

基于粒球邻域粗糙集的三支属性约简

辛现伟^{a,b}, 丁肖迪^a, 庞文莉^c, 马媛媛^a, 薛占熬^a

(河南师范大学 a.计算机与信息工程学院(人工智能学院); b.河南省教育人工智能与
个性化学习重点实验室; c.软件学院, 河南 新乡 453007)

摘要:本文提出了一种基于粒球邻域粗糙集(granular-ball neighborhood rough Set, GBNRS)的三支属性约简算法.首先,构建一种基于粒球样本分布的纯度度量准则,通过同时考虑粒球内部样本分布和粒球间的分布,对传统度量进行改进,以更准确地评估属性集合对决策分类的贡献.在此基础上,引入三支决策机制,将属性划分为核心、候选与冗余3类,实现分层筛选与渐进式约简,从而降低误判风险并提升约简质量.最后,通过在12个UCI公共数据集上与多种方法进行对比实验,结果表明所提算法能够在减少约简长度的同时保持较好的分类性能.

关键词:粒球计算;粒球邻域粗糙集;属性约简;三支决策

中图分类号:TP391

文献标志码:A

文章编号:1000-2367(2026)05-0098-07

粗糙集理论作为一种无需外部信息的推理工具,自Pawlak^[1]提出以来,已成为数据分析领域的重要理论框架.在属性约简^[2-4]中,粗糙集理论通过去除冗余和不相关属性来保留数据的分类能力.传统的属性约简方法主要依赖于等价关系来划分样本,但这种方法仅适用于离散数据,无法有效处理实值属性.为此,Hu等^[5]提出了NRS,并设计了基于此模型的前向贪心约简算法,取得了显著的成果.然而,NRS及其扩展模型如WNRS^[6]等仍依赖邻域半径的设定,半径选择不当可能影响邻域划分质量和约简效果.为缓解这一问题,粒球计算^[7-8](granular-ball computing, GBC)被引入NRS模型.GBC以粒球作为基本信息粒,能够根据数据分布自适应划分信息空间,提高算法速度^[9],并已应用于聚类^[10-11]、分类^[12]和特征选择^[13-14]等领域.基于GBC,Xia等^[15]提出了GBNRS模型,克服NRS中邻域半径选择问题的同时提高了算法运行效率.

随后,相关研究围绕GBNRS展开了多方面扩展.Ji等^[16]通过引入模糊粒球的概念,研究了基于模糊粒球的扩展粗糙集模型,增加了GBNRS在进行属性约简时的灵活性.Li等^[17]针对动态信息系统结合属性分组策略,设计了基于GBNRS的增量属性约简算法.Zhang等^[18]提出了一种快速可变粒球生成模型,该模型在处理高维数据时表现出显著的效率优势.同时,Xia等^[19]将GBNRS中的邻域拓展,衍生出了粒球粗糙集(Granular-ball rough set,GBRS),其不仅可以表示Pawlak粗糙集,还能表示邻域粗糙集,从而实现了两者的统一表示.Zhang等^[20]将GBNRS与Wu-Leung模型相结合,设计了一种多尺度粒球邻域决策表用于特征选择.

在GBNRS属性约简的过程中,纯度常用于衡量粒球内部标签一致性,并据此评价属性集合的分类质量.然而,纯度度量仅关注标签数量比例,难以刻画粒球内部不同类别样本的空间分布,也无法反映粒球整体在特征空间中的分布结构.因此,当不同属性集合生成的粒球具有相同纯度时,传统纯度度量可能无法进一步

收稿日期:2025-06-26;**修回日期:**2025-07-31.

基金项目:国家自然科学基金(62506121);河南省科技攻关项目(262102211015).

作者简介:辛现伟(1991—),男,河南舞阳人,河南师范大学讲师,博士,主要从事粒计算和认知计算研究,E-mail: xinxianwei@htu.edu.cn.

通信作者:庞文莉,E-mail: pangwenli@htu.edu.cn.

引用本文:辛现伟,丁肖迪,庞文莉,等.基于粒球邻域粗糙集的三支属性约简[J].河南师范大学学报(自然科学版),2026,54(5):98-104.(Xin Xianwei,Ding Xiaodi,Pang Wenli,et al.Three-way attribute reduction based on granular-ball neighborhood rough set[J].Journal of Henan Normal University(Natural Science Edition),2026,54(5):98-104. DOI:10.16366/j.cnki.1000-2367.2025.06.26.0004.)

区分其真实分类能力,进而造成重要属性误删或冗余属性保留.因此,仅依赖纯度进行属性约简仍存在一定局限.

针对上述问题,本文在传统纯度度量基础上区分一致粒球和不一致粒球,并分别引入粒球间与粒球内样本分布特征,以更全面地评估属性集合的分类能力.同时,引入三支决策理论^[21],将属性划分为重要、候选和无关3类^[22-24],通过多层次细化的属性筛选机制降低属性误判风险,从而提高约简质量.

1 预备知识

1.1 粒球计算

定义 1^[12] 令四元组 $DS = \langle U, AT \cup D, V, f \rangle$ 表示一个决策系统,其中 $U = \{x_1, x_2, \dots, x_n\}$ 是一组对象的集合,称为论域; $AT = \{a_1, a_2, \dots, a_n\}$ 表示条件属性的集合, D 是决策属性; V 是一组属性值的集合, $f: U \times AT \rightarrow V$ 表示映射函数.形式上,邻域决策系统可以简单地表示为一个二元组 $NS = \langle U, AT \cup D \rangle$,其中 $D = \{d\}, d \notin AT, AT \cap D = \emptyset$.

定义 2^[15] 令 U 是实空间的非空有限论域,粒球 GB 的中心 c 和半径 r 计算如下: $c = \frac{1}{N} \sum_{i=1}^N x_i, r = \frac{1}{N} \sum_{i=1}^N \|x_i - c\|$.其中, x_i 表示粒球 GB 中的对象, c 为粒球 GB 中所有对象的重心, r 是粒球 GB 中所有对象到 c 的平均距离, $\|x_i - c\|$ 表示 x_i 到 c 的距离.

定义 3^[19] 令 U 是实空间的非空有限论域, $GB = \{GB_1, GB_2, \dots, GB_k\}$ 是生成的粒球空间, $GB_1 \cup GB_2 \cup \dots \cup GB_k = U$.粒球空间中的第 j 个粒球 GB_j 可以表示为: $GB_j = \{x \mid x \in U, \Delta(x, c_j), r_j\}$.其中, $\Delta(x, c_j)$ 表示 U 中的对象 x 与中心 c_j 的距离度量, c_j 和 r_j 分别为 GB_j 的中心和半径.根据上式, r_i 越大,粒球 GB_j 越粗;反之,粒球 GB_j 越细.

定义 4^[19] 在邻域决策系统 $NS = \langle U, AT \cup D \rangle$ 中,第 j 个粒球 GB_j 的标签为该粒球中占比最大的对象所对应的决策值,记为 $d(GB_j)$.进一步,粒球的纯度为与粒球标签一致的对象占粒球中所有对象的比例,如下: $P_y(GB_j) = \frac{|\{x_i \in GB_j : d(x_i) = d(GB_j)\}|}{|GB_j|}$.其中, $|GB_j|$ 为粒球 GB_j 中对象的个数, $d(x_i) = d(GB_j)$ 表示对象 x_i 的标签与粒球标签一致.为了方便表示,使用 ρ_j 代表第 j 个粒球的纯度.

1.2 粒球邻域粗糙集

定义 5^[19] 在邻域决策系统 $NS = \langle U, AT \cup D \rangle$ 中,对任意的 $B \subseteq AT$ 和 $X \subseteq U$, X 相对于 R 的粒球上、下近似如下: $\overline{GBR}_B X = \{x_i \in U, x_i \in GB_j(B) \mid GB_j(B) \cap X \neq \emptyset\}, \underline{GBR}_B X = \{x_i \in U, x_i \in GB_j(B) \mid GB_j(B) \subseteq X\}$.其中, $GB_j(B)$ 表示在属性子集 B 下在 NS 上生成的第 j 个粒球.

定义 6^[19] 在邻域决策系统 $NS = \langle U, AT \cup D \rangle$ 中,决策 D 将论域 U 划分为 L 个等价类,表示为 $U/D = \{X_1, X_2, \dots, X_L\}$,对任意的 $B \subseteq AT$,决策属性 D 相对于 B 的粒球下、上近似分别如下:

$$\overline{GBR}_B D = \bigcup_{i=1}^L \overline{GBR}_B X_i, \underline{GBR}_B D = \bigcup_{i=1}^L \underline{GBR}_B X_i.$$

定义 7^[25] 在邻域决策系统 $NS = \langle U, AT \cup D \rangle$ 中,对任意的 $B \subseteq AT, a \in B$,若 B 同时满足以下两个条件:(1) $CRT_B(D) = CRT_{AT}(D)$, (2) $\forall a \in B, CRT_{B-\{a\}}(D) \neq CRT_{AT}(D)$.则 B 被称为 AT 的一个约简.其中, CRT 是用于属性约简的标准度量.标准度量可以是任何能反映属性相关性或重要性的测度,如常用的正域、可辨矩阵、互信息等.在 GBNRS 中,Xia 等^[15]采用粒球中独有的纯度作为标准度量,从保持粒球纯度的角度给出了属性约简的概念.

定义 8^[15] 在邻域决策系统 $NS = \langle U, AT \cup D \rangle$ 中, U 是实空间的非空有限论域,对任意的 $B \subseteq AT$,当 B 同时满足下列条件时称 B 为 AT 的一个约简:(1) $P_{y,B}^U = P_{y,AT}^U$, (2) $P_{y,B}^U \neq P_{y,B-\{a\}}^U, \forall a \in B$.其中, P_y 表示根据 B 计算出的所有粒球的平均纯度.

1.3 三支决策

为了提高决策规则对噪声的容忍度和可解释性,Yao^[21]融合粗糙集和贝叶斯最小风险决策理论,提出了

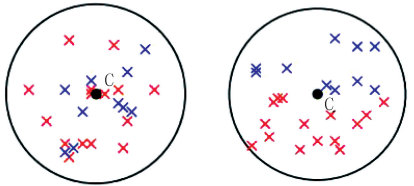
三支决策理论,即将论域划分为 3 个互不相交的区域,即正域 POS ,负域 NEG 和边界域 BND ,并分别对其采取相应的策略.相应地,对于给定的阈值参数 (α,β) ,三支决策的决策规则如下:(P) 若 $P(X | [x]) \geq \alpha$,则 $x \in POS(X)$;(B) 若 $\beta < P(X | [x]) < \alpha$,则 $x \in BND(X)$;(N) 若 $P(X | [x]) \leq \beta$,则 $x \in NEG(X)$.

2 基于粒球样本分布的三支属性约简

2.1 基于粒球样本分布的度量准则

GBNRS 中的粒球通常基于 K -means 聚类,生成纯度为 1(内部样本一致)和纯度小于 1(内部样本不一致)的粒球.不失一般性地,决策类的中心区域由属于同一决策类的样本构成(纯度为 1 的粒球),而边界区域则混杂了不同决策类的样本(纯度小于 1 的粒球).下面将分别讨论这两类粒球,以更全面地评估属性集合的分类能力.

如图 1 所示,不同属性集合下生成的两个不一致粒球虽然具有相同的样本类别、数量和纯度,但其内部样本分布不同.相比图 1(a),图 1(b)中异类样本分布更集中、类别边界更清晰,说明其分类判别能力更强.传统纯度无法刻画这一差异,虽然细化粒球可减少这种影响,但过度细分不仅耗时而且会导致分类能力差异大的属性集合纯度趋同,干扰属性约简结果.因此,本文针对边界区域粒球引入异类样本分布离散程度,定义如下:



(a) 属性集合A下的粒球 (b) 属性集合B下的粒球

图1 边界区域两个样本相同的粒球
Fig.1 Two granular balls with identical samples in the boundary region

$$BP_y = \frac{1}{m} \sum_{i=1}^m (\rho_i / D_i).$$

其中, D_i 为所有与粒球标签不同的样本之间的平均距离,用于衡量粒球内部与标签不同的样本的分布离散程度.特别的,当与粒球标签不同的样本的个数为 1 时, D_i 为 1.易知,在相同纯度下,与粒球标签不同的样本分布越集中,该粒球内部不同标签样本越易区分, BP_y 越大,该粒球的分类能力越强;反之,与粒球标签不同的样本分布越分散,该粒球内部不同标签样本越难区分, BP_y 越小,该粒球的分类能力越弱.

在评估属性集合分类能力时,除边界区域的不一致粒球外,类中心区域中同标签一致粒球之间的空间紧密程度也会影响分类效果,具体如附录图 S1 所示.一般而言,在相同纯度条件下,同类一致粒球分布越紧密,说明该属性集合对类内结构的刻画能力越强;反之,若同类一致粒球分布较分散,则可能削弱属性集合的分类表达能力.

由此,基于类中心区域粒球间样本分布的纯度计算方法如下: $CP_y = \frac{1}{n} \sum_{j=1}^n (1/S_j)$.其中, S_j 为第 j 个标签下所有纯度为 1 的粒球中心之间的平均距离,用于刻画该类别中一致粒球的空间分布的离散程度.特别的,当标签下只有一个粒球时, S_j 为 1.

综上,基于粒球样本分布的纯度公式如下: $DP_y = \frac{1}{n} \sum_{j=1}^n (1/S_j) + \frac{1}{m} \sum_{i=1}^m (\rho_i / D_i)$.

定义 9 在邻域决策系统 $NS = \langle U, AT \cup D \rangle$ 中,对任意的 $B \subseteq AT, a \in B$,当 B 同时满足下列条件时,称 B 为 AT 的一个约简:(1) $DP_{yB}^U = DP_{yAT}^U$, (2) $DP_{yB}^U \neq DP_{yB-\{a\}}^U$.其中, DP_{yB}^U 表示属性集合 B 生成的所有粒球的平均纯度.

定义 10 在邻域决策系统 $NS = \langle U, AT \cup D \rangle$ 中,对任意的 $B \subseteq AT, a \in B$,属性 a 关于 B 的内部属性重要度为: $Sig^{inner}(a, B, D) = DP_{yB}^U - DP_{yB-\{a\}}^U$.

类似的,对任意的 $B \subseteq AT, a \in (AT - B)$,属性 a 关于 B 的外部属性重要度如下:

$$Sig^{outer}(a, B, D) = DP_{yB \cup \{a\}}^U - DP_{yB}^U.$$

$Sig^{inner}(a, B, D)$ 衡量属性 a 在当前属性子集内对决策信息的贡献, $Sig^{inner}(a, B, D) \leq 0$ 表示 a 冗余. $Sig^{outer}(a, B, D)$ 则反映了属性 a 在加入属性子集后对决策信息的贡献, $Sig^{outer}(a, B, D) > 0$ 表示 a 是不可缺少的.

2.2 基于 GBNRS 的三支约简算法

考虑到与决策属性的相关性,可将属性集划分为强相关、弱相关和不相关 3 类.而约简应该包含所有强相关属性,排除所有不相关属性,暂时保留弱相关属性并对其进行进一步考察.本节给出了一种基于 DPurity(后文简称 DP)的三支约简方法.

定义 11 在邻域决策系统 $NS = \langle U, AT \cup D \rangle$ 中通过计算 AT 中每个属性及属性集合 AT 的 DP 并根据阈值 (α, β) 将所有条件属性划分为重要属性集 Ahd (Attributes with high-DP)、候选属性集 Amd (Attributes with medium-DP) 和无关属性集 Ald (Attributes with low-DP):

$$\begin{aligned} Ahd_{(\alpha, \beta)}(a) &= \{a \in AT \mid DP(a) \geq \alpha DP(AT)\}, \\ Amd_{(\alpha, \beta)}(a) &= \{a \in AT \mid \beta DP(AT) < DP(a) < \alpha DP(AT)\}, \\ Ald_{(\alpha, \beta)}(a) &= \{a \in AT \mid DP(a) \leq \beta DP(AT)\}. \end{aligned}$$

其中, $DP(a)$ 表示在属性 a 下基于样本分布的纯度, $DP(AT)$ 表示在全体属性集合下基于粒球样本分布的纯度, (α, β) 是 $[0, 1]$ 上的阈值.

实践中,需要优先选择分类能力强的重要属性,构建稳固的子集基础,避免特定度量下早期选择的重要属性对后续选择的干扰,并排除冗余或无关属性.通过引入三支决策,可使 GBNRS 有效缩小候选属性范围,提高约简的质量.基于 DP 的三支属性约简的伪代码如算法 1 所示.

算法 1 基于 DP 的三支属性约简算法

找到属性 $b = \arg \max \{Sig^{outer}(a, red, D)\}$

输入 邻域决策系统 $NS = \langle U, AT \cup D \rangle, (\alpha, \beta)$

输出 red

步骤 1 由粒球计算生成 AT 及 AT 中各属性的粒球空间

步骤 2 计算 AT 及 AT 中各属性的 DP

步骤 3 根据阈值 (α, β) 对各属性进行划分得到 Ahd 、 Amd 和 Ald

步骤 4 令 $red = Ahd$

步骤 5 While $DP(red) \leq DP(AT)$ and $Amd \neq \emptyset$

任意 $a \in Amd$, 求得 $Sig^{outer}(a, red, D)$

If $DP(red \cup \{b\}) > DP(red)$ then

$red = red \cup \{b\}, Amd = Amd - \{b\}$

end While

步骤 6 For each $a \in red$

If $Sig^{inner}(a, red, D) \leq 0$ then

$red = red - \{a\}$

步骤 7 输出 red

设样本数为 $|U|$, 条件属性数为 $|AT|$, 候选属性数为 $|Amd|$. 第 1 阶段需分别生成各单属性粒球空间并计算 DP, 时间复杂度约为 $O(|AT| \times |U|)$; 第 2 阶段在最坏情况下需遍历候选属性并重复计算属性贡献, 复杂度约为 $O(|U| \times |AT|^2)$; 第 3 阶段删除冗余属性的复杂度约为 $O(|U| \times |AT|)$; 因此, 算法总体时间复杂度为 $O(|U| \times |AT|^2)$.

3 实验分析

本节通过与其他多种方法的对比实验, 验证本文所提算法的有效性. 所有代码均在 PyCharm 中使用 Python 3.9 实现, 实验在操作系统为 Windows 10、Intel(R) Core(TM) i5-10300H CPU @ 2.50 GHz 处理器、16 GB 内存的计算机上进行.

3.1 实验数据及设置

实验使用了 12 个 UCI 数据集, 具体如表 1 所示, 以 KNN ($K=5$) 作为基分类器进行 10 折交叉验证, 并使用 Min-Max 对数据进行归一化处理.

3.2 算法性能

为了检验所提属性约简算法的性能, 将本文算法 DGBNRS 与 NRS^[5]、GBNRS^[15]、MGBNRS^[15] 和 GBRS^[19] 等算法进行了对比分析. 同时, DGBNRS 采用与 MGBNRS 相同的粒球生成方式, 该粒球生成方式能够充分利用标签信息, 更好地拟合数据分布. 此外, 为提升算法在不同数据集上的适应性, 本文所采用的阈值参数 (α, β) 并非固定设定, 而是依据数据集特征动态调整.

表 2 给出了各算法约简后的属性数量. 由表 2 可知, DGBNRS 在多数数据集上取得了较短的约简长度,

表现出较好的属性筛选能力.相比之下,GBNRS 和 MGBNRS 由于采用后向删除策略,属性删除较为保守,约简长度整体偏长;而 GBRS 和本文算法采用前向增加策略,更侧重于筛选对分类有贡献的属性,因此能够在保持分类能力的同时获得更精简的属性子集.

表 1 数据集信息
Tab. 1 The information of datasets

编号	数据集	样本	属性	分类	编号	数据集	样本	属性	分类
1	credit	690	15	2	7	iono	351	34	2
2	sonar	208	60	2	8	wdbc	569	30	2
3	german	1 000	19	2	9	wine	178	13	3
4	horse	368	23	2	10	htru2	17 898	8	2
5	heart2	303	13	5	11	abalone	4 177	8	29
6	hepatitis	155	19	2	12	letter	20 000	16	26

表 2 约简长度
Tab. 2 The reduction length

No.	Original	NRS ^[5]	GBNRS ^[15]	MGBNRS ^[15]	GBRS ^[19]	DGBNRS	No.	Original	NRS ^[5]	GBNRS ^[15]	MGBNRS ^[15]	GBRS ^[19]	DGBNRS
1	15	11	15	13	5	6	7	34	14	8	8	4	3
2	60	24	57	47	13	11	8	30	24	29	28	3	21
3	19	11	16	14	13	5	9	13	10	10	11	6	9
4	23	15	20	20	4	4	10	8	8	5	5	4	4
5	13	9	11	11	7	4	11	8	6	7	7	7	5
6	19	13	13	17	6	8	12	16	16	15	15	11	14

表 3 给出了各算法在 12 个数据集上的分类准确率.由表 3 可知,DGBNRS 在多数数据集上取得最优结果,表明所提方法在约简属性的同时能够保持较好的分类性能.对于 sonar、horse 和 letter 数据集,DGBNRS 虽未达到最优,但与最优方法的差距较小,仍表现出较强的竞争力.综合来看,本文所提出的算法在实现属性约简数目有效降低的同时,分类准确率也获得了提升.

表 3 分类准确率
Tab. 3 The classification accuracy

No.	Original	NRS ^[5]	GBNRS ^[15]	MGBNRS ^[15]	GBRS ^[19]	DGBNRS
1	0.871±0.042	0.865±0.042	0.867±0.028	0.873±0.033	0.872±0.037	0.887±0.035
2	0.816±0.065	0.841±0.046	0.841±0.072	0.841±0.069	0.893±0.039	0.879±0.043
3	0.709±0.036	0.737±0.048	0.734±0.023	0.731±0.037	0.747±0.043	0.755±0.040
4	0.771±0.081	0.796±0.082	0.825±0.051	0.834±0.042	0.858±0.042	0.840±0.044
5	0.551±0.069	0.598±0.049	0.591±0.058	0.613±0.050	0.606±0.058	0.624±0.053
6	0.800±0.112	0.846±0.063	0.851±0.077	0.876±0.090	0.882±0.068	0.890±0.077
7	0.848±0.069	0.871±0.044	0.883±0.030	0.883±0.030	0.903±0.041	0.909±0.043
8	0.967±0.025	0.972±0.023	0.972±0.019	0.974±0.020	0.972±0.022	0.977±0.016
9	0.943±0.060	0.955±0.036	0.966±0.029	0.978±0.126	0.972±0.040	0.989±0.024
10	0.977±0.004	0.914±0.046	0.978±0.004	0.978±0.004	0.978±0.004	0.979±0.004
11	0.525±0.050	0.529±0.034	0.527±0.057	0.527±0.057	0.529±0.041	0.532±0.029
12	0.979±0.004	0.979±0.004	0.979±0.004	0.979±0.004	0.983±0.004	0.979±0.004

3.3 参数分析

阈值参数(α,β)的选择对结果性能有一定的影响.以 german 数据集为例,在固定 $\beta=0.35$ 的条件下,随着 α 从 0.60 增加到 0.85,约简长度和分类准确率呈现一定波动.当 $\alpha=0.75$ 时,算法取得较好的精度与约简

长度平衡,说明阈值参数会影响初始属性划分,并进一步影响后续属性选择结果.附录图 S2 给出了本文算法在所选数据集上分类准确率随 (α, β) 变化的变化情况.其中, α 的取值范围为 $0.60 \sim 0.90$, β 的取值范围为 $0.20 \sim 0.40$,步长均为 0.05 .由于 α 与 β 分别决定了约简的初始属性集合和去除的不相关属性集合,当改变其值时,分类准确率也会随之变化.

3.4 消融实验

为了验证所提出模型中各模块的有效性和贡献,我们设计了模块增量消融实验,如附录表 S1 所示.具体设置如下:消融实验 1(在原始模型上仅增加 DP 模块).消融实验 2(在原始模型上仅增加三支约简模块).并将上述两个实验在 12 个数据集上与原始模型 MGBNRS 和完整模型 DGBNRS 作对比,实验结果如表 S1 所示.

由表 S1 可知,虽然在个别数据集上消融实验 1 和消融实验 2 相比 MGBNRS 表现略微起伏,但在多数数据集中,消融实验 1 和消融实验 2 相比 MGBNRS 表现有所提升.同时也发现,尽管模块单独引入时在部分数据集上略有提升,但整体效果不如完整模型,表明两个模块之间存在协同增强效应.

3.5 统计显著性分析

为验证所提算法在统计学意义上与其他选择的算法的性能差异,将本文算法 DGBNRS 和 5 种对比算法在所选数据集上进行了 Friedman 检验和 Nemenyi 检验,经检验,所选用的几种算法具有一定的显著性差异,其结果如附录图 S3 所示.

从附录图 S3 中可以观察到,本文所提算法与 RAW (Original)、NRS、GBNRS 的平均秩差大于 $CD=2.50$,这意味着所提算法的性能明显优于对比算法.与 MGBNRS 和 GBRS 相比虽未达到显著差异,但平均性能始终优于 MGBNRS 和 GBRS.

4 结论

本研究针对 GBNRS 属性约简时粒球纯度对属性集合分类能力判别的不足,通过结合粒球内的样本分布和粒球间的样本分布对粒球纯度进行改进,提出了一种新的度量准则.此外,在进行属性约简过程中引入三支决策,对属性三分而治,设计了基于所提度量的三支约简算法.该算法在约简过程中同时使用添加一删除策略,旨在以较少的属性长度获得质量更高的约简.在 12 个 UCI 数据上的实验结果表明,本文提出的算法能够以较少的属性个数获得质量更高的约简.本文所提算法在展现粒球样本的几何分布时会受到离群样本的干扰,造成一定误差,可能会对属性集合分类能力的判断产生影响.在以后的工作中,将使用更具鲁棒性的度量,如密度、概率等以减小离群样本误差的影响.

附录见电子版(DOI:10.16366/j.cnki.1000-2367.2025.06.26.0004).

参 考 文 献

[1] Pawlak Z.Rough sets[J].International Journal of Computer & Information Sciences,1982,11(5):341-356.

[2] Sun L,Xu F,Ding W P,et al.AFIFC: Adaptive fuzzy neighborhood mutual information-based feature selection via label correlation[J].Pattern Recognition,2025,164:111577.

[3] Ali G,Afzal M,Asif M,et al.Attribute reduction approaches under interval-valued q-rung orthopair fuzzy soft framework[J].Applied Intelligence,2022,52(8):8975-9000.

[4] Sewwandi M A N D,Li Y F,Zhang J L.A class-specific feature selection and classification approach using neighborhood rough set and K-nearest neighbor theories[J].Applied Soft Computing,2023,143:110366.

[5] Hu Q H,Yu D R,Liu J F,et al.Neighborhood rough set based heterogeneous feature subset selection[J].Information Sciences,2008,178(18):3577-3594.

[6] Hu M,Tsang E C C,Guo Y T,et al.A novel approach to attribute reduction based on weighted neighborhood rough sets[J].Knowledge-Based Systems,2021,220:106908.

[7] Xie Q,Zhang Q H,Xia S Y,et al.GBG: a fast and stable granular ball generation method for classification[J].IEEE Transactions on Emerging Topics in Computational Intelligence,2024,8(2):2022-2036.

[8] Cheng D D,Liu S S,Xia S Y,et al.Granular-ball computing-based manifold clustering algorithms for ultra-scalable data[J].Expert Sys-

tems with Applications,2024,247:123313.

[9] Cheng D D,Li Y,Xia S Y,et al.A fast granular-ball-based density peaks clustering algorithm for large-scale data[J].IEEE Transactions on Neural Networks and Learning Systems,2024,35(12):17202-17215.

[10] Xie J,Kong W Y,Xia S Y,et al.An efficient spectral clustering algorithm based on granular-ball[J].IEEE Transactions on Knowledge and Data Engineering,2023,35(9):9743-9753.

[11] Sun L,Zhang Q F,Ding W P,et al.FCPFS:fuzzy granular ball clustering-based partial multilabel feature selection with fuzzy mutual information[J].IEEE Transactions on Emerging Topics in Computational Intelligence,2025,9(1):590-606.

[12] Yang J,Liu Z Z,Xia S Y,et al.3WC-GBNRS:a novel three-way classifier with granular-ball neighborhood rough sets based on uncertainty [J].IEEE Transactions on Fuzzy Systems,2024,32(8):4376-4387.

[13] Sun L,Cai Y W,Ding W P,et al.Feature selection-based rough graph classification via graph convolutional neural networks[J].Engineering Applications of Artificial Intelligence,2025,158:111502.

[14] 丰昌武,孙林.融合 K-means 聚类 and 标记相关性的多标记 Relief 特征选择[J].聊城大学学报(自然科学版),2025,38(1):122-134.

Feng C W,Sun L.Multilabel relief-based feature selection via fusing K-means clustering and label correlation[J].Journal of Liaocheng University (Natural Science Edition),2025,38(1):122-134.

[15] Xia S Y,Zhang H,Li W H,et al.GBNRS:a novel rough set algorithm for fast adaptive attribute reduction in classification[J].IEEE Transactions on Knowledge and Data Engineering,2022,34(3):1231-1242.

[16] Ji X,Peng J H,Zhao P,et al.Extended rough sets model based on fuzzy granular ball and its attribute reduction[J].Information Sciences, 2023,640:119071.

[17] Li Y,Wu X X,Wang X Z.Incremental reduction methods based on granular ball neighborhood rough sets and attribute grouping[J].International Journal of Approximate Reasoning,2023,160:108974.

[18] Zhang J J,Sun K,Huang B,et al.Attribute reduction based on a rapid variable granular ball generation model[J].Expert Systems with Applications,2025,265:126030.

[19] Xia S Y,Wang C,Wang G Y,et al.GBRs:a unified granular-ball learning model of pawlak rough set and neighborhood rough set[J].IEEE Transactions on Neural Networks and Learning Systems,2025,36(1):1719-1733.

[20] Zhang L J,Lin G P,Wei L,et al.A novel ANN-based feature subset selection in multi-scale granular ball neighborhood decision tables[J]. Neural Networks,2025,185:107178.

[21] Yao Y Y.Three-way decisions with probabilistic rough sets[J].Information Sciences,2010,180(3):341-353.

[22] Deng J,Zhan J M,Herrera-Viedma E,et al.Regret theory-based three-way decision method on incomplete multiscale decision information systems with interval fuzzy numbers[J].IEEE Transactions on Fuzzy Systems,2023,31(3):982-996.

[23] Savchenko A V,Savchenko L V.Three-way classification for sequences of observations[J].Information Sciences,2023,648:119540.

[24] Guo D D,Jiang C M,Wu P.Three-way decision based on confidence level change in rough set[J].International Journal of Approximate Reasoning,2022,143:57-77.

[25] Gao C,Wang Z C,Zhou J,et al.Analysis of core attribute and approximate reduct based on the three-way decision[J].Applied Soft Computing,2024,150:111117.

Three-way attribute reduction based on granular-ball neighborhood rough set

Xin Xianwei^{a,b}, Ding Xiaodi^a, Pang Wenli^c, Ma Yuanyuan^a, Xue Zhanao^a

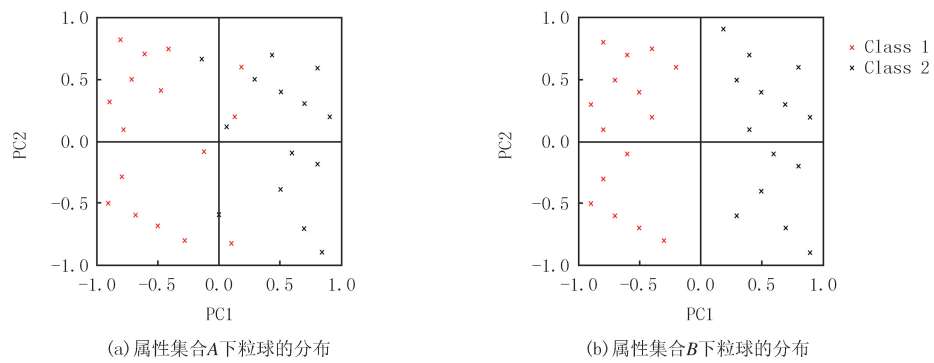
(a. School of Computer and Information Engineering(School of Artificial Intelligence); b. Henan Provincial Key Laboratory of Education Artificial Intelligence and Personalized Learning; c. College of Software, Henan Normal University, Xinxiang 453007, China)

Abstract: This paper proposes a three-way attribute reduction algorithm based on GBNRS. First, a purity measure based on granular-ball sample distribution is constructed. By simultaneously considering the sample distribution within granular balls and the distribution among granular balls, the traditional measure is improved to evaluate the contribution of attribute subsets to decision classification more accurately. On this basis, a three-way decision mechanism is introduced to divide attributes into core, candidate, and redundant attributes, enabling hierarchical screening and progressive reduction, thereby reducing the risk of misjudgment and improving reduction quality. Finally, comparative experiments with multiple methods are conducted on 12 UCI public datasets. The results show that the proposed algorithm can reduce the reduct length while maintaining good classification performance.

Keywords: granular-ball computing; granular-ball neighborhood rough set; attribute reduction; three-way decision

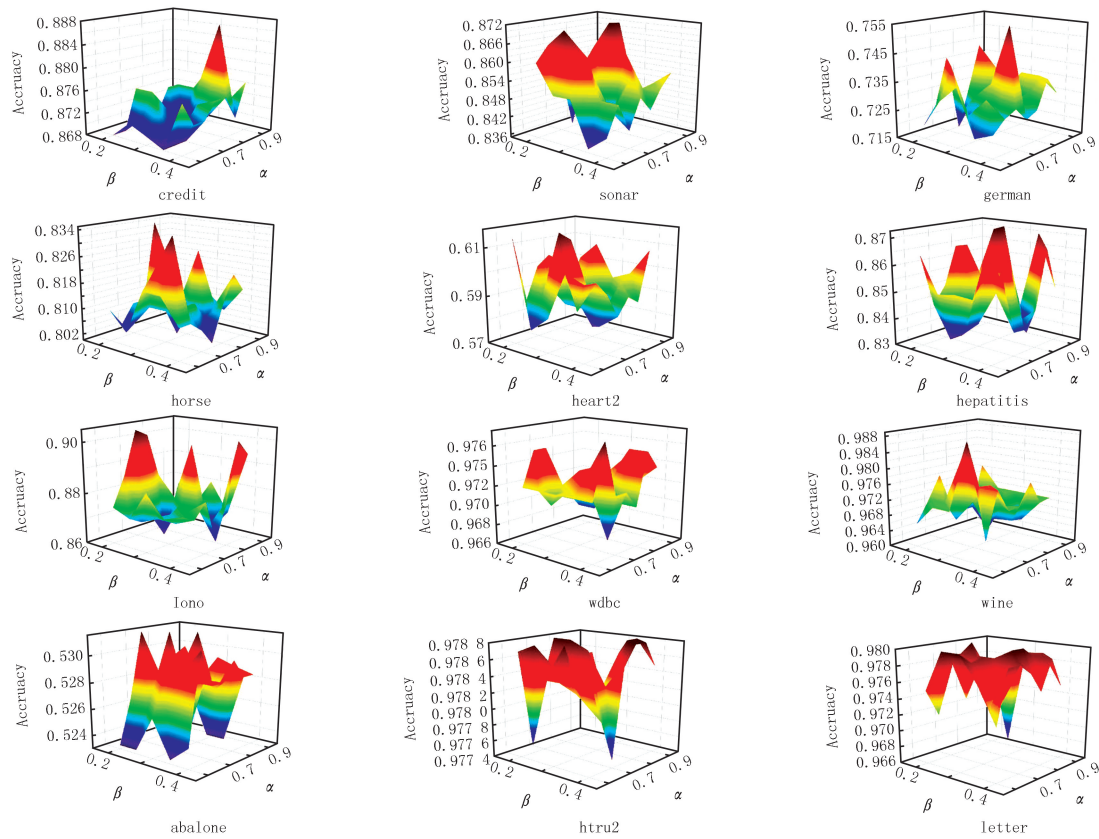
[责任编辑 陈留院 杨浦]

附 录



图S1 中心区域粒球的空间分布

Fig.S1 The distribution among granular balls in the central region

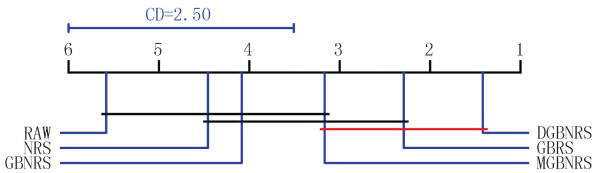


图S2 (α, β) 变化时的分类准确率

Fig.S2 The classification accuracy with (α, β) changed

表 S1 不同模块下的分类准确率
Tab. S1 The classification accuracy under different modules

No.	MGBNRS	+DP 模块	+三支约简模块	DGBNRS	No.	MGBNRS	+DP 模块	+三支约简模块	DGBNRS
1	0.873	0.877(↑)	0.875(↑)	0.887(↑)	7	0.883	0.863	0.877	0.909(↑)
2	0.841	0.846(↑)	0.850(↑)	0.879(↑)	8	0.974	0.975(↑)	0.973	0.977(↑)
3	0.731	0.729	0.742(↑)	0.755(↑)	9	0.978	0.977	0.972	0.989(↑)
4	0.834	0.815	0.823	0.840(↑)	10	0.978	0.979(↑)	0.979(↑)	0.979(↑)
5	0.613	0.600	0.621(↑)	0.624(↑)	11	0.527	0.531(↑)	0.532(↑)	0.532(↑)
6	0.876	0.851	0.851	0.890(↑)	12	0.979	0.979	0.979	0.979



图S3 不同算法在分类准确率上的Nemenyi检验

Fig.S3 The Nemenyi tests between different algorithms on classification accuracy