



· 粗糙集与形式概念分析中的约简研究 ·

【主持人语】粗糙集与形式概念分析是处理不确定性问题的两大工具,也被认为是粒计算的两种具体模型。二者的理论思想相通、研究目标接近、研究方法相似,但也有各自的研究特色。本期论文聚焦粗糙集与形式概念分析中的约简问题展开。针对粗糙集属性约简中关键属性可能被误删这一问题,给出一种基于信息熵加权的粒球邻域粗糙集属性约简方法;针对形式概念分析中的概念约简问题,探讨概念约简中元素数量与形式背景的基之间的关系,以及两类特殊概念约简的性质,并提出一种通过概念约简来还原概念格的方法;在由形式概念分析理论扩展产生的三元概念分析框架下,给出一种从三元粒概念出发进行三元概念获取的方法。

【主持人】魏玲,西北大学数学学院教授,博士生导师。

基于信息熵加权的粒球邻域粗糙集属性约简方法

曾诗淇¹, 刘 盾²

1. 西南交通大学 计算机与人工智能学院, 四川 成都 611756; 2. 西南交通大学 经济管理学院, 四川 成都 610031

摘要 粒球邻域粗糙集模型是一种融合粒计算与邻域关系的改进型粗糙集模型,能够有效处理连续值数据。然而,现有模型在属性权重不一致的情况下难以合理评估各属性的重要性,导致关键属性在约简过程中易被误删。为解决该问题,提出了一种基于信息熵加权的粒球邻域粗糙集属性约简方法。首先,结合属性内部分布特征与条件属性和决策属性之间的相关性,对属性赋予权重;其次,构建基于加权粒球邻域关系的改进模型,并定义相应的属性约简方法;最后,基于粒球平均纯度实现属性约简过程。实验在多个UCI数据集上进行,结果表明:该方法不仅具有较高的属性约简率(最高可达76.92%),还能显著提升支持向量机(SVM)和K近邻(KNN)分类器的平均准确率,分别提高1.57%和1.58%;此外,该方法在处理高维复杂数据时表现出更强的适应性和稳定性,具有良好的应用前景。

关键词 粒计算;属性约简;信息熵;加权粒球邻域关系;粒球邻域粗糙集

中图分类号: TP181 **DOI**: 10.16152/j.cnki.xdxbzr.2026-01-012

IEWGBNRS: A novel attribute reduction approach based on information entropy-weighted granular-ball neighborhood rough set

ZENG Shiqi¹, LIU Dun²

1. School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China;

2. School of Economics and Management, Southwest Jiaotong University, Chengdu 610031, China

收稿日期: 2025-07-10 修回日期: 2025-08-15

基金项目: 国家自然科学基金(62276217, 62402424); 中央高校基本科研业务费项目(2682024KJ005, 2682024ZTPY021)。

第一作者: 曾诗淇, 女, 硕士研究生, 从事粒计算与粗糙集理论及其应用研究, zengshiqi9@163.com。

通信作者: 刘盾, 男, 教授, 博士生导师, 从事粒计算与人工智能研究, newton83@163.com。

Abstract The granular ball neighborhood rough set model is an improved rough set model that integrates granular computing and neighborhood relations, capable of effectively handling continuous-valued data. However, existing models struggle to reasonably evaluate attribute importance under inconsistent attribute weights, leading to the risk of mistakenly removing key attributes during attribute reduction. To address this issue, we propose an attribute reduction algorithm for the granular ball neighborhood rough set model based on information entropy weighting. First, attribute weights are assigned by combining the internal distribution characteristics of attributes with the correlation between conditional and decision attributes. Second, an improved model based on the weighted granular ball neighborhood relation is constructed, and the corresponding attribute reduction method is defined. Finally, attribute reduction is implemented based on the average purity of granular balls. Extensive experiments are conducted on multiple UCI datasets. The results show that the proposed method not only achieves a high attribute reduction rate (up to 76.92%), but also significantly improves the average classification accuracy of classifiers such as Support Vector Machine (SVM) and K-Nearest Neighbors (KNN), increasing by 1.57% and 1.58%, respectively. Moreover, the method demonstrates stronger adaptability and stability in processing high-dimensional and complex datasets, indicating promising application potential.

Keywords granular computing; attribute reduction; information entropy; weighted granular-ball neighborhood relationships; granular-ball neighborhood rough set

粗糙集理论(rough set theory, RST)最早是由波兰数学家 Pawlak 于 1982 年提出的^[1]。作为一种能够处理不确定、模糊数据的软计算工具^[2],被广泛地应用到知识获取^[3]、图像处理^[4]、人工智能^[5]、医疗诊断^[6]等领域。然而,经典粗糙集模型在处理连续属型和复杂数据结构时存在局限性:一方面,经典粗糙集模型无法直接处理连续型或数值型数据,需对数据进行离散化,可能会导致信息损失;另一方面,经典粗糙集模型无法处理隶属度介于 0 和 1 之间的模糊情况,同时也缺乏对犹豫性的支持等。为了克服上述缺陷,国内外学者相继提出了邻域粗糙集^[7]、模糊粗糙集^[8]、直觉模糊粗糙集^[2]等,很好地扩展了粗糙集的应用范围。Yuan 等人提出一种变精度复合度量方法对局部粗糙集进行特征选择^[9];Yu 等人研究了基于变精度多粒度决策理论粗糙集的多标签学习特征选择^[10];Zhang 等人基于集成经验模态分解的核多粒度双定量粗糙集,并将其应用在股价趋势预测中^[11];Wang 等人基于多粒度云粗糙集提出一种新颖的多标准决策框架,并以此评估环经济实践中的能源互联网项目^[12]。

一般而言,大多数模糊粗糙集算法主要通过隶属函数处理连续属性^[13-16],其处理过程需要依赖数据集的先验知识,而邻域概念的提出很好地克服了这一局限性。Hu 等人基于域关系提出了

邻域粗糙集(neighborhood rough set, NRS)的概念^[7]。该模型利用邻域关系替代传统的等价关系,不仅能够有效地刻画样本间的相似性,而且不依赖于先验知识,一经提出便逐渐成为粗糙集领域的重要研究方向之一。然而,NRS 中每个相似类的邻域半径都是固定的,而邻域半径的优化依赖于网格搜索,不仅降低了效率和效果,还造成了大量时间消耗^[17]。因此,Xia 等人将粒球的概念引入邻域粗糙集中,提出了一种新颖的粒球邻域粗糙集模型(granular ball neighborhood rough set, GBNRS)^[18]。该模型通过构建基于粒球的邻域关系来改进传统粗糙集模型在处理不精确、不确定信息时的局限性,以处理连续型数据和复杂的数据结构。GBNRS 能够自适应地为每个对象生成不同的邻域,为数据处理和分类任务带来了更大的通用性和灵活性。

然而,尽管粒球邻域粗糙集在属性约简中展现出良好的鲁棒性和泛化能力^[18],但其在属性权重分配方面仍存在不足。传统 GBNRS 模型将所有属性视为同等重要,采用等权重处理方式,忽略了不同属性在数据分布、决策相关性等方面的本质差异。这种对属性重要性缺乏区分的机制,导致模型在面对特征分布不均或决策结构复杂的实际场景时,难以有效识别关键属性,在属性约简过程中可能出现冗余属性保留或关键属性遗漏的问

题,从而影响最终的分类性能。信息熵作为衡量数据不确定性的指标,在数据处理和知识发现中发挥着重要的作用^[19-21]。针对传统粒球邻域粗糙集属性约简方法的局限性,将信息熵加权引入其中具有重要的理论和实践意义。

本文旨在提出一种基于信息熵加权的粒球邻域粗糙集属性约简方法 (information entropy weighted granular ball neighborhood rough set, IEWGBNRS)。该方法引入信息熵作为属性不确定性度量工具,并结合互信息衡量属性与决策之间的相关性,构建综合指标以动态计算每个属性的权重。通过这种方式,IEWGBNRS 能够根据数据的实际分布特性自适应地调整属性的重要性排序,使属性约简过程更贴近真实的数据结构,提升了约简结果的准确性和稳定性。

与传统 GBNRS 相比,本文方法的主要创新点体现在以下 3 个方面。

1) 属性权重机制创新。将熵与互信息相结合,提出了一种新型复合度量方法,用于动态量化属性的重要性。

2) 约简策略优化。在粒球邻域粗糙集框架中引入属性加权思想,重构属性重要性评估体系,实现更细粒度的属性筛选。

3) 模型适用性增强。通过差异化赋权机制,提升模型在高维、复杂数据集上的适应能力和分类性能。

基于上述分析,本文的结构安排如下:首先,详细介绍基于信息熵加权的粒球邻域粗糙集属性约简方法的理论基础,包括粒球邻域粗糙集的相关概念和属性约简的基本原理;其次,定义新型信息熵的计算方法,提出具体的属性约简算法,包括信息熵加权的计算步骤和约简过程;最后,通过在多个 UCI 数据集上与现有属性约简方法进行对比实验,验证了所提方法在分类准确率、冗余属性去除率等方面的优势。实验结果表明,IEWGBNRS 不仅能够更有效地识别关键属性,还能提升后续分类器的性能表现,为复杂数据环境下的高效特征选择提供了新的解决方案。

1 相关知识

本节主要介绍粒球邻域粗糙集的相关定义和粗糙集的一些基础知识。

定义 1^[18] 给定一个数据集 $U = \{x_i, i = 1,$

$2, \dots, n\}$, 其中 x_i 和 n 分别表示样本和样本的数量。 $G = \{GB_1, GB_2, \dots, GB_m\}$ 是 U 上生成的一组粒球, 其中第 j 个粒球 $GB_j = \{x_i \mid i = 1, 2, \dots, k\}$ 。

一般而言, 每个生成粒球取决于两个关键参数: 中心 c_j 和半径 r_j 。粒球的引入可以使任意维度的数据都容易进行特征化。具体而言, 对于每个 GB_j , 中心 c_j 和半径 r_j 定义如下:

$$c_j = \frac{1}{k} \sum_{i=1}^k x_k \tag{1}$$

$$r_j = \max(\text{distance}(c_j, x_k)) \tag{2}$$

式中: $\text{distance}(c_j, x_k)$ 为 c_j 和 x_k 的欧式距离。

定义 2^[18] 设 $GB = \{x_i \mid i = 1, 2, \dots, k\}$ 是一个粒球, x_i 为粒球 GB 中的样本, k 为粒球 GB 中的样本个数, 设 M 为粒球 GB 中样本标签占比最大的样本数, 则可定义粒球 GB 的纯度为

$$\text{purity}(GB) = \frac{M}{k} \tag{3}$$

定义 3^[14] 设 $IS = (U, A, D)$ 是一个决策信息系统, 其中 $U = \{x_1, x_2, \dots, x_n\}$ 是一个非空有限对象集, 称为论域; $A = \{a_1, a_2, \dots, a_m\}$ 是一个非空有限条件属性集, 用于描述样本; D 为决策属性集。

设 $IS = (U, A, D)$ 是一个决策信息系统, 令 $G = \{GB_1, GB_2, \dots, GB_m\}$ 为 U 上生成的一组粒球。对于 $x_i \in GB_j, B \subseteq A, x_i$ 在特征空间 B 的粒球邻域类 $\delta_B(x_i)$ 表示为

$$\delta_B(x_i) = \{x \mid x \in U, \Delta_B(x_i, x) \leq r_j\} \tag{4}$$

式中: Δ_B 是一个距离函数, 表示为

$$\Delta_B(x_1, x_2) = \left(\sum_{a_i \in B} (f(x_1, a_i) - f(x_2, a_i))^p \right)^{1/p} \tag{5}$$

当 $p = 1$ 时, 为曼哈顿距离; 当 $p = 2$ 时, 为欧几里得距离; 当 $p = \infty$ 时, 为切比雪夫距离。本文中所有的距离均使用欧几里得距离, 即取 $p = 2$ 。

定义 4 设 $IS = (U, A, D)$ 是一个决策信息系统, $U/D = \{X_1, X_2, \dots, X_N\}$ 是 D 在 U 上的划分, $B \subseteq A, GB_j(B)$ 是在条件属性 B 下生成的第 j 个粒球, 对于 $X_l \subseteq U/D, X_l$ 关于 B 的上近似和下近似定义如下:

$$\overline{GR_B X_l} = \{x_i \in U, x_i \in GB_j \mid \delta_B(x_i) \cap X_l \neq \emptyset\} \tag{6}$$

$$\underline{GR_B X_l} = \{x_i \in U, x_i \in GB_j \mid \delta_B(x_i) \subseteq X_l\} \tag{7}$$

式中: k_j 表示 $GB_j(B)$ 中的样本数量。

那么,可以定义决策属性 D 关于条件属性 B 的上、下近似集如下:

$$\overline{GR_B D} = \bigcup_{l=1}^N \overline{GR_B X_l},$$
$$\underline{GR_B D} = \bigcup_{l=1}^N \underline{GR_B X_l} \quad (8)$$

定义 5 设 IS 是一个决策信息系统,对于条件属性集 $B \subseteq A$, B 是 A 关于上近似不变的属性约简,当且仅当满足以下条件:

$$\overline{R_B D} = \overline{R_A D} \quad (9)$$

$$\overline{R_B D} = \overline{R_{B-\{a\}} D}, \forall a \in B \quad (10)$$

定义 6^[18] 设 IS 是一个决策信息系统,对于条件属性集 $B \subseteq A$, B 是 A 的粒球纯度属性约简,当且仅当满足以下条件:

$$\text{purity}_B = \text{purity}_A \quad (11)$$

$$\text{purity}_B \neq \text{purity}_{B-\{a\}}, \forall a \in B \quad (12)$$

式中, purity_B 为基于属性子集 B 进行粒球计算得到的全部粒球的平均纯度。

Xia 等和 Cao 等提出的基于 2-means 的粒球生成方法^[22-23]如下:将论域 U 初始化为一个类簇,通过迭代对每个类簇进行 2-means 聚类,计算中心点与平均距离生成粒球及其纯度;当粒球纯度或粒球包含的样本个数达到设定阈值时终止算法,否则,继续分裂该类簇直至满足条件。本文的阈值选取与文献[23]中的一致,将纯度和样本个数均取为 1。

2 基于信息熵加权的粒球邻域粗糙集属性约简方法

在属性约简时,合理地对属性进行加权是提高模型性能的关键步骤之一。在众多加权方法中,信息熵是一种重要的赋权手段,很多学者也在这方面做了大量的工作。苗夺谦等人从互信息量角度对决策表中属性重要性进行定义,提出了基于互信息量的属性约简算法^[24];王国胤等人提出了基于信息熵的 CEBARKNC 和 CEBARKNCC 算法^[25];杨明提出了基于条件信息熵的近似约简算法^[26]。基于信息熵加权的属性约简方法能够有效提高数据处理效率、提升模型性能。然而,基于信息熵或条件熵的约简算法时间复杂度较高^[27],且未考虑属性内部的概率分布特征,无法充分挖掘数据内部的潜在信息。

匈牙利数学家 Rényi 针对于数值型数据提出了一个可以度量连续型随机变量的信息熵,称为 Rényi 熵。Rényi 熵依赖于参数 α ,它通过调整该参

数改变对数据分布中稀有事件的关注程度^[28],进而可以评估属性内部的概率分布特性。特别地,对于某些特定的 α 值(如 $\alpha = 2$),Rényi 熵具有简单的解析表达式,可以在实际应用中显著减少计算复杂度。相比之下,其他类型的熵不具备这样的特性。因此,在处理大规模数据集时,引入 Rényi 熵为属性赋予权重更为合适。此外,互信息可以衡量随机变量与条件变量之间的相关性,因此在对属性进行加权时,综合考虑互信息和 Rényi 熵,有助于平衡特征的重要性和质量。

对粒球邻域粗糙集进行属性加权可优化粒球生成过程,其通过改变特征空间增强样本的类内紧致性和类间分离性,具体表现为:首先,属性加权能够引导粒球覆盖更多同质样本,形成数量更少但质量更优的粒球结构,从而在降低计算复杂度的同时提升分类性能;其次,合理的特征权重分配可有效避免粒球冗余分裂,维持模型效率。图 1 所示案例中,三角形类粒球若在未加权的原始特征空间中过早终止分裂,将面临双重困境:其一,当前粒球纯度仅为 0.83,存在明显的类间混杂风险;其二,样本 x_2 与 x_3 因无法满足邻域覆盖条件而成为孤立点,最终导致样本信息丢失。虽然继续分裂可将纯度提升至 1.0,但会导致粒球数量增加并伴随着时间成本的增长。通过图 2 展示的属性加权解决方案,在对属性 a_1 施加 1.5 倍权重后,三角形类样本群($x_4 - x_8$)与正方形类样本群($x_1 - x_3$)的类间距增大,最终形成两个边界清晰的独立粒球,纯度均达 1.0。因此,在粒球邻域粗糙集中引入属性加权可以有效提升模型精度与效率。

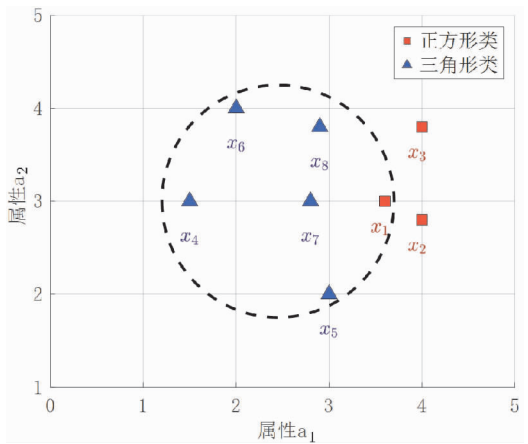


图 1 原始特征空间下的样本分布
Fig. 1 Sample distribution in native feature space

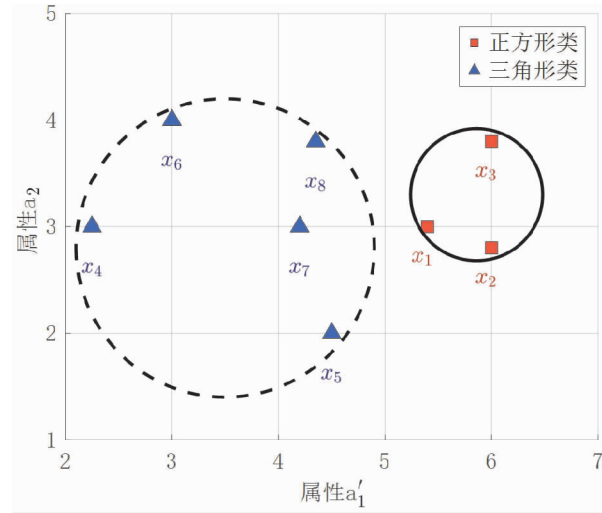


图 2 加权特征空间下的样本分布属性

Fig. 2 Sample distribution in weighted feature space

定义 7 设 $IS = (U, A, D)$ 是一个决策信息系统, $A = \{a_1, a_2, \dots, a_m\}$, 对于 $t \in a_i$ 的概率分布 $P(t)$, 属性 a_i 的 Rényi 熵 $H_\alpha(a_i)$ 定义为

$$H_\alpha(a_i) = \frac{1}{1-\alpha} \log \left(\sum_{t \in a_i} P(t)^\alpha \right) \quad (13)$$

式中: $\alpha > 0$ 且 $\alpha \neq 1$ 。特别地, 当 $\alpha = 2$ 时, $H_\alpha(a_i)$ 记为二阶 Rényi 熵, $H_2(a_i)$ 对概率分布中的小概率事件的敏感度降低, 在存在噪声或不确定性的场合下, 可能提供更稳定的结果^[28]。

定义 8 设 $IS = (U, A, D)$ 是一个决策信息系统, 对于特征值 $t \in a_i$ 和决策属性 D , 它们之间的互信息 $MI(a_i; D)$ 定义为

$$MI(a_i; D) = \sum_{t \in a_i} \sum_{X_j \in U/D} P(t, X_j) \log \left(\frac{P(t, X_j)}{P(t)P(X_j)} \right) \quad (14)$$

式中: $P(t, X_j)$ 是联合概率分布; $P(t)$ 和 $P(X_j)$ 是边缘概率分布。

定义 9 设 $IS = (U, A, D)$ 是一个决策信息系统, $\forall a_i \in A$, 属性 a_i 应同时反映其内部分布不确定性和与决策属性 D 的相关性, 于是定义属性 a_i 的综合评分 $S_{HMI}(a_i)$ 为

$$S_{HMI}(a_i) = H_\alpha(a_i) \cdot MI(a_i; D) = \frac{1}{1-\alpha} \log \sum_{t \in a_i} P(t)^\alpha \cdot \sum_{t \in a_i} \sum_{X_j \in U/D} P(t, X_j) \log \left(\frac{P(t, X_j)}{P(t)P(X_j)} \right) \quad (15)$$

进一步地, 为了消除因量纲导致数值差异悬殊带来的影响^[29], 需确保所有特征的权重总和为 1, 于是对 $S_{HMI}(a_i)$ 进行归一化处理, 则属性 a_i 的权重为

$$w(a_i) = \frac{S_{HMI}(a_i)}{\sum_{a_i \in A} S_{HMI}(a_i)} \quad (16)$$

对于任意属性 $a_i \in A$, 都有 $w(a_i) \geq 0$ 。

Rényi 熵 $H_\alpha(a_i)$ 衡量属性 a_i 自身的不确定性, 与属性的取值分布相关。对于低熵属性, 虽然取值集中、确定性高, 但可能缺乏区分不同类别的能力; 对于高熵属性, 虽然取值分散、多样性高, 可能包含更多潜在信息, 但也可能包含噪声, 提供的有效信息不足。而互信息 $MI(a_i; D)$ 衡量的是属性 a_i 对决策属性 D 的预测能力, 互信息越大, 表明 a_i 与 D 的相关性越强, 但某些特征可能在训练集中与目标变量偶然高度相关, 可能造成过拟合, 导致在测试集中的表现不稳定。

基于上述分析, 可以得到以下结论: 高 Rényi 熵的属性不一定对分类有用, 可能只是噪声; 低 Rényi 熵的属性可能冗余, 但若与决策高度相关, 则可能是关键属性。若仅最大化 $H_\alpha(a_i)$, 可能选择与决策属性 D 无关的高熵特征; 若仅最大化 $MI(a_i; D)$, 可能选择与决策属性 D 相关但分布单一的特征(如二值特征), 这类特征易受噪声干扰。因此, 综合评分通过熵与互信息的乘积能够平衡属性的不确定性以及与决策属性之间的相关性, 能够抑制高不确定性但低相关性的属性, 而保留低不确定性且高相关性的属性。

定理 1 设 $IS = (U, A, D)$ 是一个决策信息系统, $\forall a_i \in A$, 有 $S_{HMI}(a_i) \geq 0$ 。

证明 $S_{HMI}(a_i) = H_\alpha(a_i) \cdot MI(a_i; D)$, 因此, 仅需证明 $H_\alpha(a_i) \geq 0$ 和 $MI(a_i; D) \geq 0$ 。首先, 证明 $H_\alpha(a_i) \geq 0$ 。

$$\begin{aligned} H_\alpha(a_i) &= \frac{1}{1-\alpha} \log \left(\sum_{t \in a_i} P(t)^\alpha \right) = \\ &= \frac{\alpha}{1-\alpha} \cdot \frac{1}{\alpha} \log \left(\sum_{t \in a_i} P(t)^\alpha \right) = \\ &= \frac{\alpha}{1-\alpha} \log \left(\sum_{t \in a_i} P(t)^\alpha \right)^{\frac{1}{\alpha}} = \\ &= \frac{\alpha}{1-\alpha} \log \sqrt[\alpha]{\sum_{t \in a_i} P(t)^\alpha} \end{aligned}$$

1) 当 $\alpha \in (0, 1)$ 时, 对任意 $t \in a_i$, 有 $P(t)^\alpha \geq P(t)$, 则 $\sum_{t \in a_i} P(t)^\alpha \geq \sum_{t \in a_i} P(t) = 1$, 故 $\log \sqrt[\alpha]{\sum_{t \in a_i} P(t)^\alpha} \geq 0$, 又因为 $\frac{\alpha}{1-\alpha} > 0$, 则 $H_\alpha(a_i) \geq 0$ 。

2) 当 $\alpha \in (1, +\infty)$ 时, 对任意 $t \in a_i$, $0 \leq$

$P(t) \leq 1$, 有 $P(t)^\alpha \leq P(t)$, 则 $0 \leq \sum_{t \in a_i} P(t)^\alpha \leq$

$\sum_{t \in a_i} P(t) = 1$, 故 $\log \sqrt[\alpha]{\sum_{t \in a_i} P(t)^\alpha} \leq 0$, 又因为 $\frac{\alpha}{1-\alpha} < 0$, 则 $H_\alpha(a_i) \geq 0$.

综上可得, $\alpha \in (0, 1) \cup (1, +\infty)$ 时, 对于任意 $t \in a_i$, 有 $H_\alpha(a_i) \geq 0$.

互信息 $MI(a_i; D)$ 衡量的是属性 a_i 与决策属性 D 的相关性, 显然有 $MI(a_i; D) \geq 0$. 因此, 有 $S_{HMI}(a_i) \geq 0$, 证毕.

定理 2 考虑两个特征 a_1 和 a_2 , 假设 $H_\alpha(a_1) = h_1, H_\alpha(a_2) = h_2, MI(a_1; D) = mi_1, MI(a_2; D) = mi_2$, 则 $S_{HMI}(a_1) > S_{HMI}(a_2)$ 的必要条件是 $h_1 > h_2$ 且 $mi_1 > mi_2$.

证明 若 $h_1 > h_2$ 且 $mi_1 > mi_2$, 则必有 $S_{HMI}(a_1) > S_{HMI}(a_2)$; 反之, 若 $S_{HMI}(a_1) > S_{HMI}(a_2)$, 则存在多种情况. 第 1 种情况是 $h_1 > h_2$ 且 $mi_1 > mi_2$; 第 2 种情况是 $h_1 < h_2$, 但 $mi_1 \gg mi_2$; 第 3 种情况是 $h_1 \gg h_2$, 但 $mi_1 < mi_2$. 因此, $h_1 > h_2$ 且 $mi_1 > mi_2$ 是 $S_{HMI}(a_1) > S_{HMI}(a_2)$ 成立的必要非充分条件.

定理 3 设 $IS = (U, A, D)$ 是一个决策信息系统, $\forall a_i \in A$,

1) 若特征 a_i 与决策属性 D 独立, 则 $S_{HMI}(a_i) = 0$;

2) 若特征 a_i 的取值单一, 则 $S_{HMI}(a_i) = 0$.

证明 根据定义显然可得, 证明如下.

1) 若特征 a_i 与决策属性 D 独立, 对 $t \in a_i, X_j \in U/D$, 联合概率 $P(t, X_j) = 0$, 故 $S_{HMI}(a_i) = 0$.

2) 若特征 a_i 的取值单一, 设特征 a_i 只有一个取值 t , 故 $P(t) = 1$, 则 $H_\alpha(a_i) = 0$, 故 $S_{HMI}(a_i) = 0$.

定义 10 设 $IS = (U, A, D)$ 是一个决策信息系统, $w = (w(a_1), w(a_2), \dots, w(a_m))$ 为属性的权重, GB_j 是 U 上生成的第 j 个粒球, 粒球 GB_j 的加权中心 c_j^w 和加权半径 r_j^w 表示为

$$c_j^w = \frac{1}{k} \sum_{i=1}^k w \circ x_i \quad (17)$$

$$r_j^w = \max(\text{distance}(c_j^w, x_k)) \quad (18)$$

定义 11 设 $IS = (U, A, D)$ 是一个决策信息系统, 令 $w = (w(a_1), w(a_2), \dots, w(a_m))$ 为属性的权重向量, $G = \{GB_1, GB_2, \dots, GB_m\}$ 为 U 上生成的一组粒球. 对于 $x_i \in GB_j, B \subseteq A, x_i$ 在特征空间 B 的加权粒球邻域类 $W_{HMI}^B(x_i)$ 表示为

$$W_{HMI}^B(x_i) = \{x \mid x \in U, \Delta_B^w(x_i, x) \leq r_j\} \quad (19)$$

$$\Delta_B^w(x_i, x) = \left(\sum_{a_i \in B} w(a_i) (f(x, a_i) - f(y, a_i))^p \right)^{1/p} \quad (20)$$

式中: $w(a_i)$ 为属性 a_i 的权重; c_j 和 r_j 分别为粒球 GB_j 的中心和半径. 当 $w(a_i) \rightarrow 1$ 时, 属性 a_i 的重要程度越大; $w(a_i) \rightarrow 0$ 时, 属性 a_i 的重要程度越小.

定义 12 给定一个决策信息系统 $IS = (U, A, D)$ 和加权粒球邻域类 $W_{HMI}^B(x_i), X_l \subseteq U/D$, 决策属性 D 关于条件属性子集 B 的上、下近似定义为

$$\begin{aligned} \overline{GR}_B^w D &= \bigcup_{l=1}^N \overline{GR}_B^w X_l, \\ \underline{GR}_B^w D &= \bigcup_{l=1}^N \underline{GR}_B^w X_l \end{aligned} \quad (21)$$

式中:

$$\begin{aligned} \overline{GR}_B^w X_l &= \{x_i \in U, x_i \in GB_j \mid \\ &W_{HMI}^B(x_i) \cap X_l \neq \emptyset\}, \\ \underline{GR}_B^w X_l &= \{x_i \in U, x_i \in GB_j \mid W_{HMI}^B(x_i) \subseteq X_l\} \end{aligned} \quad (22)$$

定义 13 设 IS 是一个决策信息系统, 对于条件属性集 $B \subseteq A$, B 是 A 的基于加权的粒球纯度属性约简, 当且仅当满足以下条件:

$$\text{purity}_B^w = \text{purity}_A^w \quad (23)$$

$$\text{purity}_B^w \neq \text{purity}_{B-\{a_i\}}^w, \forall a_i \in B \quad (24)$$

式中, purity_B^w 为基于加权属性子集 B 进行粒球计算得到的全部粒球的平均纯度.

为了更好地理解上述相关定义, 下面将详细介绍本文所提出的基于信息熵加权的粒球邻域粗糙集属性约简方法 (information entropy weighted in granular ball neighborhood rough set, IEWGBNRS) 的核心思想、计算步骤及实现细节, 并分析其时间复杂度.

在计算属性权重时, 首先计算该属性下不同取值的概率分布, 根据式 (13) 计算出其 Rényi 熵, 对于目标属性下的每一个划分, 需要计算其与属性值的联合概率分布以及该划分的概率分布, 然后根据式 (14) 计算该属性关于目标属性 D 的互信息. 最后计算其综合评分, 并对所有的综合评分进行归一化处理. 具体而言, 相关步骤如算法 1 所示.

算法 1 基于互信息和 Rényi 熵的权重计算
输入: 决策信息系统 $IS = (U, A, D)$;

输出: 属性权重 $w = (w(a_1), w(a_2), \dots, w(a_m))$.

1) 令 $\mathbf{w} \leftarrow \varnothing$;

2) for $i = 1 \rightarrow m$ // m 为条件属性集 A 的个数

3) 计算边缘概率分布 $P(a)$;

4) 根据式(13) 计算 $H_{\alpha}(a_i)$;

5) for $j = 1 \rightarrow n$ // n 为 D 在 U 上的划分个数

6) 计算 $P(X_j)$ 和联合概率分布 $P(a, X_j)$;

7) end for

8) 根据式(14) 计算 $MI(a_i; D)$;

9) 计算属性综合得分 $w_{HMI}(a_i)$;

10) end for

11) for $i = 1 \rightarrow m$

12) 计算属性权重 $w(a_i)$;

13) 将 $w(a_i)$ 添加到 $\mathbf{w}$ 中;

14) end for

下面,简要分析算法 1 的时间复杂性,步骤 1 进行数据初始化,令权重向量 $\mathbf{w}$ 为空集,步骤 2 ~ 步骤 4 计算 a_i 的 Rényi 熵,时间复杂度为 $O(|A|)$,步骤 2 ~ 步骤 8 计算属性 a_i 关于目标属性 D 的互信息,时间复杂度为 $O(|A| \cdot |U/D|)$,故算法 1 的时间复杂度为 $O(|A| \cdot |U/D|) < O(mn)$ 。

此外,在计算 IEWGBNRS 属性约简时,首先需要根据条件属性 A 得到初始粒球计算所有粒球的纯度。接着依次遍历各个条件属性,若删除属性 a_i 后粒球的平均纯度发生变化,则说明该属性重要,不可约简;反之,若平均纯度保持不变,则表示该属性不重要,可进行约简。进一步地,具体步骤如算法 2 所示。

算法 2 IEWGBNRS 属性约简

输入:信息系统 $IS = (U, A, D)$, 权重向量 $\mathbf{w} = (w(a_1), w(a_2), \cdots, w(a_m))$;

输出:属性约简结果 reduction。

1) 令 $\text{reduction} \leftarrow A$;

2) 利用 $\mathbf{w}$ 计算特征空间 A 的加权粒球邻域类 $W_{HMI}^A(x_i)$;

3) 根据 $W_{HMI}^A(x_i)$ 计算平均粒球纯度 purity_A^w ;

4) for $i = 1 \rightarrow m$ // m 为条件属性集 A 的个数

5) if $|\text{reduction}| \leq 1$

6) break;

7) end if

8) 从 reduction 中删除属性 a_i ;

9) 计算属性 $A - \{a_i\}$ 下的加权粒球邻域类 $W_{HMI}^{A - \{a_i\}}(x_i)$;

10) 根据 $W_{HMI}^{A - \{a_i\}}(x_i)$ 计算 $\text{purity}_{A - \{a_i\}}^w$;

11) if $\text{purity}_A^w \neq \text{purity}_{A - \{a_i\}}^w$

12) 将属性 a_i 重新添加到 reduction 中;

13) else

14) continue;

15) end if

16) end for

在算法 2 中,步骤 1 ~ 步骤 3 进行数据初始化,首先将属性约简结果 reduction 初始化为 A ,然后计算加权邻域粒球类和属性集 A 下的粒球平均纯度,时间复杂度为 $O(n)$,步骤 4 ~ 步骤 16 遍历所有的条件属性,计算不同特征空间下的平均粒球纯度,并与原始平均粒球纯度进行比较,时间复杂度为 $O(m)$,因此,该算法的总时间复杂度为 $O(n) + O(m) = O(\max(m, n))$ 。

3 案例分析

本节给出一个具体案例说明本文所提方法的具体操作步骤。

表 1 给出了一个决策信息系统 $IS = (U, A, D)$,其中 $U = \{x_1, x_2, \cdots, x_6\}$, $A = \{a_1, a_2, a_3\}$ 。

首先,计算 $\alpha = 2$ 时属性 a_i 的 Rényi 熵 $H_{\alpha}(a_i)$ 。

$$H_2(a_1) = -\log((\frac{2}{6})^2 + (\frac{1}{6})^2 + (\frac{1}{6})^2 +$$

$$(\frac{1}{6})^2 + (\frac{1}{6})^2) = 0.65。$$

同理,可得 $H_2(a_2) = 0.56, H_2(a_3) = 0.56$ 。

表 1 决策信息系统

Tab. 1 Decision information system

U	a_1	a_2	a_3	D
x_1	1	2	1	1
x_2	8	3	2	2
x_3	6	3	8	2
x_4	3	1	1	1
x_5	1	5	2	1
x_6	2	1	5	2

然后,计算条件属性与决策属性之间的互信息。
 $MI(a_1; D) = 0.22, MI(a_2; D) = 0,$
 $MI(a_3; D) = 0.41。$
进而可以得到属性权重 $w = (0.38, 0, 0.62)$ 。

属性加权后得到的粒球为
 $GB_1 = \{x_1, x_4, x_5\}, GB_2 = \{x_2, x_3\}, GB_3 = \{x_6\}$ 。
此时, $\text{purity}_A^w = 1$ 。
当 $B = \{a_2, a_3\}$ 时, 计算得到的粒球为
 $GB_1 = \{x_1, x_4\}, GB_2 = \{x_2, x_5\},$
 $GB_3 = \{x_5\}, GB_4 = \{x_6\}$ 。
此时, $\text{purity}_B^w = (1 + 1/2 + 1 + 1)/4 = 0.875 \neq \text{purity}_A^w$, 故 $B = \{a_2, a_3\}$ 不是该信息系统的属性约简结果。

当 $B = \{a_1, a_3\}$ 时, 得到的粒球为
 $GB_1 = \{x_1, x_5\}, GB_2 = \{x_2, x_3\},$
 $GB_3 = \{x_4\}, GB_4 = \{x_6\}$ 。
此时, $\text{purity}_B^w = 1 = \text{purity}_A^w$, 故 $B = \{a_1, a_3\}$ 是该信息系统的属性约简结果。

当 $B = \{a_1, a_2\}$ 时, 得到的粒球为
 $GB_1 = \{x_1, x_5, x_6\}, GB_2 = \{x_2, x_3\},$
 $GB_3 = \{x_4\}$ 。
此时, $\text{purity}_B^w = (2/3 + 1 + 1)/3 = 0.89 \neq \text{purity}_A^w$, 故 $B = \{a_1, a_2\}$ 不是该信息系统的属性约简结果。

4 实验

4.1 实验设置

4.1.1 数据集

本实验所用到的 8 个数据均来自 UCI 数据集官网, 具体信息如表 2 所示。为消除数据量纲带来的影响, 使得不同特征具有可比性, 在实验前将所有属性的数值型数据均标准化至 $[0, 1]$ 区间内。

表 2 实验数据集的基本信息

Tab. 2 Basic information of experimental databases			
数据集	样本数	属性数	类簇数
Iris	150	4	3
Wine	178	13	3
Seeds	210	7	3
Sale	440	7	3
Raisin	900	7	2
Iranian	3 150	13	2
Rice	3 810	7	2
Landsat	6 435	36	7

4.1.2 对比方法

本文选取了经典的基于邻域粗糙集的属性约

简算法 (NRS) 和粒球邻域粗糙集的属性约简算法 (GBNRS) 作为两种基准算法, 并在上述的 3 个测度下与本文方法进行对比分析。

基于邻域粗糙集的属性约简算法 (NRS): 该算法利用域关系将数据划分为邻域相似类, 通过邻域相似类的上近似和下近似描述数据的不确定性, 并通过属性重要度或属性依赖度衡量每个属性对决策的贡献, 采用启发式搜索方法寻找最优的属性子集。

粒球邻域粗糙集属性约简算法 (GBNRS): 该算法通过生成自适应半径的粒球来替代传统的固定邻域半径, 结合了粒球计算和 k-means 聚类, 能够在属性约简时有效减少计算复杂度。

4.1.3 评价指标

为了验证本文提出的基于信息熵加权的粒球邻域粗糙集的属性约简算法有效性, 分别在约简率、分类精度和运行时间 3 个测度下进行分析。

1) 约简率^[30]是指在特征选择过程中, 减少的特征数量占原始特征数量的比例, 用于衡量特征选择方法的有效性, 即在保留重要信息的同时, 能够去除多少冗余或不重要的特征。约简率的定义为

$$\text{red_ratio} = \frac{|A| - |\text{reduction}|}{|A|} \times 100\%$$

式中: reduction 为属性约简集。

2) 分类精度是指分类模型在测试数据集上的正确分类率, 用于衡量模型的性能, 即模型对新数据的预测能力。

3) 运行时间是指特征选择方法或分类模型在数据集上运行所需的时间, 用于衡量算法的效率, 即算法在实际应用中的时间成本。

4.2 实验结果

本文所有实验环境配置为 Intel (R) Core (TM) Ultra 5 125H (1.20 GHz) 和 32.0 GiB 的内存个人计算机, 算法运行的软件环境为 Python 3.12.7。

4.2.1 约简率比较

为了能够清晰地展示各算法在特征选择方面的性能差异, 以 NRS 和 GBNRS 作为基准算法, 并进一步比较 IEWGBNRS 与两种基准算法在属性约简上的效果, 包括约简后属性个数以及约简率。

NRS、GBNRS 和 IEWGBNRS 3 种方法在 8 个数据集上约简结果中的属性个数对比如图 3 所示。从图 3 中可以看到, GBNRS 约简后属性个数最少, NRS 最多, 而 IEWGBNRS 介于二者之间。

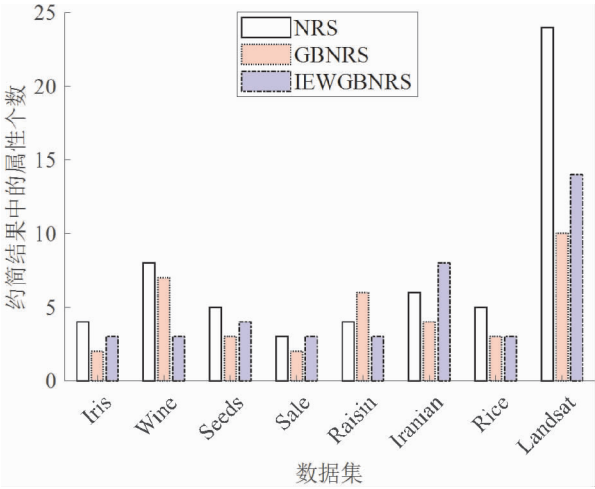


图 3 约简结果中的属性个数对比

Fig. 3 Attribution counts comparison after reduction

一般而言,约简率反应了约简方法识别冗余数据的能力,约简率越高代表其消除冗余属性的能力越强。从表 3 约简后的属性个数及约简率对比可见,GBNRS 算法在 Iris、Seeds、Sale、Iranian、Landsat 数据集上表现出最高的约简率,表明其去除冗余属性的能力最强。在 Wine 和 Raisin 数据集上,IEWGBNRS 取得最高约简率,在 Rice 数据集上,与 GBNRS 表现出相同的约简率,且高于 NRS。但在其他数据集上,其表现优于 NRS 但不及 GBNRS。其原因在于采取属性加权策略,属性的权重会促进或抑制属性的重要性。

表 3 约简结果对比

Tab. 3 Comparison of attributes

数据集	NRS		GBNRS		EWGBNRS	
	个数	约简率 /%	个数	约简率 /%	个数	约简率 /%
Iris	4	0	2	50	3	25
Wine	8	38.46	7	46.15	3	76.92
Seeds	5	28.57	3	57.14	4	42.80
Sale	3	57.14	2	71.42	3	57.14
Raisin	4	42.85	6	14.28	3	57.14
Iranian	6	53.84	4	69.23	8	38.46
Rice	5	28.57	3	57.14	3	57.14
Landsat	24	33.33	10	72.22	14	61.11

4.2.2 分类精度比较

为了系统地评估不同约简算法所选属性的分

类性能,在比较分类精度时引入了仅做标准化处理的原生特征空间(native feature space, NFS)作为基准线。通过与该基准对比,可以清晰地了解不同算法属性约简后数据集的表现是有所提升、保持不变还是出现下降,这样做的目的是确保能够准确衡量各约简算法对分类效果的实际影响。

本文采用了支持向量机(support vector machine, SVM)和 K 近邻(K-nearest neighbors, KNN)两种经典的分类器。参考文献[18],本文将 SVM 分类器中核函数设置为“poly”,正则化参数设置为 1.5。KNN 分类器中的 K 值均取为 5。对于 IEWGBNRS 算法和 GBNRS 算法,其粒球是根据粒球纯度自动生成的,因此无需设置邻域半径,对于 NRS 算法而言,将根据具体情况为其选取合适的邻域半径。图 4 为 NFS 与 3 种方法在 SVM 分类器上分类结果的对比图,表 4 为 SVM 分类器上的分类结果对比。可以看到,IEWGBNRS 方法在多个数据集上的表现优于其他方法。具体而言,IEWGBNRS 在 Iris、Seeds 和 Sale 数据集上的分类准确率分别比 NFS 提高了 0.95%、3.41%和 1.30%。对于 Iranian 数据集,IEWGBNRS 的表现与 NFS 持平,但相比 NRS 方法,其准确率提升了 0.95%。在 Wine 数据集上,NRS 方法表现欠佳,相对 NFS 下降了 6.30%,而 IEWGBNRS 方法仅降低了 0.40%,显示出更好的稳定性。在 Rice 和 Landsat 数据集上,IEWGBNRS 方法与 NFS 表现基本持平,准确率分别为 0.930 2 和 0.870 4,仅降低了 0.19%和 0.30%。

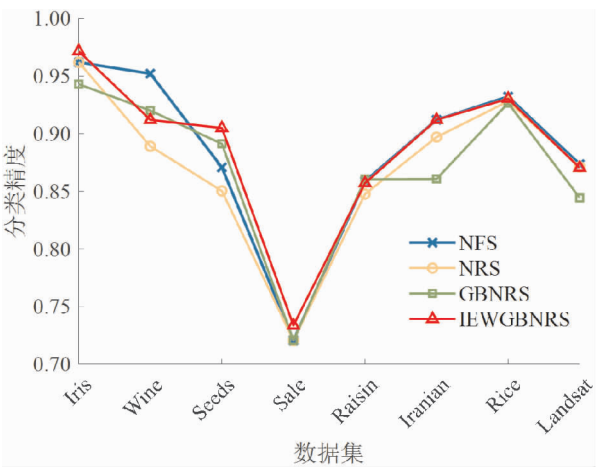


图 4 SVM 分类器上的分类精度对比

Fig. 4 Accuracy comparison on SVM classification

表 4 SVM 分类结果
Tab. 4 Classification results on SVM

数据集	NFS	NRS	GBNRS	IEWGBNRS
Iris	0.961 9	0.961 9	0.942 8	0.971 4
Wine	0.952 0	0.889 0	0.920 0	0.912 0
Seeds	0.870 7	0.850 3	0.891 1	0.904 8
Sale	0.720 7	0.720 7	0.720 7	0.733 7
Raisin	0.858 7	0.847 6	0.860 3	0.857 1
Iranian	0.912 0	0.897 0	0.860 7	0.912 0
Rice	0.932 1	0.928 0	0.926 5	0.930 2
Landsat	0.873 4	0.871 2	0.844 3	0.870 4
Average	0.885 2	0.870 7	0.870 8	0.886 5

此外,从平均准确率来看,IEWGBNRS 方法平均准确率为 0.886 5,比使用 NFS(0.885 2)提高了 0.13%,比 NRS 方法(0.870 7)提高了 1.58%,比 GBNRS 方法(0.870 8)提高了 1.57%。

图 5 为 NFS 与 3 种方法在 KNN 分类器上分类结果的对比图,表 5 为 KNN 分类器上的分类结果对比。可以看到,IEWGBNRS 方法在多个数据集上表现优于其他方法。具体来看,IEWGBNRS 方法在 Wine、Iranian、Seeds、Sale、Raisin 数据集上的分类准确率分别比 NFS 提高了 1.86%、0.77%、0.04%、3.04%、0.75%。在 iris 数据集上与 NFS 的表现一致。在 Rice 数据集上,IEWGBNRS 方法与 GBNRS 方法表现相同,均为 0.916 8。在 Landsat 数据集上,IEWGBNRS 方法比 NFS 降低了 3.00%。

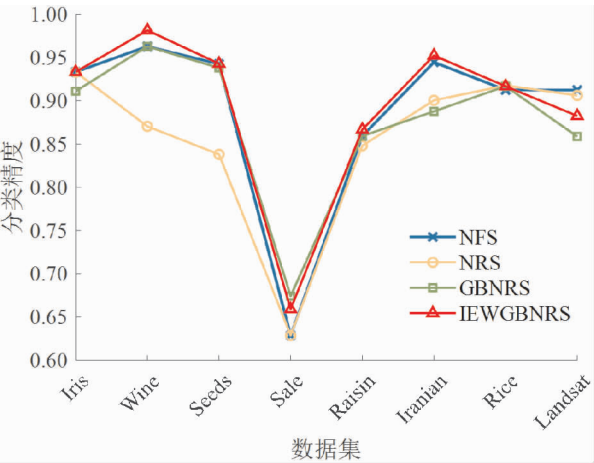


图 5 KNN 分类器上的分类精度对比图
Fig. 5 Accuracy comparison on KNN classification

表 5 KNN 分类结果
Tab. 5 Classification results on KNN

数据集	NFS	NRS	GBNRS	IEWGBNRS
Iris	0.933 3	0.933 3	0.911 1	0.933 3
Wine	0.962 9	0.870 4	0.962 9	0.981 5
Seeds	0.942 5	0.838 1	0.938 0	0.942 9
Sale	0.628 7	0.628 7	0.674 2	0.659 1
Raisin	0.859 2	0.848 1	0.859 2	0.866 7
Iranian	0.944 6	0.900 5	0.887 8	0.952 3
Rice	0.912 5	0.917 7	0.916 8	0.916 8
Landsat	0.912 4	0.906 2	0.858 6	0.882 5
Average	0.887 0	0.855 4	0.876 1	0.891 9

此外,从平均准确率来看,IEWGBNRS 方法平均准确率为 0.891 9,比使用 NFS(0.887 0)提高了 0.49%,比 NRS 方法(0.855 4)提高了 3.65%,比 GBNRS 方法(0.876 1)提高了 1.58%。

上述实验结果表明,IEWGBNRS 方法在处理这些数据集后,其平均分类正确率均较 NFS 有所提升,说明该方法能够更有效地挖掘数据特征,可以在简化数据的同时提高分类的准确性。总体而言,IEWGBNRS 方法在多数情况下具有一定的优势,尤其是在 Wine 和 Iranian 数据集上表现突出。

4. 2. 3 运行时间比较

对不同算法运行时间进行比较分析,相关结果如表 6 所示。可以看到,NRS 在 Iris 和 Wine 小型数据集上,展现出明显的优势,其运行时间显著短于 GBNRS 和 IEWGBNRS。然而,在 Iranian、Rice、Landsat 等较大规模的数据集上,NRS 的运行时间显著增加,特别是在 Landsat 数据集上,NRS 的运行时间达到了 770.523 1 s,远超其他两种模型。GBNRS 在 Seeds、Sale 等中等规模的数据集上表现良好,运行时间分别为 0.800 4 s 和 0.983 2 s。然而在大型数据集上的运行时间长于 IEWGBNRS,但仍优于 NRS,尤其是在 Landsat 数据集上,GBNRS 仅需 180.074 4 s,大大低于 NRS 的时间。IEWGBNRS 在 Raisin 和 Iranian 数据集上的运行时间略长,分别为 5.773 0 s 和 10.607 8 s。但在非常大的数据集上,如 Landsat,IEWGBNRS 表现出了显著的优化效果,仅需 149.493 0 s,明显快于 NRS 和 GBNRS。

表 6 3 种算法运行时间比较

Tab. 6 Running time comparison of three algorithms
单位: s

数据集	NRS	GBNRS	IEWGBNRS
Iris	0.030 1	0.180 7	0.190 2
Wine	0.126 9	0.691 3	1.297 9
Seeds	0.081 9	0.800 4	0.810 8
Sale	0.292 9	0.983 2	0.327 3
Raisin	1.019 3	2.361 1	5.773 0
Iranian	25.403 8	24.550 0	10.607 8
Rice	21.520 7	15.705 5	15.766 0
Landsat	770.523 1	180.074 4	149.493 0

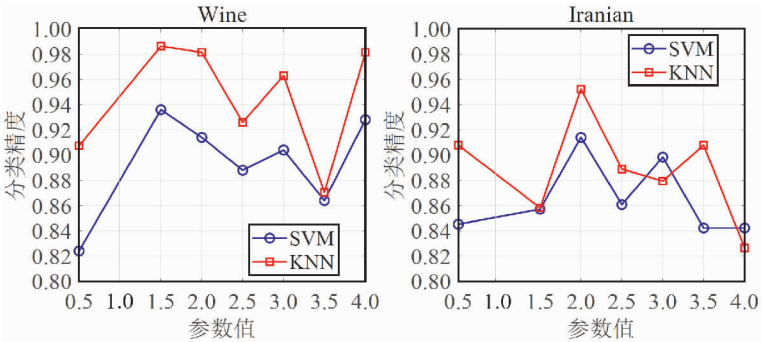


图 6 不同数据集上 α 的灵敏度分析

Fig. 6 Parameter sensitivity of α on different datasets

在对图 6 所示的 Wine 和 Iranian 数据集进行参数分析后,可以得出以下结论:对于 Wine 数据集,当 $\alpha = 1.5$ 时,SVM 和 KNN 的分类精度分别达到 93.60% 和 98.63% 的峰值,表明此时模型具有最佳的泛化能力。然而,随着参数值的进一步增加,两类模型的性能均出现不同程度的下降,特别是当 $\alpha = 3.5$ 时,SVM 和 KNN 的分类精度均降至最低点,分别为 86.40% 和 87.04%。对于 Iranian 数据集,可以观察到类似的趋势,当 $\alpha = 2$ 时,SVM 和 KNN 的分类精度达到最高点,分别为 91.20% 和 95.23%。值得注意的是,当 $\alpha = 3.5$ 时,SVM 的分类精度显著下降至 84.22%,KNN 的精度却相对上升。

综上,Wine 和 Iranian 数据集在 SVM 和 KNN 两种分类器上的分类精度均随参数值的变化而波动,且在 $\alpha = 2$ 的附近取得峰值,这是因为当 $\alpha = 2$ 时,可以通过平方项放大概率较高事件的影响,忽略小概率事件的噪声干扰。

4.4 实验结论

IEWGBNRS 模型通过综合考虑互信息和

4.3 参数灵敏度分析

参数灵敏度分析是模型优化和实际落地的重要步骤,在信息熵加权的粒球邻域粗糙集模型中需要的考虑的参数为 α ,它决定了属性的综合评分,最终影响属性的权重。本节分析了参数 α 在 $(0,4]$ 区间上不同取值对不同数据集的分类精度的影响,选取的 α 增长步长为 0.5,由于 $\alpha > 0$ 且 $\alpha \neq 1$,因此,分析了在 7 个不同取值上对分类精度的影响,这里选择对样本个数较少的 Wine 数据集以及样本个数较多的 Iranian 数据集进行参数分析,其可视化结果如图 6 所示。

Rényi 熵,可以平衡属性的重要性和质量,能够根据数据的分布和特征为属性赋予权重。引入的信息熵加权策略解决了现有模型在属性权重处理上的不足,使得在约简过程中能够更加精确地保留关键特征,使属性约简过程更加符合数据的实际特性,进而提高属性约简的准确性和效率,提升模型的整体性能。通过分析上述实验结果,可以得到如下 3 条结论。

1) IEWGBNRS 模型的约简率最高可达 76.92%,其平均约简率为 51.96%,介于 NRS (35.35%) 和 GBNRS 模型 (54.70%) 之间。在不同数据集下,IEWGBNRS 的约简率普遍高于 NRS,但在大多数情况下低于 GBNRS 模型。这表明通过为粒球邻域粗糙集的属性赋予权重,能够有效调节属性的重要性,从而在约简过程中实现更精准的特征选择。这种加权策略既能促进重要属性的作用,又能抑制不必要属性的干扰,从而提升了模型的整体性能。

2) 在 SVM 和 KNN 分类器上,IEWGBNRS 模型在两类分类器上的平均准确率分别为 88.65%

和 89.19%,均达到最高水平,与未进行属性加权的 GBNRS 模型相比分别提高了 1.57% 和 1.58%。并且 IEWGBNRS 模型在大多数情况下能够保持或提升分类精度,尤其是在具有复杂属性分布的数据集上表现优异,验证了 IEWGBNRS 方法所选择的特征对分类任务有重要贡献,说明通过属性加权,IEWGBNRS 模型能够有效地保留重要特征、去除冗余属性,从而提升分类性能。

3)IEWGBNRS 模型在 Landsat 这种数据量大且属性较为复杂的数据集上运行时间为 149.493 0 s,与 NRS 相比节省了 621.030 1 s,在时间上有显著提升。表明 IEWGBNRS 模型适合处理大型数据集,尤其是在属性更为复杂的情况下,能够更高效地处理数据,展现出更强的适应性和优越性。随着数据规模和复杂性的增加,IEWGBNRS 模型能够更好地应对挑战,保持高效的特征选择和分类性能。

5 结语

针对现有粒球邻域粗糙集模型在处理属性权重不一致问题上的局限性,提出了一种基于信息熵加权的粒球邻域粗糙集属性约简算法。该方法通过融合属性内部分布的信息熵与条件属性-决策属性间的关联强度,构建了双驱动赋权机制,从而建立了加权粒球邻域粗糙集模型,并定义了相应的属性约简方法。为验证方法有效性,本文基于 UCI 数据集设计对比实验,采用 3 组基准:仅标准化处理的原始特征空间(无约简)、基于邻域粗糙集的属性约简算法(无权重)及基于粒球邻域粗糙集的属性约简算法(无信息熵融合)。在 8 个数据集上的实验结果表明本文所提方法不仅能够有效提升 SVM 和 KNN 分类器的平均准确率,还能有效去除冗余属性,保留关键属性,从而提高了模型的分类性能。

尽管该算法在多个数据集上表现出了良好的性能,但仍有一些问题需要进一步研究。未来,基于信息熵的粒球邻域粗糙集属性约简的研究可以从下几个方向进行拓展。信息熵虽是重要的加权方法之一,但仍有待探索更优的属性赋权方式,以进一步提升约简效果;同时在实际应用中,属性的重要性不仅取决于信息量,其检测成本如数据采集难度、计算开销等同样具有决定性影响,将成本因素纳入特征选择框架,构建兼顾信息增益与代

价约束的多目标优化函数,将有望实现更具实用价值的属性约简方案。此外,可以结合多粒度粗糙集理论,深入研究信息熵在多粒度环境中的应用,进而实现更加灵活、高效的属性约简策略。针对动态数据场景,可以探索加权邻域粗糙集如何在数据变化时实现快速的属性约简与更新机制,以增强算法的适应性和实用性。

参考文献

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

[2] 桑彬彬,徐伟华. 直觉模糊序决策信息系统的分布约简[J]. 模糊系统与数学, 2017, 31(2): 164-170.

SANG B B, XU W H. Distribution reduction in an intuitionistic fuzzy ordered decision information system [J]. Fuzzy Systems and Mathematics, 2017, 31(2): 164-170.

[3] GARDINER E J, GILLET V J. Perspectives on knowledge discovery algorithms recently introduced in chemoinformatics: Rough set theory, association rule mining, emerging patterns, and formal concept analysis [J]. Journal of Chemical Information and Modeling, 2015, 55(9): 1781-1803.

[4] PETROSINO A, SALVI G. Rough fuzzy set based scale space transforms and their use in imageanalysis[J]. International Journal of Approximate Reasoning, 2006, 41(2):212-228.

[5] EMAM W, AHMMAD J, MAHMOOD T, et al. Classification of artificial intelligence tools for civil engineering under the notion of complex fuzzy rough Frank aggregation operators[J]. Scientific Reports, 2024, 14(1): 11892.

[6] EL-BABLY M K, ABU-GDAIRI R, EL-GAYAR M A. Medical diagnosis for the problem of Chikungunya disease using soft rough sets [J]. AIMS Mathematics, 2023, 8(4): 9082-9105.

[7] 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.

[8] SHEEJA T K, KURIAKOSE A S. A novel feature selection method using fuzzy rough sets[J]. Computers in Industry, 2018, 97: 111-116.

[9] YUAN K H, XU W H, MIAO D Q. A local rough set method for feature selection by variable precision composite measure [J]. Applied Soft Computing, 2024, 155: 111450.

- [10] YU Y, WAN M, QIAN J, et al. Feature selection for multi-label learning based on variable-degree multi-granulation decision-theoretic rough sets[J]. *International Journal of Approximate Reasoning*, 2024, 169: 109181.
- [11] ZHANG L, BAI J C, SUN B Z, et al. Kernel multi-granularity double-quantitative rough set based on ensemble empirical mode decomposition: Application to stock price trends prediction[J]. *International Journal of Approximate Reasoning*, 2024, 172: 109217.
- [12] WANG H, JU Y B, PORCEL C. Energy Internet project evaluation in circular economy practices: A novel multi-criteria decision-making framework with flexible linguistic expressions based on multi-granularity cloud-rough set[J]. *Computers & Industrial Engineering*, 2025, 201: 110890.
- [13] 徐伟华, 刘士虎, 张文修. 一般二元关系下基于粗糙隶属函数的程度粗糙集[J]. *重庆理工大学学报(自然科学版)*, 2010, 24(10): 101-108.
XU W H, LIU S H, ZHANG W X. The graded rough set based on membership function in general binary relation[J]. *Journal of Chongqing University of Technology*, 2010, 24(10): 101-108.
- [14] NIERADKA G, BUTKIEWICZ B. A method for automatic membership function estimation based on fuzzy measures[C]//*Foundations of Fuzzy Logic and Soft Computing*. Berlin, Heidelberg: Springer, 2007: 451-460.
- [15] JENSEN R, CORNELIS C. A new approach to fuzzy-rough nearest neighbour classification [M] // *Rough Sets and Current Trends in Computing*. Berlin, Heidelberg: Springer, 2008: 310-319.
- [16] 张小红, 裴道武, 代建华. 模糊数学与 Rough 集理论[M]. 北京: 清华大学出版社, 2013: 222-226.
- [17] YANG X, LI Y H, XIA S Y, et al. Granular-ball three-way decision[C]//*Rough Sets*. Cham: Springer, 2023: 283-295.
- [18] 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.
- [19] SHANNON C E. A mathematical theory of communication[J]. *The Bell System Technical Journal*, 1948, 27(3): 379-423.
- [20] FAYYAD U M, IRANI K B. Multi-interval discretization of continuous-valued attributes for classification learning[C]//*International Joint Conference on Artificial Intelligence*, 1993, 93(2): 1022-1029.
- [21] WITTEN I H, FRANK E. Data mining: Practical machine learning tools and techniques with Java implementations[J]. *Acm Sigmod Record*, 2002, 31(1): 76-77.
- [22] XIA S Y, DAI X C, WANG G Y, et al. An efficient and adaptive granular-ball generation method in classification problem[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, 35(4): 5319-5331.
- [23] CAO X M, YANG X, XIA S Y, et al. Open continual feature selection via granular-ball knowledge transfer[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2024, 36(12): 8967-8980.
- [24] 苗齐谦, 胡桂荣. 知识约简的一种启发式算法[J]. *计算机研究与发展*, 1999, 36(6): 681-684.
MIAO D Q, HU G R. A heuristic algorithm for knowledge[J]. *Journal of Computer Research and Development*, 1999, 36(6): 681-684.
- [25] 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): 759-766.
WANG G Y, YU H, YANG D C. Decision table reduction based on conditional information entropy[J]. *Chinese Journal of Computers*, 2002, 25(7): 759-766.
- [26] 杨明. 决策表中基于条件信息熵的近似约简[J]. *电子学报*, 2007, 35(11): 2156-2160.
YANG M. Approximate reduction based on conditional information entropy in decision tables[J]. *Acta Electronica Sinica*, 2007, 35(11): 2156-2160.
- [27] 杨传健, 葛浩, 汪志圣. 基于粗糙集的属性约简方法研究综述[J]. *计算机应用研究*, 2012, 29(1): 16-20.
YANG C J, GE H, WANG Z S. Overview of attribute reduction based on rough set[J]. *Application Research of Computers*, 2012, 29(1): 16-20.
- [28] VAN ERVEN T, HARREMOS P. Rényi divergence and kullback-leibler divergence[J]. *IEEE Transactions on Information Theory*, 2014, 60(7): 3797-3820.
- [29] JIN F F, GAO X X, ZHOU L G. Bounded rationality consensus reaching process with regret theory and weighted Moment estimation for multi-attribute group decision making[J]. *Information Fusion*, 2025, 115: 102778.
- [30] 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.