F- 邻域粗糙集及其约简

邓志轩 [1] 郑忠龙 [1] 邓大勇 [1, 2]

摘 要 邻域粗糙集可以直接处理数值型数据, F- 粗糙集是第一个动态粗糙集模型. 针对动态变化的数值型数据, 结合邻域粗糙集和 F- 粗糙集的优势, 提出了 F- 邻域粗糙集和 F- 邻域并行约简. 首先, 定义了 F- 邻域粗糙集上下近似、边界区域; 其次, 在 F- 邻域粗糙集中提出了 F- 属性依赖度和属性重要度矩阵; 根据 F- 属性依赖度和属性重要度矩阵分别提出了属性约简算法, 证明了两种约简方法的约简结果等价; 最后, 比对实验在 UCI 数据集、真实数据集和 MATLAB 生成数据集上完成, 实验结果显示, 与相关算法比较, F- 邻域粗糙集可以获得更好的分类准确率. 为粗糙集在大数据方面的应用增加了一种新方法.

关键词 邻域粗糙集, F- 粗糙集, 属性约简, 属性重要度矩阵

引用格式 邓志轩, 郑忠龙, 邓大勇. F- 邻域粗糙集及其约简. 自动化学报, 2021, **47**(3): 695−705

DOI 10.16383/j.aas.c180556

F-neighborhood Rough Sets and Its Reduction

DENG Zhi-Xuan[1] ZHENG Zhong-Long[1] DENG Da-Yong[1, 2]

Abstract Neighborhood rough sets can directly process numerical data, and F-rough sets are the first dynamic rough set model. For dynamic numerical data, combined the advantages of neighborhood rough sets and F-rough sets, F-neighborhood rough sets and its reducts are proposed. Firstly, three uncertainty regions are defined in F-neighborhood rough sets, including upper and lower approximations, and boundary regions. Secondly, F-dependence degree and an attribute significance matrix are created, and then two attribute reduction algorithms are proposed, which can deal with hybrid data. The obtained reducts with these two algorithms are proved to be equivalent. Finally, the comparison experiments are performed on UCI data sets, real data sets and MATLAB generated data sets. The experimental results show that F-neighborhood rough sets have advantages over related algorithms on the classification accuracy rates. A new method is added for the application of rough sets in big data.

Key words Neighborhood rough sets, F-rough sets, attribute reduction, attribute significance matrix

Citation Deng Zhi-Xuan, Zheng Zhong-Long, Deng Da-Yong. F-neighborhood rough sets and its reduction. *Acta Automatica Sinica*, 2021, **47**(3): 695−705

大数据时代下, 数据快速扩展, 在生产实践中获得的属性越来越多. 一部分属性可能是冗余的或与分类任务无关, 在进行任何进一步数据处理之前都需要将它们删除. 属性约简 (或特征选择) 是一种用于减少属性的技术. 其目的是找到最佳属性子集来预测样本类别. 属性约简还可以促进数据可视化和数据理解[1].

属性约简过程中存在一个关键问题: 属性评估. 如何有效地评估属性是最重要的步骤之一, 它直接影响分类器的性能. 迄今为止已经提出了许多属性评估准则, 例如信息熵[2]、依赖性[1]、相关性[3] 和一致性[4] 等. 通常, 不同的评估标准可能导致不同的最佳特征子集. 但是, 每项措施都旨在确定特征子集的区分能力.

粗糙集理论[5−6] 是一种有效属性约简工具, 产生了增量式约简[7]、动态约简[8]、多决策表约简[9] 和并行约简[10−11] 等属性约简方法. 但是, 传统粗糙集模型仅适用于非数值型数据. 需要对数值型特征进行离散化, 而离散化会带来信息损失.

研究者们通过拓展粗糙集模型来解决这一问题, 如邻域粗糙集[12−21]、模糊粗糙集[22−25] 等. 文献 [13−14] 基于邻域信息粒子逼近, 提出了邻域信息决策模型和数值型属性的选择算法, 能够无须离散化而直接处理数值型属性, 解决了离散化带来的信息损失问题, 使粗糙集模型得以更方便地处理现实生活中大量存在的数值型变量. 后继的研究者引入加权依赖度[15]、局部粗糙集[16]、模糊邻域与模糊决策[17]、Fish swarm 算法[12, 18] 等丰富了邻域粗糙集理论, 并将其应用推广于多标记数据的特征选择[19]、并行属性约简[20]、动态图像分类[21] 等方面. 但并未

收稿日期 2018-08-20 录用日期 2019-01-02
Manuscript received August 20, 2018; accepted January 2, 2019
国家自然科学基金 (61672467) 资助
Supported by National Natural Science Foundation of China (61672467)
本文责任编委 张敏灵
Recommended by Associate Editor ZHANG Min-Ling
1. 浙江师范大学数理与信息工程学院 金华 321004 2. 浙江师范大学行知学院 金华 321004
1. College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua 321004 2. Xingzhi College, Zhejiang Normal University, Jinhua 321004

考虑如何处理包含多个领域数据的数据集, 而不同类型数据的处理准则和要求有所不同, 如果放在同一个信息表中处理, 处理结果往往不尽如人意.

F- 粗糙集[26−29] 是第一个动态粗糙集模型, 其子集可以很好地表示不同情况下的概念, 从而解决处理包含多个领域数据的动态属性约简问题. 其后研究结合了模糊粗糙集[28], 初步应用于非数值型数据. F- 粗糙集比较突出的应用在于概念漂移探测[29], 但相对较缺少非数值型数据处理方面的应用.

为了更好地解决邻域粗糙集和 F- 粗糙集所遇到的问题. 本文结合邻域粗糙集和 F- 粗糙集的优势, 提出了一种新的粗糙集模型 — F- 邻域粗糙集. 首先定义 F- 邻域粗糙集的邻域关系, 使用邻域决策子系统来表示不同情况. 然后, 使用 F- 属性依赖度和属性重要度矩阵来评估属性. 充分考虑了在多种情况下同一概念的不同, 同时克服了邻域粗糙集模型和 F- 粗糙集模型的缺陷. 最后设计了两个属性约简算法, 证明了约简结果的等价性, 并说明了它们的适用范围. 实验结果表明, 相对于邻域粗糙集、F- 粗糙集和主成分分析 (Principal component analysis, PCA), 本文算法能获得更好的分类准确率.

1 基础知识

本节简单介绍邻域粗糙集[13−14] 和 F- 粗糙集[26−27] 的基本概念.

1.1 邻域粗糙集

在信息系统 $IS=(U,A)$ 中, $U=\{x_1, x_2, \cdots, x_n\}$ 为实数空间上的非空有限集合. 如果 A 为条件属性, d 为决策属性, 则称 (U,A,d) 为一个决策系统.

定义 1[13]**.** 对于 $x_k\in U$, 定义 x_k 的邻域为

$$\delta_B(x_k)=\{x_j : x_j\in U, \Delta_B(x_k,x_j)\le\delta\} \tag{1}$$

其中, Δ 代表距离, 通常用 p 范数表示距离为

$$\Delta_p(x_1,x_2)=\left(\sum_{i=1}^{N}|f(x_1,a_i)-f(x_2,a_i)|^p\right)^{\frac{1}{p}} \tag{2}$$

其中, $f(x,a_i)$ 为样本 x 在属性 a_i 上的取值.

$\Delta_p(x_1,x_2)$ 是数值型属性集. 非数值型属性 a_i 可定义为:

1) 当 x_1, x_2 在 a_i 上取值相同时, $|f(x_1,a_i)-f(x_2,a_i)|=0$;

2) 当 x_1, x_2 在 a_i 上取值不相同时, $|f(x_1,a_i)-f(x_2,a_i)|=1$.

定义 2[13]**.** 在信息系统 $IS=(U,A)$ 中, 概念 $X\subseteq U$ 相对于 $B\subseteq A$ 的邻域上近似、下近似、边界区域分别定义为

$$\overline{N}(X)=\overline{N}(IS,B,X)=\{x_k:\delta_B(x_k)\cap X\neq\emptyset,\ x_k\in U\} \tag{3}$$

$$\underline{N}(X)=\underline{N}(IS,B,X)=\{x_k:\delta_B(x_k)\subseteq X,\ x_k\in U\} \tag{4}$$

$$BN(X)=BN(IS,B,X)=\overline{N}(X)-\underline{N}(X) \tag{5}$$

1.2 F- 粗糙集

F- 粗糙集是一个拥有多个信息表 (或决策表) 的粗糙集模型, 它与其他粗糙集模型具有很好的兼容性. 下面介绍 F- 粗糙集的基本概念. 用 $FIS=\{IS_i : IS_i=(U_i,A),\ i=1,2,\cdots,n\}$ 表示信息系统簇, 与其对应的决策系统簇用 F 表示, 其中, $IS_i=(U_i,A)$, 而 $DT_i=(U_i,A,d)$.

定义 3[30]**.** 设 $FIS=\{IS_i : IS_i=(U_i,A),\ i=1,2,\cdots,n\}$ 是一个信息系统簇, $X(IS_i)\subseteq U_i$ 是一个在不同的信息系统 IS_i 中表示可能不同的概念或称为一个在不同的信息系统中意义有所变化的概念变量. 如果不引起混淆, $X(IS_i)$ 可以简记为 X. 则 X 在信息系统簇中的上下近似、边界区域定义为

$$\overline{A}(FIS,X)=\{\overline{A}(IS_i,X):IS_i\in FIS\} \tag{6}$$

$$\underline{A}(FIS,X)=\{\underline{A}(IS_i,X):IS_i\in FIS\} \tag{7}$$

$$BND(FIS,X)=\{\overline{A}(IS_i,X)-\underline{A}(IS_i,X):IS_i\in FIS\} \tag{8}$$

下近似 $\underline{A}(FIS,X)$ 也称为 F- 正区域, 记为 $POS(FIS,X)$ 或 $POS(FIS,A,X)$. $(\overline{A}(FIS,X), \underline{A}(FIS,X))$ 称为 F- 粗糙集.

注 1. $\overline{A}(IS,X)=\{x:[x]_A\cap X\neq\emptyset, x\in U\}$, $\underline{A}(IS,X)=\{x:[x]_A\subseteq X, x\in U\}$, 其中, $[x]_A$ 是等价类. $(\overline{A}(IS,X),\ \underline{A}(IS,X))$ 称为粗糙集.

如图 1 所示, 设 $FIS=\{IS_1,IS_2,IS_3\}$, X 是一个概念, 概念 X 在信息系统 IS_1 中有其上近似、下近似、边界区域, 但在信息系统 IS_2, IS_3 中概念 X 相对于 IS_1 发生了概念漂移. X 在情形 IS_1, IS_2, IS_3 的上近似、下近似、边界区域的集合, 即 FIS 的上下近似、边界区域分别用不同的块划分.

定义 4[30]**.** 设 F 是一个决策系统簇, F 的正区域定义为

$$POS(F,A,d)=\{POS(DT,A,d):DT\in F\} \tag{9}$$

定义 5[30]**.** 设 F 是一个决策系统簇, 称 $B\subseteq A$ 为 F 的并行约简, 当且仅当 $B\subseteq A$ 满足下面两个条件:

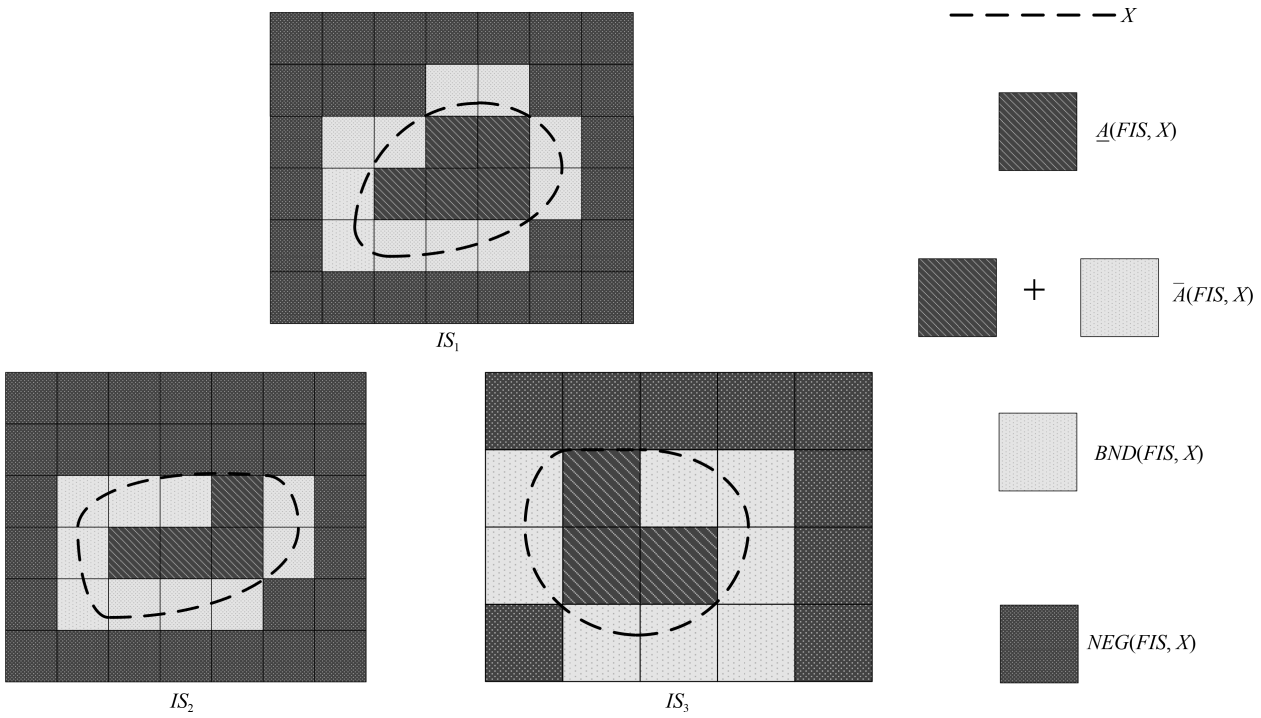

图 1　概念 X 在 FIS 中的上近似、下近似、边界区域、负区域

Fig. 1　Concept X in the FIS upper approximation, lower approximation, boundary region, and negative region

1) $POS(F,B,d)=POS(F,A,d)$;

2) 对任意 $S\subset B$, 都有 $POS(F,S,d)\neq POS(F,A,d)$.

2　F- 邻域粗糙集

本节结合邻域粗糙集和 F- 粗糙集, 定义了 F- 邻域粗糙集的基本概念, 包括 F- 邻域上下近似、边界区域等; 提出 F- 邻域依赖度并证明其单调性, 提出属性重要度矩阵.

注 2. 数据类型为数值型, 或者混合数据类型, 即有些是数值型, 另一些是非数值型.

2.1　F- 邻域粗糙集的基本概念

定义 6. 设 $FIS=\{IS_i: IS_i=(U_i,A),\ i=1,2,\cdots,n\}$ 是一个信息系统簇, $X(IS_i)\subseteq U_i$ 是一个在不同的信息系统 IS_i 中表示可能不同的概念或称为一个在不同的信息系统中意义有所变化的概念变量. $\overline{N}(IS_i,A,X)$ 与 $\underline{N}(IS_i,A,X)$ 分别为概念 X 在信息系统 IS_i 中的邻域上近似与下近似, 对于概念 $X\subseteq U$ 的 F- 邻域上近似、下近似、边界区域的定义为

$$\overline{N}(FIS,A,X)=\{\overline{N}(IS_i,A,X): IS_i\in FIS\} \tag{10}$$

$$\underline{N}(FIS,A,X)=\{\underline{N}(IS_i,A,X): IS_i\in FIS\} \tag{11}$$

$$NBND(FIS,A,X)=\{\overline{N}(IS_i,A,X)-\underline{N}(IS_i,A,X): IS_i\in FIS\} \tag{12}$$

概念 X 关于信息系统簇 FIS 的上近似、下近似、边界区域分别是 FIS 中的子信息系统关于概念 X 的上近似、下近似、边界区域的集合. 其中, 下近似 $\underline{N}(FIS,A,X)$ 也称为 F- 邻域正区域, 记为 $POS(FIS,A,X)$. 序偶 $(\underline{N}(FIS,A,X),\ \overline{N}(FIS,A,X))$ 称为概念 X 在信息系统簇 FIS 中的 F- 邻域粗糙集.

F- 邻域粗糙集继承了 F- 粗糙集的动态性. 不同的信息子系统包含的信息是不一样的, 随着时间或空间的变化而变化, F- 邻域粗糙集与 F- 粗糙集一样包含了这些变化信息, 而且可以研究这些变化.

例 1. 设 $F=\{NDT_1,NDT_2\}$, 邻域决策系统 $NDT_i=(U_i,A,d)$, a, b, c 是条件属性, d 是一个决策属性, $f(x,a)$ 表示样本在属性 a 上的取值, 具体各个样本在属性 a, b, c, d 上的取值如表 1 和表 2 所示.

概念 X 的邻域 $\delta(x)$ 在 NDT_1 和 NDT_2 中是不同的, 当指定邻域大小为 0.5 时 (为了方便计算采用欧氏距离), x 在条件属性 $\{a,b,c\}$ 下的邻域计算式为

$$(|f(x_1,a)-f(x_2,a)|^2+|f(x_1,b)-f(x_2,b)|^2+|f(x_1,c)-f(x_2,c)|^2)^{\frac{1}{2}}\leq 0.5 \tag{13}$$

表 1 邻域决策子系统 NDT_1

Table 1 A neighborhood decision subsystem NDT_1

U_1	$f(x,a)$	$f(x,b)$	$f(x,c)$	$f(x,d)$
x_1	0.1	0.6	0.1	0
x_2	1.5	1.0	0.3	0
x_3	1.6	1.2	0.4	1
x_4	0.3	0.9	0.2	0
x_5	1.3	1.5	0.5	1

表 2 邻域决策子系统 NDT_2

Table 2 A neighborhood decision subsystem NDT_2

U_1	$f(y,a)$	$f(y,b)$	$f(y,c)$	$f(y,d)$
y_1	1.1	2.1	0.6	1
y_2	1.3	1.9	2.2	1
y_3	1.2	0.5	2.4	1
y_4	1.0	0.8	2.1	0
y_5	1.1	0.6	1.6	0

则 $\delta(x_1)=\{x_1,x_4\}$; $\delta(x_2)=\{x_2,x_3\}$; $\delta(x_3)=\{x_2,x_3,x_5\}$; $\delta(x_4)=\{x_1,x_4\}$; $\delta(x_5)=\{x_3,x_5\}$. 令 $X=\{x:d(x)=1,x\in U\}$, 则根据定义 6 可以求得概念 X 在 F 中的上近似、下近似、边界区域为

$$\begin{aligned}&\overline{N}(NDT_1,A,X)=\{x_2,x_3,x_5\}\\&\underline{N}(NDT_1,A,X)=\{x_5\}\\&BN(NDT_1,A,X)=\{\overline{N}(NDT_1,A,X)-\\&\quad\underline{N}(NDT_1,A,X)\}=\{x_2,x_3\}\end{aligned}$$

同理可得

$$\begin{aligned}&\overline{N}(NDT_2,A,X)=\{y_1,y_2,y_3,y_4\}\\&\underline{N}(NDT_2,A,X)=\{y_1,y_2\}\\&BN(NDT_2,A,X)=\{\overline{N}(NDT_2,A,X)-\\&\quad\underline{N}(NDT_2,A,X)\}=\{y_3,y_4\}\end{aligned}$$

则

$$\begin{aligned}&\overline{N}(F,A,X)=\{\{x_2,x_3,x_5\},\{y_1,y_2,y_3,y_4\}\}\\&\underline{N}(F,A,X)=\{\{x_5\},\{y_1,y_2\}\}\\&NBND(F,A,X)=\{\{x_2,x_3\},\{y_3,y_4\}\}\end{aligned}$$

2.2 F- 邻域并行约简

F- 邻域粗糙集的属性约简, 结合了邻域粗糙集处理连续型数据和 F- 粗糙集的动态性的优点, 可以对数值型数据和动态变化的数据进行约简.

定义 7. 设 F 是一个邻域决策系统簇, 称 $B\subseteq A$ 为 F 的邻域并行约简, 当且仅当满足下面两个条件:

1) $POS(F,B,d)=POS(F,A,d)$;

2) 对任意 $S\subset B$, 都有 $POS(F,S,d)\neq POS(F,A,d)$.

定义 8. 在一个邻域决策系统簇 $F=\{NDT_i,\ i=1,2,\cdots,n\}$ 中, 决策属性 d 对条件属性集 B 的依赖度为

$$\gamma(F,B,d)=\frac{\sum\limits_{NDT_i\in F}^{n}Card\left(\underline{N}(NDT_i,B,d)\right)}{\sum\limits_{i=1}^{n}Card(U_i)}\tag{14}$$

定义 9. 在一个邻域决策系统簇 F 中, $NDT_i=(U_i,A,d)\in F$, $i=1,2,\cdots,n$, $B\subseteq A$, 定义属性 $a\in B$ 或 $a\in A-B$ 相对于 B 的 F- 邻域属性重要度为

$$\sigma(B,a)=\gamma(F,B,d)-\gamma(F,B-\{a\},d)\tag{15}$$

或

$$\sigma'(B,a)=\gamma(F,B\cup\{a\},d)-\gamma(F,B,d)\tag{16}$$

定义 10. 在一个邻域决策系统簇 F 中, $NDT_i=(U_i,A,d)\in F$, $i=1,2,\cdots,n$, $B\subseteq A$, F 的所有邻域并行约简组成的集合为 $NRED$, 则 F- 邻域并行约简的核定义为

$$NCORE=\bigcap NRED\tag{17}$$

通过定义 7 将并行约简的概念扩展到邻域决策系统中, 定义 8 和定义 9 是对邻域决策系统属性重要度的扩展, 假设 F 中只含有一个邻域决策系统, 那么, F- 邻域属性重要度即为该决策系统的邻域属性重要度. F- 邻域粗糙集的属性重要度有以下性质:

性质 1. B_1 是邻域决策子系统 $NDT\in F$ 的一个约简, 则存在一个 F- 邻域并行约简 B_2 使得 $B_1\subseteq B_2$.

性质 2. 如果 a 为一个邻域决策子系统 $NDT\in F$ 的核属性, 则 a 为 F- 邻域并行约简的核属性.

性质 3. 如果 a 为 F- 邻域并行约简的核属性, 则存在一个邻域决策子系统 $NDT\in F$, 使得 a 为 NDT 的核属性.

性质 1 ∼ 3 可以根据 F- 邻域并行约简、核属性的定义直接得出.

定理 1. 在一个邻域决策系统簇 F 中, F- 邻域依赖度 $\gamma(F,B,d)$ 具有单调性, 如果 $B_1\subseteq B_2\subseteq\cdots\subseteq A$, 则 $\gamma(F,B_1,d)\leq\gamma(F,B_2,d)\leq\cdots\leq\gamma(F,A,d)$.

证明. 根据文献 [13] 的定理 3, 我们可以得出: $POS(NDT_i, B_1, d) \subseteq POS(NDT_i, B_2, d) \subseteq \cdots \subseteq POS(NDT_i, A, d)$. 而

$$POS(F, B, d) = \{POS(NDT_i, B, d) : NDT_i \in F\}$$

于是有

$$Card(POS(F, B_1, d)) \subseteq Card(POS(F, B_2, d)) \subseteq \cdots \subseteq Card(POS(F, A, d))$$

由于

$$\gamma(F, B, d) = \frac{Card(POS(F, B, d))}{\sum_{i=1}^{n} Card(U_i)}$$

则有

$$\gamma(F, B_1, d) \leq \gamma(F, B_2, d) \leq \cdots \leq \gamma(F, A, d)$$

□

因为 $\gamma(F, B, d)$ 的单调性, 根据文献 [30] 中的定律 1, $\gamma(F, B, d)$ 可以作为属性约简准则, 由此可得到定理 2.

定理 2. 在一个邻域决策系统簇 F 中, $B \subseteq A$ 是 F 的邻域并行约简, 当且仅当 $B \subseteq A$ 满足下面两个条件:

i) $\gamma(F, B, d) = \gamma(F, A, d)$;

ii) 对于任意 $S \subset B$, 有 $\gamma(F, S, d) \neq \gamma(F, A, d)$.

证明. 条件 i) 确保了 F- 邻域正区域保持不变; 条件 ii) 确保了 F- 邻域并行约简的最小性.

1) 证明 $\gamma(F, B, d) = \gamma(F, A, d) \Leftrightarrow POS(F, B, d) = POS(F, A, d)$.

a) $\gamma(F, B, d) = \gamma(F, A, d) \Rightarrow POS(F, B, d) = POS(F, A, d)$ 显然成立;

b) 假设存在 $\gamma(F, B, d) = \gamma(F, A, d)$ 成立, 而 $POS(F, B, d) \neq POS(F, A, d)$. 由于 $\gamma(F, B, d) = Card(POS(F, B, d))/\sum_{i=1}^{n} Card(U_i)$, 在同一信息系统簇 F 中 $\sum_{i=1}^{n} Card(U_i)$ 是一个定值, 因为 $\gamma(F, B, d)$ 的单调性, 当 $\gamma(F, B, d) = \gamma(F, A, d)$ 时, 必然有 $POS(F, B, d) = POS(F, A, d)$, 与假设矛盾.

2) 证明确保 F- 邻域并行约简的最小性.

假设存在 $S \subset B$, 使得 $\gamma(F, S, d) = \gamma(F, A, d)$. 根据 1) 可知: $POS(F, S, d) = POS(F, A, d)$, 即 S 是 F 的 F- 邻域并行约简, 与 $B \subseteq A$ 是 F 的 F- 邻域并行约简矛盾. □

根据以上性质和定理, 还可以得到以下两个命题:

命题 1. 给定一个邻域决策子系统簇 F, $a \in B \subseteq A$. 如果 $\sigma(B, a) = 0$, 则属性 a 可以被约简.

$\sigma(B, a) = 0$ 表明, 如果属性 a 被约简, F 的所有决策子系统也能保持正域不变.

命题 2. 给定一个邻域决策子系统簇 F, $a \in A$, 若 $\sigma(A, a) > 0$, 则属性 a 为 F- 邻域并行约简的核属性.

$\sigma(A, a) > 0$ 表明, 如果属性 a 被约简, 至少有一个邻域决策子系统不能保持正区域不变, 所以属性 a 为 F- 邻域并行约简的核属性.

2.3 属性重要度矩阵

第 2.1 节已经构建了 F- 邻域并行约简的概念. 本小节引入属性重要度矩阵并证明 F- 邻域属性重要度和属性重要度矩阵构建的约简准则等价.

文献 [27] 所提出的属性重要度矩阵是求并行约简的一种方法, 基于此我们构造了 F- 邻域并行约简, 属性重要度矩阵的定义如下:

定义 11. F 是一个邻域决策系统簇, $NDT_i = (U_i, A, d) \in F$, $i = 1, 2, \cdots, n$, $B \subseteq A$, B 关于 F 的属性重要度矩阵定义为

$$H[B, F] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nm} \end{bmatrix} \tag{18}$$

其中, $\sigma_{ij} = \sigma(a_j, U_i) = \gamma_i(U_i, B, d) - \gamma_i(U_i, B - \{a_j\}, d)$, $a_j \in B$, n 表示 F 中邻域决策子系统的个数, m 表示条件属性的个数. 矩阵 $H[B, F]$ 的行表示不同的属性在同一邻域决策子系统下的属性重要度, 列表示相同的属性在不同邻域决策子系统下的属性重要度.

定理 3. 在一个邻域决策系统簇 F 中, $B \subseteq A$ 是 F 的 F- 邻域并行约简, 当且仅当 $B \subseteq A$ 满足下面两个条件:

i) $POS(F, B, d) = POS(F, A, d)$;

ii) 属性重要度矩阵 $H[B, F]$ 中没有全零的列.

证明. 条件 i) 确保了 F- 邻域正区域保持不变; 条件 ii) 确保了 F- 邻域并行约简的最小性.

1) 条件 i) 由定义 9 直接得出;

2) 证明确保 F- 邻域并行约简的最小性.

反设: 属性重要度矩阵 $H[B, F]$ 中有全零的列, 使得 $B \subseteq A$ 是 F 的 F- 邻域并行约简. 由于 $\sigma_{ij} = \gamma_i(U_i, B, d) - \gamma_i(U_i, B - \{a_j\}, d)$, $\sigma_{ij} = 0$ 说明属性 a_j 在 U_i 中对依赖度无影响, 若 a_j 所对应的列元素全为零, 表明 $\sigma_j = \gamma(F, B, d) - \gamma(F, B - \{a_j\}, d) = 0$, 则有 $B - \{a_j\} \subset B$, $\gamma(F, B - \{a_j\}, d) = \gamma(F, B, d)$, 与定理 2 矛盾. □

定理 2 的约简准则等价于定理 3 的约简准则, 定理 2 中第 1 部分的证明已得出定理 2 与定理 3 的

条件 i) 等价; $H[B,F]$ 中没有全零的列, 由定理 3 可知 B 中所有属性对 $\gamma(F,B,d)$ 都有影响, 则有任意 $S \subset B$, $\gamma(F,S,d) \neq \gamma(F,B,d) \Leftrightarrow \gamma(F,S,d) \neq \gamma(F,A,d)$.

因为定理 2 和定理 3 的约简准则等价, 所以可以用 F-邻域属性重要度和邻域属性重要度矩阵来求得 F-邻域并行约简, 两种方法求得的约简结果是相同的, 具体算法可见第 3.2 节.

为了求属性约简, 需要定义 H 的改进矩阵 H', 改进矩阵 H' 定义如下.

定义 12. F 是一个邻域决策系统簇, $NDT_i = (U_i, A, d) \in F$, $i = 1, 2, \cdots, n$, $B \subseteq A$, B 关于 F 的改进属性重要度矩阵定义为

$$H'[B,F] = \begin{bmatrix} \sigma'_{11} & \sigma'_{12} & \cdots & \sigma'_{1m} \\ \sigma'_{21} & \sigma'_{22} & \cdots & \sigma'_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma'_{n1} & \sigma'_{n2} & \cdots & \sigma'_{nm} \end{bmatrix} \tag{19}$$

H' 是 H 的改进矩阵, 若 $a_j \in B$, 则 $\sigma'_{ij} = 0$, 这意味着随着 B 中包含的属性越多, H' 就越稀疏. 如果 B 中的属性随时间变化而增多, 直到 $POS(F, B, d) = POS(F, A, d)$, 也就是 H' 为零矩阵为止, 这就是一个增量式约简过程.

3 约简算法

属性约简是粗糙集理论最重要的应用之一, 而并行约简是属性约简的一个重要延伸. 并行约简是在若干个信息子系统 (或决策子系统) 中寻找稳定的、泛化能力强的条件属性约简. 基于属性重要度, 有以下 F-邻域并行约简算法 (算法 1), 本算法借鉴了文献 [27] 算法的思想, 根据属性集 A 中各元素在邻域决策子表簇 F 中的属性重要度找到属性核, 然后通过属性重要度找到其他属性.

算法 1. 基于 F-邻域属性重要度的 F-邻域并行约简算法 (NPRAS).

输入. $F \subseteq P(DS)$;

输出. F 的一个邻域并行约简 S;

步骤 1. $S = \emptyset$;

步骤 2. $\forall a \in A$, 计算 F-邻域属性重要度 $\sigma(A, a)$, 如果 $\sigma(A,a) > 0$, 那么 $S = S \cup \{a\}$;

步骤 3. $C = A - S$;

步骤 4. 重复以下步骤, 直至 C 为空集:

步骤 4.1. 对任意 $a \in C$, 计算 $\sigma'(S,a)$ //$\sigma'(S, a) = \gamma(F, S \cup \{a\}, d) - \gamma(F, S, d)$;

步骤 4.2. 对任意 $a \in C$, 如果 $\sigma'(S,a) = 0$, 那么 $C = C - \{a\}$ //将属性 a 从属性集 C 中删除;

步骤 4.3. 选择 F-邻域属性重要度最大的非零元素 $a \in C$, $S = S \cup \{a\}$, $C = C - \{a\}$ //将属性集 C 中 F-邻域属性重要度最大的非零元素添加到邻域并行约简 S 中;

步骤 5. 输出邻域并行约简 S.

算法 1 首先从局部的邻域决策子表中计算出决策属性对条件属性的依赖度和条件属性的属性重要度, 得出各个子表的核属性, 然后, 从邻域决策子表簇整体出发, 计算出条件属性的属性重要度, 最后, 得出原属性集的一个 F-邻域并行约简.

算法 1 的时间复杂度主要由 F-邻域属性重要度和步骤 4 的时间复杂度决定. 其中计算一个条件属性的 F-邻域属性重要度的时间复杂度为 $O(m\sum_{U\in F} U \times \sum_{U\in F} \log U)$[26], U 代表决策子表中数据的个数, m 代表条件属性的个数. 在最坏的情况下, 步骤 4 需计算 $\frac{1}{2}m(m-1)$ 次 F-邻域属性重要度. 因此算法 1 的时间复杂度为 $O(m^3 \sum_{U\in F} U \times \sum_{U\in F} \log U)$.

基于属性重要度矩阵, 有以下 F-邻域并行约简算法 (算法 2), 根据属性重要度矩阵 $H(A,F)$ 找到属性核 B, 然后通过建立 B 的改进属性重要度矩阵 H' 找到其他属性, 直到 $H'(P,F)$ 为零矩阵为止.

算法 2. 基于属性重要度矩阵的 F-邻域并行约简算法 (NPRMS)

输入. $F \subseteq P(DS)$;

输出. F 的一个邻域并行约简 P;

步骤 1. 建立属性重要度矩阵 $H'[A,F]$;

步骤 2. $P = \bigcup_{j=1}^{m}\{a_j : \exists \sigma_{kj}(\sigma_{kj} \in H(A,F) \wedge \sigma_{kj} \neq 0)\}$ //P 是所有子系统的属性核;

步骤 3. 计算 $H'[P,F]$;

步骤 4. 重复以下步骤, 直至 $H'[P,F]$ 为零矩阵:

步骤 4.1. For $j = 1$ to m **do** $s_j = 0$; //m 为条件属性个数, s_j 为 $H'[P,F]$ 第 j 列中非零元素个数;

步骤 4.2. For $j = 1$ to m **do**

For $k = 1$ to n **do**

If $t'_{kj} \neq 0$ **then** $s_j = s_j + 1$

//计算 $H'(P,F)$ 中每一列中非零元素个数;

步骤 4.3. $P = P \cup \{a_j : \exists s_j(s_j \neq 0 \wedge \forall s_p(s_j \geq s_p))\}$ //将 $H'[P,F]$ 中非零元素个数最多的列所对应的属性加入 P 中;

步骤 5. 输出邻域并行约简 P.

算法 2 是根据定义 13 和定义 14 构造 F-邻域并行约简算法. 邻域并行约简 P 先从空集开始, 通过计算不同 (相同) 的属性在同一 (不同) 邻域决策子系统下的属性重要度建立属性重要矩阵, 先从中选出所有子系统中属性重要度都不为零 (即矩阵 H

中没有零元素的列) 所对应的属性加入 P 中, 然后计算改进属性重要度矩阵 H' 把非零元素个数最多的列所对应的属性加入 P 中, 直到 H' 为零矩阵. 该算法保证了对正区域有影响的属性不会被删除.

算法 2 的时间复杂度主要是由建立矩阵以及改进矩阵组成, 使用与算法 1 相同的方法计算属性重要度, 它的时间复杂度为 $\mathrm{O}(mU\log U)$, 其中, U 代表决策子表中数据的个数, m 代表条件属性的个数, 那么建立一个属性重要度矩阵的时间复杂度为 $\mathrm{O}(nm^2U'\log U')$, 其中, U' 代表 F 中最大子表的数据个数, n 代表子表个数. 在最坏的情况下, 改进的矩阵的个数为 m, 因此算法 2 的时间复杂度为 $\mathrm{O}(nm^3U'\log U')$, 略高于算法 1.

算法 1 和算法 2 的约简结果是相同的, 所以在大部分情况下使用算法 1 或算法 2 并没有区别. 算法 1 使用的 F- 邻域属性重要度表明属性对邻域决策系统簇整体的影响; 算法 2 使用的属性重要度矩阵表明属性对邻域决策系统簇中各个子系统的影响. 当实验需要测量属性对邻域决策系统簇的影响时应该使用算法 1, 实验需要测量属性对各个决策子系统的影响时应该使用算法 2, 实验对以上两个数据都需要时应该同时使用算法 1 和算法 2.

4　实验结果

本节在 UCI 数据集、真实数据集以及 MATLAB 生成数据集上进行实验, 通过对比实验, 验证了相对于邻域粗糙集、F- 粗糙集和 PCA, F- 邻域粗糙集在分类准确率上都具有优势.

4.1　数据集

本节设计了一些实验来测试所提出的 F- 邻域粗糙模型的性能, 使用 UCI 数据库 (http://archive.ics.uci.edu/ml/datasets.html) 中 Iris 等 9 个数据集; 两个真实数据集 Cevaluation, Rapequality; 一个 MATLAB 生成数据集 Generated data. Cevaluation 数据集来自于国内某高校研二、研三的研究生综合测评数据, 包含 240 个样本和 26 个属性, 其中研二和研三的综合测评计分规则不同. Rapequality 数据集用于描述油菜加工品质, 包含 138 个样本和 26 个属性. 如表 3 所示.

4.2　实验设置

实验的目的是验证 F- 邻域粗糙集约简的有效性, 并通过对比约简后属性子集的分类质量揭示它的可行性. 由于本次实验使用算法 1 与算法 2 的约简结果的等价性, 在实验中使用的算法为算法 2. 实验的评估是使用决策树模型中比较稳定的 CART 分类器, 以 10 折交叉验证计算分类准确率, 以分类准确率的高低为标准. 同时为了构造邻域决策系统簇 F, 对每个数据集进行了分块, 每一块都作为一个邻域决策系统. 实验选取 $\delta = 0.1$, $\delta = 0.05$, $\delta = 0.01$ 三个邻域参数进行验证, 所有算法在 MATLAB R2018a 上实现.

表 3　数据集描述

Table 3　Description of datasets

名称	样本量	属性量	分类数目
Iris	150	4	3
wpbc	198	33	2
soy	47	35	4
sonar	208	60	2
wine	178	13	3
abalone	4 177	8	3
spambase	4 601	57	2
debrecen	1 151	19	2
EEGEye	14 980	14	2
Cevaluation	240	26	2
Rapequality	138	10	2
Generated data	1 000	40	2

4.3　实验结果与分析

先与 NRS 进行比较, 目的是为了比较 NRS 与 NPRMS 的约简质量. 因此, 计算两种算法基于 CART 分类器在 $\delta = 0.1$, $\delta = 0.05$, $\delta = 0.01$ 下的分类准确率, 如表 4～6 所示.

从表 4～6 中可以发现, 在 3 个参数下, 除了 abalone 数据集在邻域 0.1 和 0.05 下没有约简, 其他情况下两种方法都能有效地减少属性. 虽然在除 soy, Iris, wine 和 Cevaluation 之外的数据集上, F- 邻域并行约简 (NPRMS) 的约简属性子集数目要高于 NRS, 但是在 wpbc, sonar, debrecen, EEGEye 和 Generated data 数据上, NPRMS 仅仅多出了一个属性, 在分类准确率上则提升了 5～17.5 个百分点不等. 其在这些数据集上分类准确率的大幅提升可以表明, NPRMS 在这些数据集上的性能要优于 NRS. 值得注意的是在 Rapequality 数据集上, 当参数为 0.1 和 0.5 时, NPRMS 和 NRS 的约简完全相同; 当参数为 0.01 时, NPRMS 的约简子集还是没有变化, NRS 则减少了两个属性, 降低了 3 个百分点的分类准确率. 在 Rapequality 数据集上, NPRMS 有效地保留了决策子系统中的有效信息, 而 NRS 约简掉两个属性时分类准确率出现了明显降低, 说明 NRS 在参数为 0.01 的约简中丢失了有效信息, NPRMS 的表现符合我们为了有效地保留

表 4 $\delta = 0.1$ 时两种算法约简的结果

Table 4 Results of two algorithm reductions when $\delta = 0.1$

数据集	NRS		NPRMS (或 NPRAS)	
	属性数目	分类准确率	属性数目	分类准确率
Iris	4	0.93333	3	**0.93333**
wpbc	6	0.625	7	**0.65**
soy	2	**1**	2	**1**
sonar	5	0.64286	10	**0.69048**
wine	5	0.86111	4	**0.88889**
abalone	8	0.83713	8	**0.83713**
spambase	8	0.88587	9	**0.89239**
debrecen	3	0.60435	4	**0.62609**
EEGEye	4	0.71996	5	**0.8004**
Cevaluation	2	0.89583	4	**0.91667**
Rapequality	4	0.92857	4	**0.92857**
Generated data	4	0.565	5	**0.665**

表 5 $\delta = 0.05$ 时两种算法约简的结果

Table 5 Results of two algorithm reductions when $\delta = 0.05$

数据集	NRS		NPRMS (或 NPRAS)	
	属性数目	分类准确率	属性数目	分类准确率
Iris	3	0.86667	3	**0.93333**
wpbc	4	0.675	6	**0.725**
soy	2	**1**	2	**1**
sonar	4	**0.71429**	7	**0.69048**
wine	3	0.77778	5	**0.83333**
abalone	8	0.83713	8	**0.83713**
spambase	7	0.87065	9	**0.87065**
debrecen	3	0.57391	3	**0.63043**
EEGEye	4	0.71996	5	**0.8004**
Cevaluation	2	0.8125	3	**1**
Rapequality	4	**0.92857**	4	**0.92857**
Generated data	3	0.635	5	**0.67**

有效信息而把 F- 粗糙集引入邻域粗糙集的初衷.

由于 F- 粗糙集并行约简 (OPRMAS)[29] 不能直接处理数值型数据, 先把数据进行离散化处理, 再通过 OPRMAS 算法约简, 根据约简结果从原数据中挑选出这些属性数据, 经过 CART 分类器判别得到最后结果. 我们选取邻域参数 $\delta = 0.01$ 时, NRS 与 NPRMS 的结果与之进行比较. PCA 是经典的特征选择方法, 其对于条件属性的特征选择不需要决策属性, 所以我们在使用 PCA 进行降维时, 去除了数据中的决策属性, 在判别分类准确率时, 再将决策属性加入已降维的数据进行判别; 其还可以控制保留属性的数目, 为了方便比较, 将 PCA 保留属性的数目设置为与 NPRMS 相同.

从表 7 和图 2 中可以看出, 在 NRS 和 NPRMS

表 6 $\delta = 0.01$ 时两种算法约简的结果

Table 6 Results of two algorithm reductions when $\delta = 0.01$

数据集	NRS		NPRMS (或 NPRAS)	
	属性数目	分类准确率	属性数目	分类准确率
Iris	3	0.86667	3	**0.93333**
wpbc	3	0.675	4	**0.85**
soy	2	**1**	2	**1**
sonar	3	0.64286	4	**0.7381**
wine	3	0.86111	3	**0.94444**
abalone	5	0.83832	6	**0.8479**
spambase	8	0.87283	9	**0.87609**
debrecen	2	0.54783	3	**0.6913**
EEGEye	4	0.71996	5	**0.8004**
Cevaluation	2	0.8125	2	**1**
Rapequality	2	0.89286	4	**0.92857**
Generated data	3	0.595	4	**0.64**

表 7 在各个数据集中三种算法约简的结果

Table 7 Results of three algorithmic reductions in each dataset

数据集	NRS		OPRMAS		PCA		NPRMS (或 NPRAS)	
	属性数目	分类准确率	属性数目	分类准确率	属性数目	分类准确率	属性数目	分类准确率
Iris	3	0.86667	3	0.9	3	**0.96667**	3	0.93333
wpbc	3	0.675	9	0.725	4	**0.55**	4	**0.85**
soy	2	**1**	2	0.66667	2	0.77778	2	**1**
sonar	3	0.64286	7	**0.80952**	4	0.61905	4	0.7381
wine	3	0.86111	4	0.77778	3	0.91667	3	**0.94444**
abalone	5	0.83832	8	0.83713	6	0.48862	6	**0.8479**
spambase	8	0.87283	20	0.92283	9	0.87174	9	0.87609
debrecen	2	0.54783	11	0.6087	3	0.56522	3	**0.6913**
EEGEye	4	0.71996	14	**0.83678**	5	0.72664	5	0.8004
Cevaluation	2	0.8125	2	**1**	2	0.8125	2	**1**
Rapequality	2	0.89286	6	0.89286	4	0.89286	4	**0.92857**
Generated data	3	0.595	15	0.575	4	0.57	4	**0.64**

选择合理的邻域参数的情况下, NRS, OPRMAS, PCA, NPRMS 四种方法的属性数目和分类准确率比较. NPRMS 较于 NRS 分类准确率有所提升, 且约简子集数目并未显著增加, 特别是在 Cevaluation 上, 由于其研一、研二两部分测评规则的不同, NRS 表现得并不好, 而 OPRMAS 和 NPRMS 的约简效果明显优于 NRS 和 PCA; 虽然 OPRMAS 在 sonar, spambase 和 EEGEye 上分类准确率要高于 NPRMS, 但其在 sonar 和 spambase 上的约简子集中属性数目多于 NPRMS, 在 EEGEye 上更是并未减少数据集原本的属性数目, 而在 Cevaluation 上两种方法虽然约简结果相同, 但 NPRMS 比 OPRMAS 少了离散化的步骤, 因此具有一定优势.

造成以上实验结果的原因有: 1) NRS 算法由于

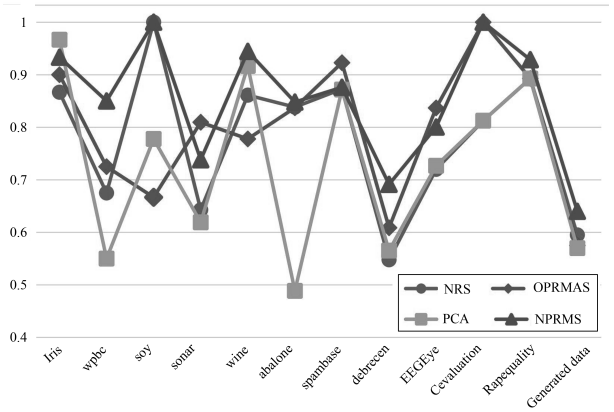

图 2 在各个数据集中算法的分类准确率

Fig. 2 Classification accuracy of algorithms in each dataset

邻域半径造成的信息丢失等原因, 所得的并不是最优约简, 而 NPRMS 是动态约简, 可以有效地减少信息损失; 2) 在样本数量多属性数目少的数据集中, 离散化带来信息损失尤为明显, 使 OPRMAS 在这类数据集上的约简效果较差, 甚至可能并无约简效果, 而 NPRMS 不需要离散化, 保留了必要的信息, 从而可以实现较好的约简; 3) 在有的数据集中包含多种规则, NRS 等大部分算法并没有考虑这种情况, 只是把其当作一般的数据集一样约简, 所以其约简后的分类准确率并不理想; 而 OPRMAS 正是基于这种情况而诞生的算法, 这种数据集下它的约简效果较好是可以预见的; NPRMS 结合了 OPRMAS 这方面的优点, 也能较好地适用于该类数据集的约简.

NPRMS (或 NPRAS) 相较于 NRS 和 OPRMAS, 准确率有所提升, 其性质又决定了其具有更广泛的适用范围, 因此 F- 邻域并行约简更具优势.

5 结论与展望

减少冗余属性可以提高分类性能并降低分类成本. 在本文中, 首先介绍了两种粗糙集模型: F- 粗糙集和邻域粗糙集. 由于两种粗糙集模型都具有自身的优势, 但双方都未考虑对方的优点, 因此提出了 F- 邻域粗糙集. 该模型结合了两个粗糙集模型的优势, 是一个无需离散化处理数值型数据的动态粗糙集模型. 最后, 用 F- 属性重要度和属性重要度矩阵来评估属性, 使用它们来设计属性约简算法, 并说明两种算法的相同点和不同点. 实验结果表明两种算法能获得较高的分类准确率. 实验中还发现决策子系统的划分对所提出的两种属性约简算法的性能的影响较大. 应该根据属性数目和数据项数目为每个数据集选择合适的决策子系统划分.

未来的工作可能包括: 1) 如何将所提出的模型应用于具有不确定性的分类学习和推理领域; 2) 在所提出的模型中, 在数据集中划分决策子系统对所提出算法的性能具有重要影响. 它需要由用户提前划分. 如何为每个数据集自动自动划分决策子系统的最佳解决方案也是一项有意义的工作.

References

1 Jensen R, Shen Q. Fuzzy-rough attribute reduction with application to web categorization. *Fuzzy Sets and Systems*, 2004. **141**(3): 469−485

2 Hu Q H, Yu D R, Xie Z X. Information-preserving hybrid data reduction based on fuzzy-rough techniques. *Pattern Recognition Letters*, 2006, **27**(5): 414−423

3 Hall M A. Correlation-based feature selection for discrete and numeric class machine learning. In: Proceedings of the 17th International Conference on Machine Learning. Morgan Kaufmann Publishers Inc, 2015. 359−366

4 Dash M, Liu H. Consistency-based search in feature selection. *Artificial Intelligence*, 2003, **151**(1−2): 155−176

5 Pawlak Z. Rough sets. *International Journal of Information and Computer Science*, 1982, **11**(5): 341−356

6 Sakai H, Chakraborty M K, Hassanien A E, Sakai H, Chakraborty M K, Hassanien A E, Ślęzak D, Zhu W. Rough sets, fuzzy sets, data mining and granular computing. In: Proceedings of the 12th International Conference on RSFDGrC. Delhi, India: Springer, 2009.

7 Yin Lin-Zi, Yang Chun-Hua, Wang Xiao-Li, Gui Wei-Hua. An incremental algorithm for attribute reduction based on labeled discernibility matrix. *Acta Automatica Sinica*, 2014, **40**(3): 397−404
(尹林子, 阳春华, 王晓丽, 桂卫华. 基于标记可辨识矩阵的增量式属性约简算法. 自动化学报, 2014, **40**(3): 397−404)

8 Wang Jia-Yang, Chen Song-Qiao, Luo An. Study for dynamic reduct based on rough set. *Mini-micro Systems*, 2006, (11): 2056−2060
(王加阳, 陈松乔, 罗安. 粗集动态约简研究. 小型微型计算机系统, 2006, (11): 2056−2060)

9 Xu Fei-Fei, Lei Jing-Sheng, Bi Zhong-Qin, Miao Duo-Qian, Du Hai-Zhou. Approaches to approximate reduction with interval-valued multi-decision tables in big data. *Journal of Software*, 2014, **25**(9): 2119−2135
(徐菲菲, 雷景生, 毕忠勤, 苗夺谦, 杜海舟. 大数据环境下多决策表的区间值全局近似约简. 软件学报, 2014, **25**(9): 2119−2135)

10 Deng Da-Yong, Xu Xiao-Yu, Huang Hou-Kuan. Conceptual drifting detecti for categorical evolving data based on parallel reducts. *Journal of Computer Research and Development*, 2015, **52**(5): 1071−1079
(邓大勇, 徐小玉, 黄厚宽. 基于并行约简的概念漂移探测. 计算机研究与发展, 2015, **52**(5): 1071−1079)

11 Chen Ze-Hua, Ma He. Granular matrix based rapid parallel reduction algorithm for MIMO truth table. *Journal of Electronics and Information Technology*, 2015, **37**(5): 1260−1265
(陈泽华, 马贺. 基于粒矩阵的多输入多输出真值表快速并行约简算法. 电子与信息学报, 2015, **37**(5): 1260−1265)

12 Wang F, Xu J, Li L. A novel rough set reduct algorithm to feature selection based on artificial fish swarm algorith. *Advances in Swarm Intelligence*, ICSI 2014. Lecture Notes in Computer Science, Springer, 2014. 24−33

13 Hu Qing-Hua, Yu Da-Ren, Xie Zong-Xia. Numerical attribute reduction based on neighborhood granulation and rough approximation. *Journal of Software*, 2008, (3): 640−649
(胡清华, 于达仁, 谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简. 软件学报, 2008, (3): 640−649)

14 Hu Qing-Hua, Zhao Hui, Yu Da-Ren. Efficient symbolic and numerical attribute reduction with neighborhood rough sets. *Pattern Recognition and Artificial Intelligence*, 2008, **21**(6): 732−738
(胡清华, 赵辉, 于达仁. 基于邻域粗糙集的符号与数值属性快速约简算法. 模式识别与人工智能, 2008, **21**(6): 732−738)

15 Xu Bo, Zhang Xian-Yong, Feng Shan. Weighted dependence of neighborhood rough sets and its heuristic reduction algorithm. *Pattern Recognition and Artificial Intelligence*, 2018, **31**(3): 256−264
(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. 模式识别与人工智能, 2018, **31**(3): 256−264)

16 Wang Q, Qian Y H, Liang X Y, Guo Q, Liang J Y. Local neighborhood rough set. *Knowledge-Based Systems*, 2018, **151**: 53−64

17 Wang C Z, Shao M W, He Q, Qian Y H, Qi Y L. Feature subset selection based on fuzzy neighborhood rough sets. *Knowledge-Based Systems*, 2016, **111**: 173−179

18 Chen Y M, Zeng Z Q, Lu J W. Neighborhood rough set reduction with fish swarm algorithm. *Soft Computing*, 2016, **21**(23): 1−12

19 Duan Jie, Hu Qing-Hua, Zhang Ling-Jun, Qian Yu-Hua, Li De-Yu. Feature selection for multi-label classification based on neighborhood rough sets. *Journal of Computer Research and Development*, 2015, **52**(1): 56−65
(段洁, 胡清华, 张灵均, 钱宇华, 李德玉. 基于邻域粗糙集的多标记分类特征选择算法. 计算机研究与发展, 2015, **52**(1): 56−65)

20 Chen H M, Li T R, Cai Y, Luo C, Fujita H. Parallel attribute reduction in dominance-based neighborhood rough set. *Information Sciences*, 2016, **373**: 351−368

21 Kumar S U, Inbarani H H. PSO-based feature selection and neighborhood rough set-based classification for BCI multiclass motor imagery task. *Neural Computing and Applications*, 2017, **28**(11): 3239−3258

22 Meng D, Zhang X H, Qin K Y. *Soft Rough Fuzzy Sets and Soft Fuzzy Rough Sets*. Pergamon Press, 2011.

23 Zhao Tao, Xiao Jian. Interval type-2 fuzzy rough sets based on inclusion measures. *Acta Automatica Sinica*, 2013, **39**(10): 1714−1721
(赵涛, 肖建. 基于包含度的区间二型模糊粗糙集. 自动化学报, 2013, **39**(10): 1714−1721)

24 Aggarwal M. Probabilistic variable precision fuzzy rough sets. *IEEE Transactions on Fuzzy Systems*, 2016, **24**(1): 29−39

25 Lin Y J, Li Y W, Wang C X, Chen J K. Attribute reduction for multi-label learning with fuzzy rough set. *Knowledge-Based Systems*, 2018, **152**: 51−61

26 Deng D Y, Yan D X, Chen L. Attribute significance for F- parallel reducts. In: Proceedings of the 2011 IEEE International Conference on Granular Computing. Kaohsiung, China: IEEE, 2012. 156−161

27 Deng Da-Yong, Chen Lin. Parallel reduction and F-rough Sets. *Cloud Model and Granular Computing*. Beijing: Science Press, 2012. 210−228
(邓大勇, 陈林. 并行约简与 F- 粗糙集. 云模型与粒计算. 北京: 科学出版社, 2012: 210-228)

28 Deng Da-Yong, Xu Xiao-Yu, Pei Ming-Hua. F-fuzzy rough sets and its reducts. *Journal of Zhejiang Normal University (Natural Science)*, 2015, **38**(1): 58−66
(邓大勇, 徐小玉, 裴明华. F- 模糊粗糙集及其约简. 浙江师范大学学报 (自然科学版), 2015, **38**(1): 58−66)

29 Deng Da-Yong, Li Ya-Nan, Huang Hou-Kuan. Concept drift and attribute reduction from the viewpoint of F-rough sets. *Acta Automatica Sinica*, 2018, **44**(10): 1781−1789
(邓大勇, 李亚楠, 黄厚宽. F- 粗糙集视角的概念漂移与属性约简. 自动化学报, 2018, **44**(10): 1781−1789)

30 Deng Da-Yong, Lu Ke-Wen, Huang Hou-Kuan, Deng Zhi-Xuan. Attribute reduction for concepts and concept drifting detection in heterogeneous data. *Acta Electronica Sinica*, 2018, **46**(5): 1234−1239
(邓大勇, 卢克文, 黄厚宽, 邓志轩. 概念的属性约简及异构数据概念漂移探测, 电子学报, 2015, 2018, **46**(5): 1234−1239)

邓志轩 浙江师范大学硕士研究生. 2016 年获得河南师范大学电气工程及其自动化专业学士学位. 主要研究方向为粗糙集与图像识别技术.
E-mail: zhixuandenga@163.com
(**DENG Zhi-Xuan** Master student at Zhejiang Normal University. He received his bachelor degree from Henan Normal University in 2016. His research interest covers rough sets and image recognition.)

郑忠龙 浙江师范大学数学与计算机学院教授. 2005 年获得上海交通大学模式识别与智能系统专业博士学位. 主要研究方向为机器学习与模式识别. 本文通信作者.
E-mail: zhonglong@zjnu.edu.cn
(**ZHENG Zhong-Long** Professor at the College of Mathematics and Computer Science, Zhejiang Normal University. He received his Ph. D. degree from Shanghai Jiao Tong University in 2005. His research interest covers machine learning and pattern recognition. Corresponding author of this paper.)

邓大勇 浙江师范大学行知学院副教授. 2007 年获得北京交通大学计算机应用技术专业博士学位. 主要研究方向为粗糙集理论及应用.
E-mail: dayongd@163.com
(**DENG Da-Yong** Associate professor at Xingzhi College, Zhejiang Normal University. He received his Ph. D. degree from Beijing Jiaotong University in 2007. His research interest covers rough set theory and its application.)