DOI:10.11992/tis.201704032
网络出版地址:http://kns.cnki.net/kcms/detail/23.1538.TP.20170703.1854.016.html

连续型数据的辨识矩阵属性约简方法

冯丹[1,2],黄洋[2],石云鹏[2],王长忠[2]
(1. 国网葫芦岛供电公司 信息通信分公司,辽宁 葫芦岛 125000; 2. 渤海大学 数理学院,辽宁 锦州 121000)

摘 要:属性约简是粗糙集理论在数据处理方面的重要应用,已有的针对连续型数据的属性约简算法主要集中在基于正域的贪心算法,该方法只考虑了一致样本和其他样本的可辨识性,而忽略了边界样本点间可区分性。为了克服基于正域算法的缺点,提出了连续型数据的辨识矩阵属性约简模型,该模型不但考虑了正域样本的一致性,同时考虑了边界样本的可分性。基于该模型,分析了属性约简结构,定义了辨识矩阵来刻画特征子集的分类能力,构造了实值型数据的属性约简启发式算法,并利用 UCI 标准数据集进行了验证。理论分析和实验结果表明,提出的算法能够有效地处理连续型数据,提高了数据的分类精度。

关键词:邻域关系;粗糙集;属性约简;辨识矩阵;启发式算法
中图分类号:TP391;TP274 **文献标志码**:A **文章编号**:1673-4785(2017)03-0371-06

中文引用格式:冯丹,黄洋,石云鹏,等.连续型数据的辨识矩阵属性约简方法[J]. 智能系统学报, 2017, 12(3): 371-376.
英文引用格式:**FENG Dan, HUANG Yang, SHI Yunpeng, et al. A discernibility matrix-based attribute reduction for continuous data[J]. CAAI transactions on intelligent systems**, 2017, 12(3): 371-376.

A discernibility matrix-based attribute reduction for continuous data

FENG Dan[1,2], HUANG Yang[2], SHI Yunpeng[2], Wang Changzhong[2]
(1. Information and Communication Branch, State Grid Power Supply Company of Huludao, Huludao 125000, China; 2. College of Mathematics and Physics, Bohai University, Jinzhou 121000, China)

Abstract: In data processing, attribute reduction is an important application of rough set theory. The existing methods for continuous data mainly concentrate on the greedy algorithms based on the positive region. These methods take account of only the identifiability between consistent samples and other samples while ignoring distinguishability among the boundary samples. To overcome the disadvantage based on the positive domain algorithm, this paper proposed a new method for attribute reduction using a discernibility matrix. The model considers not only the consistency of samples in the positive region but also the reparability of boundary samples. On this basis, this paper analyzes the structure of attribute reduction and defines a discernibility matrix to characterize the discernibility ability of a subset of attributes. Next, an attribute reduction algorithm was designed based on the discernibility matrix. The validity of the proposed algorithm was verified using UCI standard data sets and theoretical analysis.

Keywords: neighborhood relation; rough set; attribute reduction; discernibility matrix; heuristic algorithm

粗糙集理论是由波兰数学家 Z. Pawlak[1] 于 1982 年提出的,它是一种处理不确定性知识的数据分析理论。目前粗糙集理论已被广泛应用于人工智能、过程控制、数据挖掘、决策支持以及知识发现等领域。属性约简是粗糙集理论的研究内容之一,它是利用粗糙集进行数据挖掘、规则抽取的理论基础。其主要思想是在确保信息表的分类能力不变的情况下,删除掉不必要属性,保留必要属性,从而导出问题的分类规则。属性约简无疑是在获取规则的过程中最重要的核心问题,历来受到大家的关注。

传统的粗糙集理论[1-3]是基于等价关系来描述的,由等价关系粒化样本空间形成信息粒子,构造论域上的上、下近似算子,进而研究知识约简与知识获取问题。但是经典粗糙集只适用于离散型的数据,对于连续型数据,必须通过离散化才能处理。然而,数据的离散化会导致大量信息丢失,从而使

收稿日期:2017-04-23. 网络出版日期:2017-07-03.
基金项目:国家自然科学基金项目(61572082, 61673396, 61473111, 61363056);辽宁省教育厅项目(LZ2016003);辽宁省自然科学基金项目(2014020142);辽宁省高校创新团队计划项目(LT2014024).
通信作者:王长忠.E-mail:changzhongwang@126.com.

计算的结果不能准确地反映分类信息。为此,经典粗糙集理论被进行了多角度的推广,其中包括邻域粗糙集模型[4-11]、优势粗糙集模型[12-13]、覆盖粗糙集模型[14-16]、模糊粗糙集模型[17-22]等。邻域粗糙集模型是重要的推广模型之一,基于此模型,许多学者研究了不同的依赖度函数,并设计了相应的属性约简算法[4-11]。例如,Hu[5]利用邻域的概念定义了样本空间的决策正域,构造了邻域依赖度函数来刻画属性的分类能力,用于处理混合数据的属性约简;Zhao[7]根据数据精度设计了一个自适应性的邻域粗糙集模型,并给出了基于该模型的代价敏感属性约简算法;Chen[8]利用邻域粗糙集模型和信息测度为肿瘤分类进行特征选择;Zhu[9]对邻域粗糙集模型的边界进行分布优化,从而为粒度的选取和组合提供了新方法。然而,邻域粗糙集模型中的决策正域只考虑了一致性样本与其他样本的可辨识性,忽略了边界样本的可分性。因此,基于正域的依赖度函数不能正确地刻画一个属性子集的分类能力。为了克服基于正域算法的缺点,本文提出了基于辨识矩阵的属性约简算法。该算法克服了依赖度算法的局限性。

1 邻域关系决策表的属性约简

设(U,A,F)为信息表,其中,$U=\{x_1,x_2,\cdots,x_n\}$为样本集合,$A=\{a_1,a_2,\cdots,a_m\}$是属性集合,$F=\{f_j: j\leqslant m\}$为 U 和 A 的关系集,$f_j:U\to V_j, j\leqslant m$,$V_j$ 为属性 a_j 的值域。

定义 1 设 $U=\{x_1,x_2,\cdots,x_n\}$ 为样本集,$B\subseteq A$,令

$$M_B=\{(x_i,x_j)\in U\times U:|f_a(x_i)-f_a(x_j)|\leqslant\varepsilon,\forall a\in B\}$$

则称 M_B 为 U 上的邻域关系,称(U,A,F)为基于邻域关系的信息系统,简称邻域信息表,ε 表示相似度阈值。对于任意 $x_i\in U$,令

$$\delta_B(x_i)=\{x_j\in U:(x_i,x_j)\in M_B\}$$

则称 $\delta_B(x_i)$ 为 x_i 关于 M_B 的邻域。

邻域关系可以用一个关系矩阵来表示。设属性 $a_l\in A$,对于任意 $x_i,x_j\in U$,如果 $x_j\in\delta_{a_l}(x_i)$,那么记 $\boldsymbol{N}_l(i,j)=1$,否则记 $\boldsymbol{N}_l(i,j)=0$。因此,一个属性 a_l 唯一地对应一个关系矩阵 $\boldsymbol{N}_l$。而属性集合 A 的关系矩阵可以由公式 $\boldsymbol{N}_A=\cap_{a_l\in A}\boldsymbol{N}_l$ 计算。

定义 2 设$(U,A,F,D=\{d\})$为决策表,其中(U,A,F)为信息表,$A=\{a_1,a_2,\cdots a_m\}$,D 为决策属性。对于任意 $x\in U$,令

$$M_D=\{(x_i,x_j)\in U\times U|d(x_i)=d(x_j)\}$$

$$[x_i]_D=\{x_j\in U|d(x_i)=d(x_j)\}$$

则称 M_D 为 U 上的决策等价关系,$[x_i]_D$ 为 x_i 的关于 M_D 的决策等价类。称(U,A,F,D)为基于邻域关系的决策信息表,简称邻域决策表。若 $M_A\subseteq M_D$,则称(U,A,F,D)是协调的,否则,称它是不协调的。U 上的所有决策等价类构成了 U 的一个划分,记为$U/D=\{[x_i]_D|x_i\in U\}$。

同理,决策等价关系 $\boldsymbol{M}_D$ 可以用关系矩阵 $\boldsymbol{N}_D$ 来表示。即对于任意 $x_i,x_j\in U$,如果 $x_j\in[x_i]_D$,那么记 $\boldsymbol{N}_D(i,j)=1$,否则记 $\boldsymbol{N}_D(i,j)=0$。

设(U,A,F,D)为决策表,$U/D=\{X_1,X_2,\cdots,X_r\}$为决策划分,$B\subseteq A$,对于任意 $X_k\in U/D$,定义 X_k 的下近似为 $\underline{B}(X_k)=\{x_i\in U:\delta_B(x_i)\subseteq X\}$,$D$ 关于 B 的正域定义为 $\mathrm{POS}_B(D)=\cup_{k=1}^{r}\underline{B}(X_k)$。显然,$\mathrm{POS}_B(D)\subseteq\mathrm{POS}_A(D)$。

设 $a_i\in B\subseteq A$,如果 $\mathrm{POS}_B(D)=\mathrm{POS}_{B-\{a_i\}}(D)$,则称 a_i 相对于 D 是 B 中不必要的属性;否则,称 a_i 是 B 中必要的属性。如果 $\mathrm{POS}_B(D)=\mathrm{POS}_A(D)$且 B 中每一个属性相对于 D 都是 B 中必要的,则称 B 是 A 的一个属性约简。

定理 1 设$(U,A,F,D=\{d\})$为决策表,$B\subseteq A$,则 $\mathrm{POS}_B(D)=\mathrm{POS}_A(D)$ 的充要条件是:对于任意 $x_i,x_j\in U$,如果满足以下条件之一,即

1) $x_i\in\mathrm{POS}_A(D)$,$x_j\notin\mathrm{POS}_A(D)\wedge d(x_i)\neq d(x_j)$;

2) $x_i,x_j\in\mathrm{POS}_A(D)\wedge[x_i]_D\cap[x_j]_D=\varnothing$;

则有 $x_j\notin\delta_A(x_i)\Rightarrow x_j\notin\delta_B(x_i)$。

证明 充分性证明。设 $\forall x_i,x_j\in U$,如果 $x_i\in\mathrm{POS}_A(D)$ 和 $x_j\notin\mathrm{POS}_A(D)$ 且 $d(x_i)\neq d(x_j)$,则 $\exists X_0\in U/D$使得 $x_i\in\underline{A}(X_0)$以及 $x_j\notin X_0$,于是 $\delta_A(x_i)\subseteq X_0$,从而 $x_j\notin\delta_A(x_i)$。由于 $\mathrm{POS}_B(D)=\mathrm{POS}_A(D)$,所以对于任意$X_k\in U/D$,都有 $\underline{B}(X_k)=\underline{A}(X_k)$,当然也有 $\underline{A}(X_0)=\underline{B}(X_0)$。由 $x_i\in\underline{A}(X_0)$,可得 $x_i\in\underline{B}(X_0)$,于是 $\delta_B(x_i)\subseteq X_0$,因此 $x_j\notin\delta_B(x_i)$。

如果 $x_i,x_j\in\mathrm{POS}_A(D)$且$[x_i]_D\cap[x_j]_D=\varnothing$,则存在 X_0 和 $X_1\in U/D$ 使得 $X_0\neq X_1$ 且满足 $x_i\in\underline{A}(X_0)$以及 $x_j\notin X_0$。由 $x_i\in\underline{A}(X_0)$可得 $\delta_A(x_i)\subseteq X_0$,从而 $x_j\notin\delta_A(x_i)$。类似地,由于 $\mathrm{POS}_B(D)=\mathrm{POS}_A(D)$,所以 $\underline{A}(X_0)=\underline{B}(X_0)$。由 $x_i\in\underline{A}(X_0)$,可得 $x_i\in\underline{B}(X_0)$。从而 $\delta_B(x_i)\subseteq X_0$,因此 $x_j\notin\delta_B(x_i)$。

必要性证明。设 $x_i\in\mathrm{POS}_A(D)$,所以存在 $X_0\in U/D$ 使得 $\delta_A(x_i)\subseteq X_0$。对于满足 $x_j\notin\mathrm{POS}_A(D)$ 且 $d(x_i)\neq d(x_j)$ 的任意 $x_j\in U$,有 $x_j\notin X_0$,因此 $x_j\notin\delta_A(x_i)$。由于 $x_j\notin\delta_A(x_i)\Rightarrow x_j\notin\delta_B(x_i)$,所以 $\delta_B(x_i)\subseteq X_0$,从而 $x_i\in\mathrm{POS}_B(D)$。于是 $\mathrm{POS}_A(D)\subseteq\mathrm{POS}_B(D)$,

而 $\mathrm{POS}_B(D)\subseteq\mathrm{POS}_A(D)$ 显然成立,因此 $\mathrm{POS}_B(D)=\mathrm{POS}_A(D)$。

另设 $x_i,x_j\in\mathrm{POS}_A(D)$ 且 $[x_i]_D\cap[x_j]_D=\varnothing$,则存在 $X_0,X_1\in U/D(X_0\neq X_1)$ 使得 $\delta_A(x_i)\subseteq X_0=[x_i]_D$ 和 $\delta_A(x_j)\subseteq X_1=[x_j]_D$。由于 $x_j\notin\delta_A(x_i)\Rightarrow x_j\notin\delta_B(x_i)$,所以 $\delta_B(x_i)\subseteq X_0$,从而 $x_i\in\mathrm{POS}_B(D)$。于是 $\mathrm{POS}_A(D)\subseteq\mathrm{POS}_B(D)$,而 $\mathrm{POS}_B(D)\subseteq\mathrm{POS}_A(D)$ 是显然成立的,因此 $\mathrm{POS}_B(D)=\mathrm{POS}_A(D)$。综上所述,结论成立。

根据定理 1 可以定义如下的辨识矩阵。

定义 3　设 $(U,A,F,D=\{d\})$ 为决策表,$U=\{x_1,x_2,\cdots,x_n\}$,$A=\{a_1,a_2,\cdots a_m\}$,令

$$\mathrm{DIS}=\{(x_i,x_j)\mid x_i\in\mathrm{POS}_A(D),x_j\notin\mathrm{POS}_A(D)\wedge d(x_i)\neq d(x_j)\}\cup\{(x_i,x_j)\mid x_i,x_j\in\mathrm{POS}_A(D)\wedge[x_i]_D\cap[x_j]_D=\varnothing\}$$

则称 DIS 为决策表 (U,A,F,D) 的可辨识域。对于任意的样本对 $(x_i,x_j)\in\mathrm{DIS}$,记

$$\mathbf{DM}(i,j)=\begin{cases}\{a_l\in A:x_j\notin\delta_{a_l}(x_i)\}, & (x_i,x_j)\in\mathrm{DIS}\\ A, & (x_i,x_j)\notin\mathrm{DIS}\end{cases}$$

则称 $\mathbf{DM}(i,j)$ 为 x_i,x_j 的辩识集合,称 **DM** 为基于邻域关系的辨识矩阵。

定理 2　设 **DM** 为决策表 $(U,A,F,D=\{d\})$ 的辨识矩阵,$B\subseteq A$,则 B 是决策表的一个约简的充要条件是:B 满足 $B\cap\mathbf{DM}(i,j)\neq\varnothing$, $\forall x_i,x_j\in U$ 的最小子集。

定理 2 说明通过辨识矩阵可以等价地刻画决策表的属性约简。下面给出决策表的属性约简的辨识公式。通过析取和合取运算可以获得决策表的全部约简。

定义 4　设 **DM** 为决策表 (U,A,F,D) 的辨识矩阵,$U=\{x_1,x_2,\cdots,x_n\}$,辨识函数定义为

$$f(U,A,F,D)=\bigwedge_{i,j=1}^{n}(\vee\mathbf{DM}(i,j))$$

定理 3　设 $f(U,A,F,D)$ 为决策表 (U,A,F,D) 的辨识函数,如果通过析取和合取运算,有

$$f(U,A,F,D)=\bigvee_{k=1}^{l}(\wedge B_k)$$

式中:$B_k\subseteq A$,且 B_k 中每个属性只能出现一次。则称 $\{B_k:k\leqslant l\}$ 是 A 的所有约简组成的集类。

A 的所有约简组成的集类记为 $\mathrm{RED}_D(A)=\{B_k:k\leqslant l\}$。

下面通过一个具体的实例来说明应用辨识矩阵方法如何求解邻域决策表的属性约简。

例 1　表 1 是具有 4 种症状 a_1、a_2、a_3、a_4 的某些病例信息,具体描述如表 1 所示。

表 1　病例决策信息表

Table 1　Decision information for cases

序号	a_1	a_2	a_3	a_4	D
x_1	0.66	0.45	0.20	0.82	1
x_2	0.47	0.30	0.06	0.65	1
x_3	0.05	0.80	0.40	0.10	2
x_4	0.35	0.51	0.00	0.52	2
x_5	0.31	0.20	0.15	0.70	1
x_6	0.00	1.00	0.20	0.00	2

取 $\varepsilon=0.25$,根据定义 1 和定义 2,计算关系矩阵 $\boldsymbol{N}_l(i\leqslant4)$、$\boldsymbol{N}_A$ 以及决策关系矩阵 $\boldsymbol{N}_D$ 分别为

$$\boldsymbol{N}_1=\begin{bmatrix}110000\\110110\\001001\\010110\\010110\\001001\end{bmatrix},\boldsymbol{N}_2=\begin{bmatrix}110110\\110110\\001001\\110100\\110010\\001001\end{bmatrix},\boldsymbol{N}_3=\begin{bmatrix}111111\\110111\\101011\\110111\\111111\\111111\end{bmatrix}$$

$$\boldsymbol{N}_4=\begin{bmatrix}110010\\110110\\001001\\010110\\110110\\001001\end{bmatrix},\boldsymbol{N}_A=\begin{bmatrix}110000\\110110\\001001\\010100\\010010\\001001\end{bmatrix},\boldsymbol{N}_D=\begin{bmatrix}110010\\110010\\001101\\001101\\110010\\001101\end{bmatrix}$$

说明 $\boldsymbol{N}_A\not\subset\boldsymbol{N}_D$。由以上计算知,$\mathrm{POS}_A(D)=\{x_1,x_3,x_5,x_6\}$。根据定义 3,得到辨识矩阵如表 2 所示。

表 2　病例决策信息表的辨识矩阵

Table 2　Discernibility matrix of case decisions

序号	x_1	x_2	x_3	x_4	x_5	x_6
x_1	A	A	a_1,a_2,a_4	a_1,a_4	A	a_1,a_2,a_4
x_2	A	A	A	A	A	A
x_3	a_1,a_2,a_4	A	A	A	a_1,a_2,a_4	A
x_4	A	A	A	A	A	A
x_5	A	A	a_1,a_2,a_4	A	A	a_1,a_2,a_4
x_6	a_1,a_2,a_4	a_1,a_2,a_4	A	A	a_1,a_2,a_4	A

所以,可得决策表的辨识函数为

$$f=(a_1\vee a_4)\wedge(a_1\vee a_2\vee a_4)\wedge a_2=(a_1\wedge a_2)\vee(a_2\wedge a_4)$$

因此 $\{a_1,a_2\}$ 和 $\{a_2,a_4\}$ 是病例决策表的两个约简。

2　属性约简算法

经典粗糙集算法是以等价关系作为聚类标准

的。对于数值型数据集,首先进行离散化并构造等价关系。若两个样本在所有属性上取值一样,那么这两个样本就为一类,否则不是一类。而在离散化过程中,避免不了信息流失,而这恰恰是当今研究的热点。本算法通过定义一个距离参数,考虑某两个样本在属性上的相似程度。如果两个样本之间距离小于阈值,则这两个样本就可以聚为一类,这比经典粗糙集的理论显然更多地考虑了样本之间的联系,避免大量信息的流失,从而提高了属性约简的精度。由于利用定理 3 去搜索决策表的全部约简是一个 NP 问题,因此利用本文提出的辨识矩阵的概念来构造一个启发式算法。

算法 辨识矩阵属性约简算法(DISRS)。

输入 $(U,A,D=\{d\})$,阈值 θ //θ 是用于算法停止搜索的阈值。

输出 约简 RED。

1)计算正域 $\mathrm{POS}_A(D)$。

2)计算出关系矩阵 $\boldsymbol{N}_D$,即对于任意的样本 x_i, $x_j\in U$,若 $d(x_i)\neq d(x_j)$,令 $\boldsymbol{N}_D(i,j)=0$, 否则,令 $\boldsymbol{N}_D(i,j)=1$。

3)$\forall a_l\in A$,计算可辨识矩阵 $\boldsymbol{N}_l$,对于任意 $x_i\in \mathrm{POS}_A(D)$ 和任意 $x_j\in U$, 当 $\boldsymbol{N}_D(i,j)=0$ 时,若 $|f_l(x_i)-f_l(x_j)|>\varepsilon$,令 $\boldsymbol{N}_l(i,j)=1$,否则令 $\boldsymbol{N}_l(i,j)=0$;当 $\boldsymbol{N}_D(i,j)=1$,令 $\boldsymbol{N}_l(i,j)=0$。

4)计算 $\boldsymbol{N}_A=\cup\boldsymbol{N}_l$,RED←$A$。

5)$\forall a_l\in A$,计算 $\boldsymbol{N}_{\{\mathrm{red}-a_l\}}$ 和 $\mathrm{sum}(N_{\{\mathrm{red}-a_l\}})$,其中 $\mathrm{sum}(N_{\{\mathrm{red}-a_l\}})$ 表示对矩阵 $\boldsymbol{N}_{\{\mathrm{red}-a_l\}}$ 的行列求和。

6)选择满足 $\mathrm{sum}(\boldsymbol{N}_{\{\mathrm{red}-a_k\}})=\max\limits_i(\mathrm{sum}(N_{\{\mathrm{red}-a_i\}}))$ 的 a_k,令 RED←$\{A-a_k\}$。

7)如果 $(\mathrm{sum}(\boldsymbol{N}_A)-\mathrm{sum}(\boldsymbol{N}_{\mathrm{red}}))/|U|<\theta$,输出约简 RED。否则,转到 5)。

了更好地理解所提出的算法,下面给出该算法的流程图,如图 1。

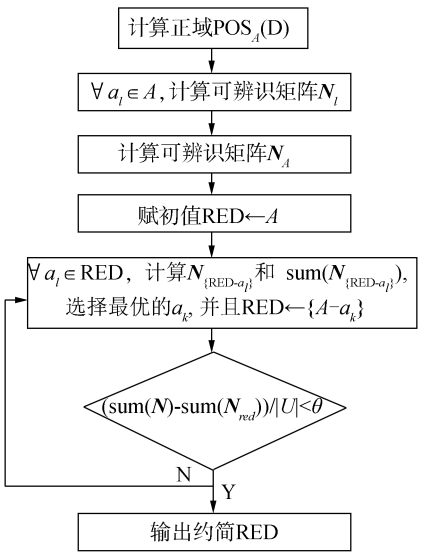

图 1 属性约简算法流程图

Fig. 1 Flow chart of attribute reduction

3 实验分析

为了验证算法的有效性,从一些文献中选出 4 个相关的属性约简方法与本文所提的算法作比较。这 4 个算法分别是经典粗糙集算法(FCMRS)[1]、邻域粗糙集算法(NBRS)[5]、邻域粗糙信息测度算法(NBIM)[8]以及自适应邻域粗糙集模型(APTNB)[7]。本实验主要从算法选择的属性数目和相应的分类精度两个方面进行比较。计算机运行的环境参数为:奔腾双核,CPU E5200 1.90 GHz, RAM 4.0 GB,软件为 MATLAB 2007。

本实验采用 K 近邻(KNN,$K=3$)和支持向量机(RBF-SVM)分类器来评估这些属性约简算法的优劣,RBF-SVM 分类器中参数采用默认值。数据的分类精度是基于十折交叉验证方法计算的。从 UCI 数据集中选择了 5 个数值型的数据集,其属性和分类信息描述如表 3 所示。在本文所提的算法(DISRS)中,停止搜索的阈值 θ 设置为 $\theta=0.005$。为了说明该算法的有效性和可行性,首先对各个数据集进行了属性约简,选取约简数据的最高分类精度所对应的属性数目进行比较,具体结果如表 4 所示,其中 ε 列表示对应数据约简后所取得的最高分类精度的相似度阈值的取值。表 5 和表 6 分别给出了各个数据集属性约简前后的分类精度。

表 3 数据描述

Table 3 Data information

序号	Data sets	Samples	Numerical	Classes
1	Glass	214	10	6
2	Colon	62	1 224	2
3	Wine	178	13	3
4	Wdbc	569	30	2
5	Prostate	102	12 625	2

从表 4 中可以看出,这 5 种属性约简方法都能够有效地对数据集进行属性约简。FCMRS 算法选取的属性数目最少,DISRS 算法次之。表 5 和表 6 表明,DISRS 算法的分类精度最高,FCMRS 算法的分类精度最低。同时,DISRS 算法所对应的分类精度明显好于其他方法。10 次分类任务,DISRS 算法

获得了 8 次最高分类精度,APTNB 获得了 2 次,NBIM 获得了 1 次,而 FCMRS 和 NBRS 算法没有获得最高分类精度。但是 NBRS 算法的性能要比 FCMRS 方法好很多。这说明邻域粗糙集模型在处理连续型数据时比经典粗糙集方法具有更大的优势。造成 FCMRS 算法选择的属性数目少、分类精度低的原因,可能是由于 FCMRS 算法固有的离散化步骤破坏了原有数据的分类信息。而 NBRS、NBIM 和 APTNB 算法避免了离散化步骤,直接利用相似关系粒化数据空间,构造分类目标的依赖度函数用于分类,因此属性约简后的数据分类精度明显高于 FCMRS 算法。然而,NBRS、NBIM 和 APTNB 算法只考虑一致样本和其他样本的可辨识性,忽略了边界样本点间可区分性,因此这 3 种算法的分类精度低于 DISRS 算法精度。而 DISRS 算法克服了 NBRS、NBIM 和 APTNB 算法的缺点,因此,在数据实验分析中取得了较好的性能,即 DISRS 算法的分类精度不仅高于其他算法,而且有效地删减了属性。根据数据实验分析表明,本文提出的算法是行之有效的,达到了理论预期的效果。

表 4　属性约简的结果

Table 4　Result of attribute reduction

Data sets	Raw data	FCMRS	NBRS	NBIM	APTNB	DISRS	ε
Glass	10	5	8	8	7	6	0.225
Colon	1 224	4	10	13	8	8	0.275
Wine	13	6	9	10	6	7	0.425
Wdbc	30	7	16	18	19	12	0.325
Prostate	12 625	2	4	3	4	3	0. 25
平均值	2 780.4	4.8	9.4	10.4	8.8	7.2	

表 5　约简数据的 SVM 后的分类精度

Table5　Classification accuracy of reduced data with SVM　%

Data sets	Raw data	FCMRS	NBRS	NBIM	APTNB	DISRS
Glass	91.58 ± 11. 02	89.33 ± 3. 06	92.32 ± 6. 86	92.56 ± 6.12	93.07 ± 5.33	94.43 ± 4.36
Colon	76.17 ± 17.23	81.07 ± 14.28	82.67 ± 11.78	83.11 ± 12.32	83.46 ± 10.69	84.86 ± 10.11
Wine	95.56 ± 3. 33	92.11 ± 3. 56	96.18 ± 2. 72	95.27 ± 3.64	96.78 ± 2. 22	96.78 ± 3.89
Wdbc	94.03 ± 4. 83	93.06 ± 8. 47	96.81 ± 4. 80	97.20 ± 3.59	97.00 ± 2.94	97.17 ± 3. 12
Prostate	81.23 ± 15. 83	81.86 ± 13. 47	85.86 ± 10. 81	86.55 ± 12.35	88.39 ± 11.88	88.86 ± 9.98
平均值	87.71 ± 10.45	86.09 ± 7.14	90.77 ± 8.04	90.94 ± 7.60	91.74 ± 6.61	92.42 ± 6.29

表 6　约简数据的 3NN 分类精度

Table 6　Classification accuracy of reduced data with 3NN　%

Data sets	Raw data	FCMRS	NBRS	NBIM	APTNB	DISRS
Glass	89.73 ± 6. 72	89.10 ± 5. 00	91.43 ± 5. 49	91.37 ± 4. 87	91.44 ± 3.98	93.34 ± 3. 00
Colon	76.15 ± 16.55	78.33 ± 13.06	81.88 ± 12.08	82.85 ± 11.77	82.76 ± 10.98	83.71 ± 10.23
Wine	94.52 ± 5. 63	92.45 ± 6. 82	96.78 ± 1. 67	97.19 ± 1. 08	96.43 ± 3. 56	97.22 ± 1. 98
Wdbc	94.62 ± 2. 57	92.26 ± 7. 33	97.00 ± 1. 82	97.00 ± 3. 16	97.00 ± 3. 05	97.17 ± 2. 18
Prostate	89.73 ± 6. 72	89.10 ± 5. 00	91.43 ± 5. 49	92.37 ± 4. 87	93.44 ± 3.98	93.34 ± 3. 00
平均值	87.18 ± 9.50	86.80 ± 9.31	90.29 ± 6.86	90.87 ± 6.71	91.02 ± 6.62	91.75 ± 5.57

4　结束语

邻域粗糙集中基于正域的贪心算法只考虑了区分一致性样本和异类样本,忽略了边界样本间的区分性。事实上,一个属性子集的分类能力不仅与一致样本有关,也与边界样本有关。而辨识矩阵的概念正好反映了一组特征的区分能力。本文研究了基于邻域辨识矩阵的属性约简方法,设计了启发式属性约简的算法,并通过 UCI 数据集验证了该算法的有效性。未来的工作将讨论该方法在分类决

策中的应用。

参考文献：

[1] PAWLAK Z. Rough sets [J]. International journal of computer and information sciences, 1982, 11(5): 341-356.

[2] SKOWRON A, RAUSZER C. The discernibility matrices and functions in information systems [C]// Slowinski R. (Ed.), Intelligent Decision Support. Dordrecht, Kluwer Academic Publishers, 1992: 331-362.

[3] MI J S, WU W Z, ZHANG W X. Approaches to knowledge reduction based on variable precision rough sets model [J]. Information sciences, 2004, 159(3/4): 255-272.

[4] WU W Z, ZHANG W X. Neighborhood operator systems and approximations [J]. Information sciences, 2002, 144(1/4): 201-217.

[5] HU Q H, YU D, LIU J F, et al. Neighborhood-rough-set based heterogeneous feature subset selection [J]. Information sciences, 2008, 178(18): 3577-3594.

[6] KIM D. Data classification based on tolerant rough set [J]. Pattern recognition, 2001, 34(8): 1613-1624.

[7] ZHAO H, WANG P, HU Q H. Cost-sensitive feature selection based on adaptive neighborhood granularity with multi-level confidence [J]. Information sciences, 2016, 366: 134-149.

[8] CHEN Y, ZHANG Z, ZHENG J, et al. Gene selection for tumor classification using neighborhood rough sets and entropy measures [J]. Journal of biomedical informatics, 2017, 67:59-68

[9] ZHU P, HU Q H. Adaptive neighborhood granularity selection and combination based on margin distribution optimation [J]. Information sciences, 2013, 249:1-12.

[10] 鲍丽娜,丁世飞, 许新征, 等. 基于邻域粗糙集的极速学习机算法[J]. 济南大学学报, 2015, 29(5): 367-371.
BAO Lina, DING Shifei, XU Xinzheng, et al. Extreme learning machine algorithm based on neighborhood rough sets [J]. Journal of jinan university, 2015, 29(5): 367-371.

[11] 谢娟英, 李楠, 乔子芮. 基于邻域粗糙集的不完整决策系统特征选择算法[J]. 南京大学学报, 2016, 47: 384-390.
XIE Juanying, LI Nan, QIAO Zirui. A feature selection algorithm based on neighborhood rough sets for incomplete information systems [J]. Journal of Nanjing university, 2016, 47:384-390.

[12] 徐伟华. 序信息系统与粗糙集[M]. 北京: 科学出版社, 2013.

[13] GRECO S, MATARAZZO B, SLOWINSKI R. Rough sets methodology for sorting problems in presence of multiple attributes and criteria[J]. European journal of operational research, 2002, 38:247-259.

[14] WANG C, HE Q, CHEN D G, et al. A novel method for attribute reduction of covering decision tables[J]. Information sciences, 2014, 254: 181-196.

[15] WANG C, SHAO M, SUN B, et al. An improved attribute reduction scheme with covering based rough sets[J]. Applied soft computing, 2015, 26(1): 235-243.

[16] ZHU W, WANG F Y. Reduction and maximization of covering generalized rough sets[J]. Information sciences, 2003, 152: 217-230.

[17] DUBOIS D, PRADE H. Rough fuzzy sets and fuzzy rough sets[J]. International journal of general systems, 1990, 17: 191-208.

[18] WANG C, QI Y, HU Q, et al. A fitting model for feature selection with fuzzy rough sets[J]. IEEE transaction on fuzzy systems, 2016,99: 1-1.

[19] WANG C, SHAO M, QIAN Y. Feature subset selection based on fuzzy neighborhood rough sets[J]. Knowledge-based systems, 2016, 111(1): 173-179.

[20] CHEN D G, ZHANG L, ZHAO S Y, et al. A novel algorithm for finding reducts with fuzzy rough sets[J]. IEEE transaction on fuzzy systems, 2013, 20(2): 385-389.

[21] WANG X Z, ZHAI J H, LU S X. Induction of multiple fuzzy decision trees based on rough set technique[J]. Information sciences, 2008, 178(16): 3188-3202.

[22] ZHAO S Y, TSANG C C, CHEN D. Building a rule-based classifier by using fuzzy rough set technique[J]. IEEE transaction on knowledge and data engineering, 2010, 22(5): 624-638.

作者简介：

冯丹,女,1977 年生,高级工程师,主要研究方向为计算机信息管理、数据挖掘。已发表学术论文 10 余篇。

黄洋,女,1994 年生,硕士研究生,主要研究方向为粒计算与数据挖掘。

石云鹏,男,1994 年生,硕士研究生,主要研究方向为粒计算与数据挖掘。