NCM}_C(D)$. 则称 B 为条件属性 C 在邻域决策信息系统中相对于决策属性 D 的一个相对约简. C 的所有约简的交集称为 C 的核 (core).

定义 9 给出了基于邻域组合测度约简的定义. 其中: 条件 1) 保证了约简后的决策信息系统与原决策信息系统具有相同的信息量; 条件 2) 保证了所得的约简是最紧凑的. 由定义 9 可知, 基于邻域组合测度的约简算法的目标就是寻找和原决策信息系统具有相同邻域组合测度的最小条件属性集.

3 基于邻域组合测度的属性约简算法

根据邻域组合测度的单调性原理, 以基于邻域组合测度的特征重要度为启发信息, 构建基于邻域组合测度的属性约简算法.

首先给出计算邻域组合测度步骤如下.

算法 1 计算邻域组合测度.

输入: 邻域决策信息系统 $I=(U,C,D,\delta)$, $B\subseteq C$, $U/\mathrm{IND}(D)=\{D_1,D_2,\cdots,D_m\}$, $U=\{x_1,x_2,\cdots,x_n\}$, 邻域半径 δ;

输出: 邻域组合测度 $\mathrm{NCM}_B(D)$.

Step 1: 计算邻域集合及邻域粒度测度.

Step 1.1: 初始化 $\mathrm{NG\_tmp}=0$, 对于每个对象 x_i $(1\leqslant i\leqslant n)$ 执行:

1) 计算 x_i 的 δ 邻域

$$N_B^\delta(x_i)=\{y|x_i,y\in U,D_B(x_i,y)\leqslant\delta\},$$

其中 $D_B(x_i,y)$ 为对象 x_i 与 y 之间的距离;

2) 计算 $P(x_i)$;

3) 计算 $\mathrm{NG\_tmp}=\mathrm{NG\_tmp}+P(x_i)\times P(x_i)$.

Step 1.2: 计算邻域粒度测度

$$\mathrm{NG}(B)=1-2\times\frac{1-\mathrm{NG\_tmp}}{|U|}.$$

Step 2: 对于每个 D_j $(1\leqslant j\leqslant m)$, 计算 $\underline{N}_B^\delta(D_j)$ 和 $\overline{N}_B^\delta(D_j)$.

Step 3: 计算邻域近似精度

$$\mathrm{Acapp}_B^\delta(D)=\frac{\sum\limits_{D_j\in U/D}|\underline{N}_B^\delta(D_j)|}{\sum\limits_{D_j\in U/D}|\overline{N}_B^\delta(D_j)|}.$$

Step 4: 计算邻域组合测度

$$\mathrm{NCM}_B(D)=\frac{\mathrm{Acapp}_B^\delta(D)}{\mathrm{NG}(B)}.$$

Step 5: 返回 $\mathrm{NCM}_B(D)$.

因为算法 1 中 Step 1.1 的 1) 在计算两个对象的距离时, 既考虑了完备数据又考虑了不完备数据, 所以由算法 1 得到的邻域组合测度对于不完备数据集也适用.

算法 2 基于邻域组合测度的属性约简算法 (NCMAR).

输入: 邻域决策信息系统 $I=(U,C,D,\delta)$, 误差参数 ctrl_k, 邻域半径 δ;

输出: 约简 Red.

Step 1: 初始化 $\mathrm{Red}=\varnothing$, $\mathrm{NCM}_{\mathrm{Red}}(D)=0$.

Step 2: 计算条件属性集 C 的邻域组合测度 $\mathrm{NCM}_C(D)$.

Step 3: 如果 $\mathrm{NCM}_{\mathrm{Red}}(D)\neq\mathrm{NCM}_C(D)$, 则执行:

1) $\forall a\in C-\mathrm{Red}$, 计算 $\mathrm{NCM}_{\mathrm{Red}\bigcup\{a\}}(D)$;

2) 计算 $\mathrm{Sig}(a,\mathrm{Red},D)$;

3) 选择最大属性重要度对应的属性 a_k 作为约简属性, 如果同时存在多个属性 a_k 满足要求, 则选择满足条件的第 1 个属性;

4) $\mathrm{Red}=\mathrm{Red}\bigcup\{a_k\}$;

5) 计算邻域组合测度 $\mathrm{NCM}_{\mathrm{Red}}(D)$;

6) 如果 $\mathrm{Sig}(a_k,\mathrm{Red},D)<\mathrm{ctrl}\_k$, 则转至 Step 4, 否则转至 1).

Step 4: 输出 Red.

考虑到计算过程中的误差, 在约简算法中引入了误差参数 ctrl_k, 当 NCM 的增量值小于 ctrl_k 时, 则认为 NCM 不再增长, 即算法找到了最终约简, 算法终止.

4 实验结果

为了验证本文所提算法的有效性, 将 NCMAR 算法与如下 3 个有代表性的约简算法在约简属性数量和分类精度方面进行比较:

1) 基于邻域依赖度的属性约简算法 (NDFS)[6];

2) 基于邻域熵的属性约简算法 (NEFS)[13];

3) 基于相对决策熵的属性约简算法 (RDEAR)[11].

选用 UCI 机器学习库中的 9 个数据集进行测试, 数据集见表 1. 其中: Heart 和 Hepatitis 数据集的条件

属性是混合型的(既包括数值型属性又包括符号型属性); Breast-cancer、Zoo 和 Lymphography 数据集的条件属性是符号型数据集; Iris、Ecoli、Wdbc 和 Sonar 数据集的条件属性是数值型的; Breast-cancer 和 Hepatitis 数据集的条件属性中含有缺失值, 是不完备数据集.

表 1　UCI 数据集描述

序号	数据集名称	对象数	属性数	属性类型	完备性	类别数
1	Iris	150	4	数值	完备	2
2	Heart	270	13	混合	完备	2
3	Ecoli	336	7	数值	完备	8
4	Wdbc	569	30	数值	完备	2
5	Sonar	208	60	数值	完备	2
6	Lymphography	148	18	符号	完备	4
7	Zoo	101	16	符号	完备	7
8	Breast-cancer	699	9	符号	不完备	2
9	Hepatitis	155	19	混合	不完备	2

由于 NEFS 算法和 RDEAR 算法只能用于完备信息系统, 在进行属性约简之前, 需对不完备数据进行完备化处理. 本文将数据集中的未知值替换为相应属性值的平均值; RDEAR 算法只能处理符号型数据, 需要对数值型数据集先进行离散化处理, 本文直接采用 WEKA 软件中的 weka.filters.unsupervised.attribute.Discretize 方法对其离散化, 量化间距选为 6. 文中参数 $\delta = 0.14$, $\mathrm{ctrl}\_k = 0.001$ (参数 δ 和 ctrl_k 通过多次实验得出). 为了减少属性量纲不一致对约简结果带来的影响, 计算样本邻域时, 将所有数值型属性标准化到 $[0,1]$ 区间. 本文算法采用 Matlab 语言实现, 硬件环境为 Intel 处理器 2.0 GHz, 2 GB 内存.

首先比较 4 种算法在不同数据集上约简结果的紧凑性, 结果如表 2 所示. 由表 2 可知, 4 种算法都有效地实现了对数据的约简. 相比于其他 3 种约简算法, NCMAR 算法在大部分数据集上(除了 Breast-cancer 数据集之外) 都获得了较紧凑的约简.

表 2　不同算法属性约简结果比较

数据集	约简所包含的属性数目			
	NDFS	NEFS	RDEAR	NCMAR
Iris	4	4	4	4
Heart	9	12	8	8
Ecoli	6	6	6	6
Wdbc	12	22	10	8
Sonar	6	10	6	6
Lymphography	6	13	7	6
Zoo	5	12	5	5
Breast-cancer	3	8	4	4
Hepatitis	3	6	3	2

现有约简算法主要考虑了算法的复杂度和数据约简程度, 但对于分类问题, 选择的属性的分类能力更为重要[16]. 为了进一步验证约简结果的分类能力, 本文使用 Weka 软件自带的 JRIP 和 oneR 分类器来评价 4 种算法属性约简的质量, 分类精度如表 3 和表 4 所示.

表 3　4 种算法在 JRIP 下的分类精度　　%

数据集	分类精度			
	NDFS	NEFS	RDEAR	NCMAR
Iris	97.5	97.5	97.5	97.5
Heart	93.9	96.1	93.0	99.3
Ecoli	89.6	89.6	89.6	89.6
Wdbc	97.4	98.1	96.3	98.1
Sonar	86.1	85.0	84.8	92.0
Lymphography	86.0	90.2	87.1	90.5
Zoo	96.5	86.9	97.2	99.1
Breast-cancer	72.0	76.2	72.0	72.0
Hepatitis	63.0	80.2	63.0	83.2
Average	86.89	88.87	86.72	91.26

表 4　4 种算法在 oneR 下的分类精度　　%

数据集	分类精度			
	NDFS	NEFS	RDEAR	NCMAR
Iris	96.2	96.2	96.2	96.2
Heart	78.3	78.3	75.5	78.3
Ecoli	65.2	65.2	65.2	65.2
Wdbc	92.4	92.4	92.4	92.4
Sonar	75.9	76.0	75.9	75.9
Lymphography	75.6	75.6	75.6	75.6
Zoo	70.7	70.7	70.7	70.7
Breast-cancer	72.0	70.3	72.0	72.0
Hepatitis	79.3	79.3	79.3	79.8
Average	78.40	78.22	78.09	78.46

从表 3 和表 4 可以看出:

1) 对于包含有数值属性的 Iris、Heart、Ecoli、Wdbc、Sonar 和 Hepatitis 数据, 由于 RDEAR 算法在属性约简之前需要对数值型属性进行离散化, 散化步骤不仅加大了算法的复杂度, 而且会引起信息的丢失, 最终导致分类精度下降[17]. 由表 3 和表 4 可知, 由于 NDFS 算法、NEFS 算法和 NCMAR 算法可以直接处理数值型数据, 在平均分类精度上明显优于 RDEAR 算法. 本文算法在大部分数据集下的分类精度均优于或等于其他 3 类算法.

2) 对于不完备数据集 Breast-cancer 和 Hepatitis, 本文算法的分类精度与其他 3 类算法相当. 当用分类器 JRIP 对数据集 Breast-cancer 进行分类时, 本文算法的分类精度低于 NEFS 算法, 这是由于 NEFS 算法求得的约简属性个数为 8 个, 远远多于本文约简属性个数 4 个. 需要指出的是, NEFS 算法和 RDEAR 算法不能直接处理不完备数据, 因而约简前要进行完备化, 一定程度上加剧了算法的复杂度.

3) 对于完备的符号型数据集 Lymphography 和 Zoo, 相比于 NDFS 算法、NEFS 算法和 RDEAR 算法, 本文算法在获得紧凑约简的同时, 在两种分类器下的分类精度也相对较高.

4) 结合集合不确定性和知识不确定性的邻域组合测度所包含的信息量更丰富, 保证了本文约简算法

获得紧凑约简的同时, 具有较高的分类精度.

由以上的实验分析可知, 本文提出的约简算法不仅能够处理混合型数据, 而且适用于不完备信息系统, 在获得紧凑约简的同时, 保持了较高的分类精度.

5 结 论

粗糙集理论在不完备信息系统中的应用是将其进一步推向实用的关键之一[14], 目前对不完备信息系统的直接处理仍缺乏完备的理论支持. 本文将邻域系统中两种互补的不确定性度量进行组合, 定义了一种更加全面的属性集成不确定性度量; 在此基础上提出了一种基于邻域组合测度的属性约简 (NCMAR) 算法, 该算法不仅能够处理符号和数值属性共存的混合信息系统, 也可以用于不完备信息系统. 通过 UCI 数据集上的实验表明, 本文所提出的约简算法不仅能够获得紧凑的约简, 而且具有较好的分类性能.

参考文献(References)

[1] Pawlak Z. Rough sets[J]. Int J of Computer and Information Science, 1982, 11(5): 341-356.

[2] Liu G L. Using one axiom to characterize rough set and fuzzy rough set approximation[J]. Information Science, 2013, 223: 285-296.

[3] Zhang Y L, Luo M K. Relationships between covering-based and relation-based rough sets[J]. Information Science, 2013, 225: 55-71.

[4] Lin T Y. Neighborhood systems: A qualitative theory for fuzzy and rough sets[C]. Advances in Machine Intelligence and Soft-Computing. Durham: Duke University, 1997: 132-155.

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

[6] Hu Q H, Liu J F, Wu C X. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Science, 2008, 178(18): 3577-3594.

[7] Jing S, She K, Ali S. A universal neighbourhood rough sets model for knowledge discovering from incomplete heterogeneous data[J]. Expert Systems, 2013, 30(1): 89-96.

[8] Hu Qinghua, Pedrycz W, Yu D, et al. Selecting discrete and continuous features based on neighborhood decision error minimization[J]. IEEE Trans on Systems, Man, and Cybernetics, Part B, 2010, 40(1): 137-150.

[9] Teng Shu-hua, Lu Min, Yang A-feng, et al. Efficient attribute reduction from the viewpoint of discernibility[J]. Information Sciences, 2016, 326: 297-314.

[10] 江峰, 王莎莎, 杜军威, 等. 基于近似决策熵的属性约简[J]. 控制与决策, 2015, 30(1): 65-70.
(Jiang F, Wang S S, Du J W, st al. Attribute reduction based on approximation decision entropy[J]. Control and Decision, 2015, 30(1): 65-70.)

[11] Jiang Feng, Sui Yue-fei, Zhou Lin. A relative decision entropy-based feature selection approach[J]. Pattern Recognition, 2015, 48(7): 2151-2163.

[12] Hu Qinghua, Zhang Lei, David Zhang, et al. Measuring relevance between discrete andcontinuous features based on neighborhood mutual information[J]. Expert Systems with Applications, 2011, 38(9): 10737-10750.

[13] Yumin Chen, Keshou Wu, Xuhui Chen, et al. An entropy-based uncertainty measurement approach in neighborhood systems[J]. Information Science, 2014, 279: 239-250.

[14] 徐久成, 沈钧毅, 王国胤. Rough 集之间的相似度量[J]. 计算机科学, 2003, 30(10): 55-60.
(Xu J C, Shen J Y, Wang G Y. Measure of Similarity between Rough Sets[J]. Computer Science, 2003, 30(10): 55-60.)

[15] 唐朝晖, 陈玉明. 邻域系统的不确定性度量方法[J]. 控制与决策, 2014, 29(4): 691-695.
(Tang C H, Chen Y M. Neighborhood system uncertainty measurement approaches[J]. Control and Decision, 2014, 29(4): 691-695.)

[16] 滕书华, 鲁敏, 杨阿峰, 等. 基于一般二元关系的粗糙集加权不确定性度量[J]. 计算机学报, 2014, 37(3): 649-665.
(Teng S H, Lu M, Yang A F, et al. A Weighted uncertainty measure of rough sets based on general binary relation[J]. J of Computer, 2014, 37(3): 649-665.)

[17] 谢宏, 程浩忠, 牛东晓. 基于信息熵的粗糙集连续属性离散化算法[J]. 计算机学报, 2005, 28(9): 1570-1574.
(Xie H, Cheng H Z, Niu D X. Discretization of Continuous Attributes in Rough Set Theory Based on Information Entropy[J]. Chinese J of Computer, 2005, 28(9): 1570-1574.)

(责任编辑: 齐 霁)