基于密度峰值聚类的超区间粒化方法及其分类模型

吴成英[1),2),4)] 张清华[1),2),3)] 赵 凡[1),2)] 程云龙[1),2)]
谢 秦[1),2)] 夏书银[1),2),3)] 王国胤[1),2),3)]

[1)](重庆邮电大学计算智能重庆市重点实验室 重庆 400065)
[2)](旅游多源数据感知与决策技术文化和旅游部重点实验室 重庆 400065)
[3)](重庆邮电大学大数据智能计算重点实验室 重庆 400065)
[4)](湖北民族大学 湖北 恩施 445000)

摘 要 大数据挖掘时代,数据丰富与知识贫乏之间的矛盾日趋突出.粒计算是解决大规模、复杂问题的新范式,其核心任务是粒化.粗糙集是经典粒计算模型之一,在数据挖掘领域已广泛应用.遗憾的是基于不可区分关系的粒化条件很严格,造成粗糙集在粒化定量数据时会失效.因此,本文首先从一维属性的区间划分出发,定义多维属性组合生成的超区间粒,并基于超区间粒提出新颖的粗糙集模型有效地将定量数据和定性数据统一到一个框架;其次,从决策属性的视角考虑条件属性之间的相关性提出基于密度峰值聚类的超区间粒化算法,算法输出的超区间粒不仅是论域的划分,且每个划分块都是同质信息粒;最后,受近邻分类算法的启发,融合多数投票分类机制和近邻分类准则基于超区间粒提出自适应近邻分类模型(IGANN),并在 UCI 数据集上与 8 个经典分类模型进行实验对比,4 个指标下的对比结果均表明 IGANN 模型具有更强的稳定性和更高的鲁棒性.

关键词 粒计算;粗糙集;粒化;信息粒;密度峰值聚类
中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2023.01620

Hyper-Interval Granulation Approach Based on the Density Peaks Clustering for Classification

WU Cheng-Ying[1),2),4)] ZHANG Qing-Hua[1),2),3)] ZHAO Fan[1),2)] CHENG Yun-Long[1),2)]
XIE Qin[1),2)] XIA Shu-Yin[1),2),3)] WANG Guo-Yin[1),2),3)]

[1)](*Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing* 400065)
[2)](*Key Laboratory of Tourism Multisource Data Perception and Decision, Ministry of Culture and Tourism, Chongqing* 400065)
[3)](*Key Laboratory of Big Data Intelligent Computing, Chongqing University of Posts and Telecommunications, Chongqing* 400065)
[4)](*Hubei Minzu University, Enshi, Hubei* 445000)

Abstract In the era of big data mining, the contradiction between rich data and poor knowledge becomes more and more prominent. Data reduction is still an indispensable technique for knowledge discovery, which aims to extract useful information from large-scale data and reduces the redundancy of data. Granular computing is a new paradigm for solving large-scale and complex problems, and its core task is granulation. Granulation is the process of decomposing a large-scale dataset into several subsets through a given granulation strategy to construct desirable information granules (IGs). Rough set is one of the classical models in granular computing and has been widely used in

收稿日期:2022-11-23;在线发表日期:2023-04-11.本课题得到国家自然科学基金(62276038,62221005)、国家重点研究发展计划(2020YFC2003502)、重庆市自然科学基金(cstc2019jcyj-cxttX0002)、重庆市研究生科研创新项目(CYB21207)和重庆邮电大学博士人才培养计划(BYJS202012)资助.**吴成英**,博士研究生,主要研究方向为知识发现和粒计算. E-mail: 381047936@qq.com. **张清华**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为粗糙集、模糊集、粒计算以及不确定性信息处理. E-mail: zhangqh@cqupt.edu.cn. **赵 凡**,博士研究生,主要研究方向为不确定性信息处理、模糊集和粗糙集. **程云龙**,博士研究生,主要研究方向为粒计算、三支决策以及粗糙集. **谢 秦**,博士研究生,主要研究方向为数据挖掘、知识发现以及粒计算. **夏书银**,博士,教授,博士生导师,主要研究领域为数据挖掘和粒计算. **王国胤**,博士,教授,博士生导师,中国计算机学会(CCF)会士,长江学者,主要研究领域为粒计算、人工智能和认知计算.

the field of data mining. However, the condition of granulation based on the indiscernibility relations is too strict, which makes rough set invalid when granulating quantitative data. Although granular ball computing (GBC) can efficiently use for the granulation of quantitative data, it has three problems as follows: (1) GBC uses a granular ball to represent an information granule, which may lose some boundary information; (2) In the process of granulation, the selection for the center of granular ball is not accurate enough; (3) Granular ball cannot adaptively characterize the sizes of IGs according to the differences among attributes, which is susceptible to the data distribution and boundary objects. Therefore, in this paper, an interval division is defined from the perspective one-dimensional attribute, and then interval equivalence relationship is defined under multi-dimensional attributes to induce a partition of a universe. Each partition unit is represented by a hyper-interval granule that overcomes the problem of information loss in GBC. With hyper-interval granules, a novel rough set model based on the hyper-interval granule is proposed first, which can effectively unify quantitative and qualitative data into one framework. Next, based on the density peaks clustering, a hyper-interval granulation algorithm (DPCIG) is proposed under considering the correlation between condition attributes from the perspective of decision attribute. DPCIG selects a object with the highest density as the initial center of a hyper-interval granule, which solves the problem that the center of granulation in GBC cannot be accurately determined. Besides, DPCIG can partition quantitative attribute values into a small number of intervals under the condition of all objects in a partition unit with the same labels. All objects in a partition unit have the same characterization ability from the perspective of decision attribute. For this reason, all objects in a partition unit defined as a hyper-interval granule, which can effectively achieve data reduction. Compared to granular balls generated by GBC, a set of hyper-interval granules outputted by the DPCIG is not only a partition of the universe, but also each information granule can adaptively adjust the interval length according to different attributes. Finally, inspired by the nearest neighbor classification algorithm, integrating the majority voting mechanism and nearest neighbor criterion, an adaptive nearest neighbor classification model based on hyper-interval granules (IGANN) is proposed. Compared with eight classical classification models on 21 widely-used UCI benchmark datasets, the experimental results under four indexes demonstrate that IGANN has stronger stability and higher robustness.

Keywords　granular computing; rough set; granulation; information granule; peaks density clustering

1　引　言

随着世界数字化进程的加快，各种形态的数据呈几何级数快速增长. 如今，数据丰富与知识贫乏之间的矛盾日趋突出. 数据质量和数据挖掘的效率仍是知识发现的瓶颈. 尽管算力和模型的发展已显著提高了知识发现的效率以及准确性，但不能完全满足大规模数据处理的需求. 数据约简仍是知识发现不可或缺的技术，旨在从大规模数据集中提取有意义、有用的信息，同时减少数据的冗余和复杂性.

粒计算(Granular Computing，GrC)是智能信息处理领域中通过模拟人脑的多粒度认知机理求解复杂问题的新范式[1]，推动了人工智能的发展. GrC 发轫于 1979 年美国工程院士 Zadeh 教授发表的论文“Fuzzy sets and information granularity”[2]，文中指出很多领域存在信息粒的概念. 历经 40 余年的发展，GrC 引起了学者的广泛关注，并针对不同问题提出了不同的粒计算模型，比如，1982 年波兰科学院院士 Pawlak 教授针对边界不确定性问题，提出了粗糙集模型[3]；1985 年 Hobbs 教授针对信息粒的分解与合并，提出了产生不同大小信息粒的模型[4]；1992 年清华大学张钹院士针对复杂问题，提出了分层求解的商空间理论[5]；1995 年李德毅院士针对模

糊与随机共存的不确定性问题，提出了云模型[6].

粒化是 GrC 的核心[7]，其主要任务是通过给定的粒化策略将复杂数据分解成若干子集，从而形成信息粒，以达到数据约简的目的. 基于不可区分关系[8-9]、相似关系[10-11]或近邻关系[12-13]是常用的粒化策略. 信息粒是 GrC 的基本计算单元，是数据建模和知识表示的关键组成部分，用信息粒代替样本作为模型的输入不仅可以简化问题、提高运行效率，而且可显著增强模型的抗噪性和可解释性[1,4,14]. 目前，粒计算已成功应用于机器学习[15]、数据挖掘[16-17]、知识发现[7,18-19]以及智能决策[20]等.

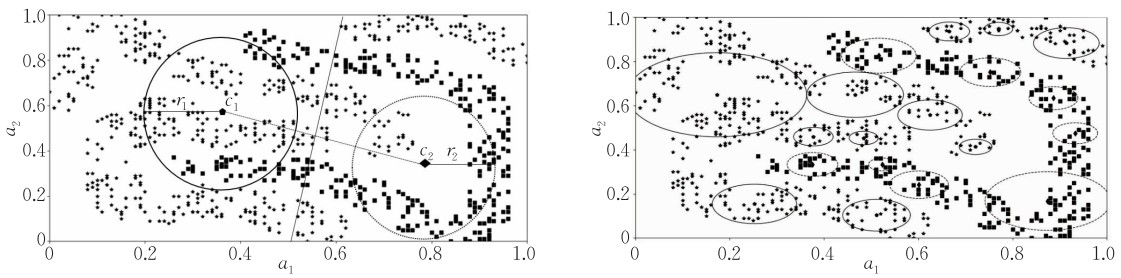
作为经典的 GrC 模型之一，粗糙集[3]是解决不精确、不确定性问题的有效理论与方法. 该模型首先基于不可区分关系将数据空间划分成互不相交的等价类；然后，基于等价类计算两个确定的集合，即上近似集和下近似集，逼近不确定的目标概念. 经过 40 年的发展，粗糙集在属性约简[21-23]、知识表示[24-25]、不确定性推理[26-27]、医疗诊断[28]等领域已取得广泛应用. 由于基于不可区分关系的粒化条件很严格，造成模型只能很好地粒化定性数据. 当粒化定量数据时，模型的性能会显著下降，甚至在粒化连续数据时会失效.

为了在定量的数据空间上实现数据粒化，2008 年 Hu 等人[29]将不可区分关系泛化到邻域关系提出了邻域粗糙集(NRS). 基于邻域关系生成的信息粒称为邻域粒，其大小取决于邻域半径. 在定量属性的数据空间上，当邻域半径取值为 0 时，邻域粒可视为一个等价类. 相比粗糙集中的信息粒(等价类)，NRS 中的邻域粒是数据空间上的覆盖而非划分. 近几年，NRS 已成功应用于规则学习[30]、推荐系统[31]、属性约简[32-33]以及医疗诊断[34]等. 但是，求解数据空间上的最优邻域半径是一个 NP 难问题，这在一定程度上阻碍了模型的发展. 另外，数据空间上的邻域粒相互之间存在大量覆盖甚至嵌套，使得信息之间存在冗余，进而影响规则学习的性能和效率.

针对 NRS 的两个问题，2020 年 Xia 等人[35]提出了粒球邻域粗糙集模型(GBNRS). GBNRS 基于 k-means 聚类算法自适应的在数据空间上生成大小不同的信息粒，即粒球. 生成粒球的过程为粒球计算[36]. 粒球计算简单易实现，能很好的逼近球状的数据分布，实现数据约简. 相比 NRS，GBNRS 具有更高的计算效率和更强的普适性. 目前，粒球计算是粒计算领域中新颖、高效的粒化方法，并成功应用于属性约简[35]、不平衡数据分类[37-38]、噪声检测[39]. 然而，粒球计算仍有以下三点不足：

(1) 粒球计算使用粒球表示信息粒，会损失一定的边界信息. 在粒化过程中，粒球计算首先将数据集划分成若干个数据子集；然后每个数据子集对应计算一个粒球. 粒球的中心点是数据子集的中心，粒球半径是数据子集中的对象到中心点的平均距离. 粒球半径的计算方法决定了粒球无法包含一个数据子集中到中心点距离大于粒球半径的对象. 换言之，一定存在边界对象在粒球之外. 为了便于理解，以 UCI 数据集 fourclass 为例，将基于 2-means 的粒球生成算法[37]的迭代过程可视化，算法第一次迭代如图 1(a)，最后一次迭代如图 1(b). fourclass 有 863 个对象、2 个条件属性和 2 个类标签，图中的坐标点代表对象在属性 a_1(横轴)和属性 a_2(纵轴)下的取值，不同的形状代表不同类的标签. 如图 1(a)所示，当确定两个中心点后，数据集将沿着两个中心点的垂直平分线分成两个子集. 显然，根据两个子集计算的两个粒球不能构成数据空间的覆盖.

(2) 在计算粒球的中心时，目前粒球计算是随机选择或按类别均值计算，前者存在很大的不确定性降低了算法效率，后者在处理不对称数据集时，计算的均值中心可能并非类簇中心. 如图 1(a)所示，fourclass 数据集按五角星类和正方形类计算的类簇中心分别为 c_1 和 c_2. 显然，c_2 离密度最高的区域有一定的距离，而且 c_2 处于边界位置. 若以 c_2 为正方形类的粒化中心，容易受五角星类的影响. 若能更准确



(a) 第一次迭代 (b) 最后一次迭代

图 1 fourclass 数据集在粒球计算算法的输出

地找到类别的密度中心，不仅有助于提高粒化的效率，而且能生成半径更大的信息粒，有助于提高模型的鲁棒性.

(3) 粒球在每个属性维度下的表示都严格围绕中心对称，从空间结构来看粒球的大小易受数据分布和边界对象的影响.

为了解决上述问题，本文首先基于区间划分定义能覆盖数据空间的超区间粒，并进一步提出超区间粒粗糙集模型. 然后，受 1982 年陈霖院士提出的“大范围首先”处理机制[40]和粒计算的“粒度思维”的启发[2]，基于密度峰值聚类算法[41]提出了超区间粒化算法(DPCIG)；最后，受 k 最近邻算法(k-Nearest Neighbor，kNN)的启发，融合多数投票分类机制和近邻分类准则，基于超区间粒提出自适应近邻分类模型(IGANN). 综上，本文主要有以下 3 个贡献点：

(1) 从一维属性出发定义定量属性的区间划分，并进一步定义多维属性下的区间等价关系诱导论域的划分，每个划分用超区间粒表示，克服了用粒球表示信息粒带来信息损失的问题. 相比粗糙集模型，提出的超区间粒粗糙集模型能有效地将定量属性和定性属性统一到一个框架. 研究区间划分的性质，得出区间等价关系是不可区分关系的推广.

(2) 提出的 DPCIG 算法首先基于密度峰值聚类算法，按决策类簇分别计算密度最大的对象作为超区间粒的初始中心，解决了粒球计算中的中心点无法准确确定的问题；然后，以生成同质信息粒为条件，围绕中心点沿着上下两个方向，从标签的视角在整个数据空间上求解多维定量条件属性形成区间的上边界和下边界，有效地将多个实数或者连续数据划分为少量的区间. 同一个区间内的数据相对决策标签而言视为相同的数据，即具有相同的刻画能力. 在属性区间划分的基础上生成的超区间粒在各个维度上能根据数据分布自适应地调整区间长度，解决了粒球在每个维度上的取值长度都相等的问题. 相比粒球计算的结果，DPCIG 算法输出的同质超区间粒不再围绕初始中心点对称，并构成了论域的划分.

(3) 实现了超区间粒到分类规则的转换；另外，为了解决新的测试对象不能与规则完全匹配的问题，提出了 IGANN 模型. 最后通过对比实验在 4 个分类指标下验证了 IGANN 的有效性.

本文第 2 节介绍与本文相关的背景知识，主要涉及粗糙集以及密度峰值聚类的基础定义；第 3 节介绍本文提出的区间划分、超区间粒、超区间粒化算法以及分类学习模型；第 4 节主要通过实验对比验证 IGANN 模型的鲁棒性和稳定性.

2　背景知识

为了便于理解，本节首先归纳了本文使用的概念符号，如表 1 所示. 然后，回顾了本文模型涉及的基础知识，主要包括决策系统、不可区分关系、粗糙集和密度峰值聚类.

表 1　概念符号

符号	概念
DS	DS 是决策系统，由论域 U、条件属性集 C、决策属性 d、属性的值域 V 以及映射函数 f 确定，见定义 1
$\mathcal{R}_B$	条件属性子集 $B\subseteq C$ 下的不可区分关系，见定义 2
$\mathcal{F}$	决策属性 d 下的不可区分关系 $\mathcal{R}_d$ 诱导论域 U 的划分，见第 4 页右栏 4 段
I_a^i	属性 $a\in B$ 的第 i 个区间划分块，由上边界 $\overline{V}_a^i$ 和下边界 $\underline{V}_a^i$ 确定，$\overline{V}_a^i,\underline{V}_a^i\in V_a$，见定义 5
$\mathcal{I}_B$	条件属性子集 B 下的区间等价关系，见定义 6
$\mathcal{G}_B$	$\mathcal{I}_B$ 诱导 U 的划分，见定义 7
G_B^j	G_B^j 为 $\mathcal{G}_B$ 中的第 j 个超区间粒，见定义 7
$\rho(X_j')$	决策类子集 X_j' 的密度中心，见定义 10

2.1　粗糙集

为了便于描述和研究，监督学习任务中的数据空间用决策系统表示，而无监督环境下的数据空间用信息系统表示.

定义 1.　决策系统[9]. 决策系统定义为五元组 $DS=(U,C,d,V,f)$，其中 U 是一个非空有限的集合，称为论域，论域中的元素 $x\in U$ 称为对象；C 是条件属性集合，d 是决策属性；$V=V_C\cup V_d$，$V_C=\{V_a\mid a\in C\}$ 是 C 的值域，V_a 是条件属性 a 的值域，$V_d=\{l_1,l_2,\cdots,l_z\}$ 是 d 的值域或称标签集，$f:U\times(C\cup\{d\})\to V$ 是一个映射函数使得 $f(x,a)\in V_a$，$f(x,d)\in V_d$.

决策系统是信息系统 $IS=(U,C,V,f)$ 的特殊情况. 相比决策系统，信息系统没有决策属性，即对象没有标签.

根据属性的取值不同，通常将条件属性分为定性属性和定量属性两大类，如图 2 所示. 名义有序属性的取值离散且有序，如等级评定属性的优、良、中、

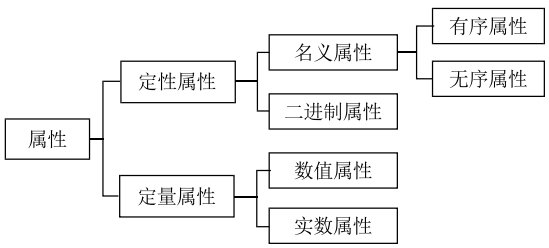


图 2　条件属性的分类

差;名义无序属性的取值离散且无序,如颜色属性;二进制属性仅两个状态,如性别;定量属性包括用整数表示的数值属性和用实数表示的连续属性.

定义 2. 不可区分关系[9]. 给定一个决策系统 $DS=(U,C,d,V,f)$,对于任意一个属性子集 $B\subseteq C$,则在论域 U 上由 B 诱导的不可区分关系 $\mathcal{R}_B$ 定义为

$$\mathcal{R}_B=\{(x,y)\in U^2\mid\forall_{a\in B}(f(x,a)=f(y,a))\} \quad (1)$$

由 $\mathcal{R}_B$ 诱导 U 的划分记为 $U/\mathcal{R}_B=\{E_1,E_2,\cdots,E_s\}$,其中 $E_i=\{x\mid x,y\in U,\forall_{a\in B}(f(x,a)=f(y,a))\}$ 为等价类.

特别地,由决策属性的不可区分关系 $\mathcal{R}_d$ 诱导 U 的划分简记为 $\mathcal{F}$,即 $\mathcal{F}=\{X_1,X_2,\cdots,X_z\}$,其中 $X_j=\{x\mid x\in U,f(x,d)=l_j\}(j=1,2,\cdots,z)$ 是标签为 l_j 的对象集合.

定义 3. 粗糙集[3]. 给定一个决策系统 $DS=(U,C,d,V,f)$,对于任意一个属性子集 $B\subseteq C$,$U/\mathcal{R}_B=\{E_1,E_2,\cdots,E_s\}$,$\mathcal{F}=\{X_1,X_2,\cdots,X_z\}$. $\mathcal{F}$ 关于 $\mathcal{R}_B$ 的上近似集和下近似集分别定义如下:

$$\overline{\mathcal{R}}_B\mathcal{F}=\bigcup_{j=1}^{z}\overline{\mathcal{R}}_BX_j,$$
$$\underline{\mathcal{R}}_B\mathcal{F}=\bigcup_{j=1}^{z}\underline{\mathcal{R}}_BX_j \quad (2)$$

其中 $\overline{\mathcal{R}}_BX_j=\{E_i\mid E_i\cap X_j\neq\varnothing\}$,$\underline{\mathcal{R}}_BX_j=\{E_i\mid E_i\subseteq X_j\}$,$i=1,2,\cdots,s$.

X_j 关于 $\mathcal{R}_B$ 的上近似集和下近似集将论域 U 划分为三个互不相交的区域,即正域 $POS_{\mathcal{R}_B}(X_j)$、边界域 $BND_{\mathcal{R}_B}(X_j)$ 和负域 $NEG_{\mathcal{R}_B}(X_j)$,其中:

$$POS_{\mathcal{R}_B}(X_j)=\underline{\mathcal{R}}_BX_j,$$
$$BND_{\mathcal{R}_B}(X_j)=\overline{\mathcal{R}}_BX_j-\underline{\mathcal{R}}_BX_j,$$
$$NEG_{\mathcal{R}_B}(X_j)=U-\overline{\mathcal{R}}_BX_j.$$

正域中的对象确定地属于 X_j;边界域中的对象具有不确定性,可能属于 X_j,即对象在属性集 B 下具有不一致性;负域中的对象一定不属于 X_j.

2.2 密度峰值聚类

密度峰值聚类(Density Peaks Clustering,DPC)是经典无监督学习算法,能有效计算出类簇的密度中心. 为了更好的理解本文基于监督学习环境提出的算法 1,给出 DPC 局部密度的计算方法.

定义 4. DPC 的局部密度[41]. 给定一个信息系统 $IS=(U,C,V,f)$,对于 $\forall x_i\in U$ 的局部密度 $\rho(x_i)$ 定义为

$$\rho(x_i)=\sum_{x_j\in U}\chi(\Delta(x_i,x_j)-\Delta_c) \quad (3)$$

令 $t=\Delta(x_i,x_j)-\Delta_c$,其中 $\chi(t)=\begin{cases}1, & t<0\\ 0, & t\geq 0\end{cases}$,$\Delta(x_i,x_j)$ 为 x_i 与 x_j 的欧式距离,Δ_c 为截断距离. Δ_c 是为了提算法效率而设计的参数. $\rho(x_i)$ 是计算 U 中的对象与 x_i 之间的距离小于 Δ_c 的对象数目.

3 超区间粒化方法及其分类模型

为了在定量属性的数据空间上有效实现信息粒化,本文 3.1 节定义了区间划分和区间等价关系,以此为基础诱导论域上的划分,每个划分用超区间粒表示;然后,定义超区间粒粗糙集. 3.2 节基于密度峰值聚类算法提出了无参的超区间粒化算法,算法输出的同质超区间粒不再围绕初始中心点对称,并构成了论域的划分. 3.3 节首先给出了超区间粒转换为分类规则的定义;同时为了增加模型的普适性提出了基于超区间粒的自适应近邻分类模型.

3.1 区间划分与超区间粒粗糙集

粗糙集中基于不可区分关系的粒化方法能很好地实现定性属性的论域划分,进而实现数据到知识之间的转换. 除了定性属性,现实中定量属性大量存在,比如医疗领域中的血压、血糖、体温,大气科学中的空气质量分指数 PM2.5、SO_2、NO 等. 不可区分关系要求对象之间的属性值完全相等,面对定量属性的粒化则会失效. 因此,本节首先从一维属性出发定义定量属性的区间划分,并进一步定义多维属性下的区间等价关系,基于此诱导论域上的划分,每个划分用超区间粒表示;然后基于超区间粒定义粗糙集模型.

定义 5. 区间划分. 给定一个决策系统 $DS=(U,C,d,V,f)$,$a\in C$ 是定量属性,则 a 上的区间划分定义为

$$I_a=\{I_a^1,I_a^2,\cdots,I_a^{q_a}\} \quad (4)$$

其中 $I_a^i=[\underline{V}_a^i,\overline{V}_a^i](i=1,2,\cdots,q_a)$ 是 a 的第 i 个区间划分块. I_a^i 的下边界 $\underline{V}_a^i\in V_a$ 和上边界 $\overline{V}_a^i\in V_a$ 满足 $V_a^{\min}=\underline{V}_a^1\leq\overline{V}_a^1<\underline{V}_a^2\leq\overline{V}_a^2<,\cdots,<\underline{V}_a^{q_a}\leq\overline{V}_a^{q_a}=V_a^{\max}$,其中 $V_a^{\min}$ 和 $V_a^{\max}$ 分别是集合 V_a 中的最小值和最大值. $q_a\leq|V_a|$,当 a 的取值连续时,q_a 的取值小于等于 $|U|$,集合 V_a 是 a 的值域如定义 1;$|\cdot|$ 表示集合"·"的势.

从决策的视角来看,连续的属性值表现出的决策结果并不一定连续,比如医学中将连续的血压值划分为两个区间来诊断病人是否患有高血压,并将高血压区间再细分成 3 个区间来诊断高血压病人所属的级别;同样,连续的空气质量分指数通过融合计算得出空气质量指数,并将其分成六个区间来判断空气质量. 在此应用背景下,本文定义区间划分,目的是为了从决策属性的视角,将定量属性的多个实

数或者无数个连续的数值划分为少量的区间，有利于归纳学习和形成可解释的知识.

从定义 5 可看出，确定区间划分块的上下边界是区间划分的关键. 因此，本文在监督学习环境下，考虑属性之间的相关性来计算各个属性相对决策结果表现出的区间特性，具体实现算法见 3.2 节. 本节仅给出了基于区间划分的相关理论.

定义 6. 区间等价关系. 给定一个决策系统 $DS=(U,C,d,V,f)$，对于任意属性子集 $B\subseteq C$，由 B 诱导 U 的区间等价关系 $\mathcal{I}_B$ 定义为

$$\mathcal{I}_B=\{(x,y)\in U^2\mid \forall_{a\in B}(\underline{V}_a^i\le f(x,a)\le\overline{V}_a^i\wedge \underline{V}_a^i\le f(y,a)\le\overline{V}_a^i)\} \tag{5}$$

其中 $[\underline{V}_a^i,\overline{V}_a^i]=I_a^i$，$i=1,2,\cdots,q_a$ 如定义 5. 如果 $(x,y)\in\mathcal{I}_B$，则说明 x 和 y 在 $\forall a\in B$ 下的取值都属于同一个区间，即满足区间等价关系. 当属性 a 下的区间块满足 $\underline{V}_a^i=\overline{V}_a^i$ 时，区间等价关系 $\mathcal{I}_a$ 退化为不可区分关系 $\mathcal{R}_a$. 显然，区间等价关系是不可区分关系的推广，主要应用于定量属性尤其是连续属性的粒化. 如同不可区分关系，区间等价关系满足性质 1.

性质 1. 给定一个决策系统 $DS=(U,C,d,V,f)$，$B\subseteq C$，$\mathcal{I}_B$ 是 U 上由 B 诱导的区间等价关系. 对于 $\forall x,y,z\in U$ 满足以下性质：

(1) 自反性：$\mathcal{I}_B(x,x)=1$；

(2) 对称性：$\mathcal{I}_B(x,y)=\mathcal{I}_B(y,x)$；

(3) 传递性：若 $(x,y)\in\mathcal{I}_B$ 且 $(y,z)\in\mathcal{I}_B$，则 $(x,z)\in\mathcal{I}_B$.

定义 7. 超区间粒. 给定一个决策系统 $DS=(U,C,d,V,f)$，$B\subseteq C$. 设 $\mathcal{I}_B$ 诱导 U 的划分为 $U/\mathcal{I}_B=\{G_B^1,G_B^2,\cdots,G_B^{q_B}\}$，$U/\mathcal{I}_B$ 简记为 $\mathcal{G}_B$，其中 G_B^j $(j=1,2,\cdots,q_B)$ 为 $\mathcal{G}_B$ 中的第 j 个超区间粒且 $G_B^j=\{x\mid x,y\in U,\forall_{a\in B}(\underline{V}_a^i\le f(x,a)\le\overline{V}_a^i\wedge\underline{V}_a^i\le f(y,a)\le\overline{V}_a^i)\}$，其中 $[\underline{V}_a^i,\overline{V}_a^i]=I_a^i(i=1,2,\cdots,q_a)$ 如定义 5.

性质 2. 给定一个决策系统 $DS=(U,C,d,V,f)$，$B\subseteq C$. $\mathcal{I}_B$ 诱导 U 的划分为 $\mathcal{G}_B=\{G_B^1,G_B^2,\cdots,G_B^{q_B}\}$. 对于 $\forall G_B^i,G_B^j\in\mathcal{G}_B$，满足：

(1) $G_B^i\cap G_B^j=\varnothing$ 且 $i\ne j$；

(2) $\bigcup_{i=1}^{q_B}G_B^i=U$；

(3) $G_B^i\ne\varnothing$；

其中 $i,j=1,2,\cdots,q_B$.

定义 8. 超区间粒的标签与纯度. 给定一个决策系统 $DS=(U,C,d,V,f)$，$B\subseteq C$. $\mathcal{I}_B$ 诱导 U 的划分为 $\mathcal{G}_B$，则 $\forall G_B\in\mathcal{G}_B$ 的标签 $L(G_B)$，隶属标签 l 的纯度 $\mu_l(G_B)$，以及在 $\forall a\in B$ 下形成的区间块 $[\underline{V}_a(G_B),\overline{V}_a(G_B)]$ 定义为

$$L(G_B)=\arg\max_{l\in V_d}|\{x\in G_B\mid f(x,d)=l\}|,$$
$$\mu_l(G_B)=\frac{|\{x\in G_B\mid f(x,d)=L(G_B)\}|}{|G_B|} \tag{6}$$
$$\underline{V}_a(G_B)=\min\{f(x,a)\mid x\in G_B\},$$
$$\overline{V}_a(G_B)=\max\{f(x,a)\mid x\in G_B\} \tag{7}$$

由式(6)知，超区间粒的标签与其中大多数对象的标签一致. 另外从式(7)可看出超区间粒是由各个属性下的区间块组合诱导而成，即每个属性任意给定一个区间块即可诱导一个超区间粒. 相反，已知一个属性集 B 下的超区间粒，可计算 $\forall a\in B$ 的一个区间块. 鉴于此，可从不同角度出发提出不同的超区间粒化算法.

定义 9. 超区间粒粗糙集. 给定一个决策系统 $DS=(U,C,d,V,f)$，$B\subseteq C$. $\mathcal{F}=\{X_1,X_2,\cdots,X_z\}$ 是 $\mathcal{R}_d$ 诱导 U 的划分，其中 X_j 是 U 上决策属性值为 l_j 的对象集合，$\mathcal{G}_B=\{G_B^1,G_B^2,\cdots,G_B^{q_B}\}$ 是 $\mathcal{I}_B$ 诱导 U 划分，则 $\mathcal{F}$ 关于 $\mathcal{I}_B$ 的上近似集和下近似集分别定义为

$$\overline{\mathcal{I}}_B\mathcal{F}=\bigcup_{j=1}^{z}\overline{\mathcal{I}}_BX_j,$$
$$\underline{\mathcal{I}}_B\mathcal{F}=\bigcup_{j=1}^{z}\underline{\mathcal{I}}_BX_j \tag{8}$$

其中 $\overline{\mathcal{I}}_BX_j=\{G_B^i\mid G_B^i\cap X_j\ne\varnothing\}$，$\underline{\mathcal{I}}_BX_j=\{G_B^i\mid G_B^i\subseteq X_j\}$，$i=1,2,\cdots,q_B$，$j=1,2,\cdots,z$.

区间等价关系是不可区分关系的推广，因此基于超区间粒的粗糙集模型是粗糙集的推广.

3.2 基于密度峰值的超区间粒化方法

从定义 5 知，计算区间的上下边界是区间划分的关键，也是基于区间等价关系生成超区间粒的基础. 本文以决策系统为应用背景，从决策属性的视角考虑条件属性之间的相关性，提出基于密度峰值聚类的超区间粒化算法(DPCIG).

DPCIG 算法以输出超区间粒为目标，首先基于 DPC 算法按决策类簇计算其中密度最大的对象作为超区间粒的初始中心；进而围绕中心沿着上下两个方向从标签的视角在论域上求解多维定量条件属性形成区间块的上边界和下边界，有效地将多个实数或者无数个连续的数值划分为少量的区间；最后由区间等价关系诱导生成超区间粒，逐次迭代直到超区间粒形成论域上的划分为止. DPCIG 详细设计如算法 1，计算密度中心和从中心点开始外扩生成纯度为 1 的超区间粒是算法的两个主要步骤.

《计算机学报》

算法 1. DPCIG.

输入：$DS=(U,C,d,V,f)$

输出：$\mathcal{G}_C$

1. 初始化超区间粒集合，$\mathcal{G}_C\leftarrow\varnothing$；
2. 初始化对象的粒化标志符 $F_U=\{F_{x_1},F_{x_2},\cdots,F_{x_n}\}$，其中 $F_{x_i}=0,i=1,2,\cdots,n$；
3. $U/\mathcal{R}_d=\{X_1,X_2,\cdots,X_z\}$；
4. WHILE $|X_j|\neq 0$ and $j=1,2,\cdots,z$
5. 按照定义 10 计算决策类 X_j 的密度中心 $\rho(X_j)$，令 $\rho(X_j)=x_v$；
6. 初始化 $G_C\leftarrow\{x_v\}$，初始化 G_C 在 $a_k(k=1,2,\cdots,m)$ 的上下界分别为 $\underline{V}_{a_k}(G_C)\leftarrow f(x_v,a_k)$，$\overline{V}_{a_k}(G_C)\leftarrow f(x_v,a_k)$；
7. $F_{x_v}=1$；//x_v 已经包含于 G_C，则将其标志符置 1
8. 初始化 G_C 每个属性的上下界标志，即 $\overline{F}_{a_k}=0$，$\underline{F}_{a_k}=0$；
9. 计算 $\Delta(x_v,x_i)$ 且 $x_i\neq x_v$，按距离从小到大对 $n-1$ 个对象排序，假设 x_1 为离 x_v 最近的点，x_{n-1} 为离 x_v 最远的点；
10. FOR i from 1 to $n-1$
11. IF $f(x_v,d)=f(x_i,d)$ and $F_{x_i}=0$
12. $G_C\leftarrow G_C\cup\{x_i\}$，$F_{x_i}=1$；//$x_i$ 加入超区间粒，并标记 x_i 已粒化；
13. 如果 $f(x_i,a_k)<\underline{V}_{a_k}(G_C)$，则 $f(x_i,a_k)\rightarrow\underline{V}_{a_k}(G_C)$；
14. 如果 $f(x_i,a_k)>\overline{V}_{a_k}(G_C)$，则 $f(x_i,a_k)\rightarrow\overline{V}_{a_k}(G_C)$；
15. ELSE
16. $\overline{ind}=\{k|\overline{F}_{a_k}=0\}$，$\underline{ind}=\{k|\underline{F}_{a_k}=0\}$；//计算上下界未确定的属性索引集
17. IF $\min\limits_{k\in\underline{ind}}abs(f(x_i,a_k)-\underline{V}_{a_k}(G_C))>\min\limits_{k\in\overline{ind}}abs(f(x_i,a_k)-\overline{V}_{a_k}(G_C))$ //$abs(\psi)$ 为求 ψ 的绝对值
18. $s=\arg\min\limits_{k\in\overline{ind}}abs(f(x_i,a_k)-\overline{V}_{a_k}(G_C))$，$\overline{F}_{a_s}\leftarrow 1$；
19. 如果 $f(x_i,a_s)<\overline{V}_{a_s}(G_C)$，则 $\overline{V}_{a_s}(G_C)\leftarrow f(x_i,a_s)$；
20. ELSE
21. $s=\arg\min\limits_{k\in\underline{ind}}abs(f(x_i,a_k)-\underline{V}_{a_k}(G_C))$，$\underline{F}_{a_s}\leftarrow 1$；
22. 如果 $f(x_i,a_s)>\underline{V}_{a_s}(G_C)$，则 $\underline{V}_{a_s}(G_C)\leftarrow f(x_i,a_s)$；
23. END FOR
24. //检测粒化的对象是否介于超区间粒 G_C 的区间范围
25. FOR $\forall x_i\in G_C$
26. IF $f(x_i,a_k)<\underline{V}_{a_k}(G_C)$ or $f(x_i,a_k)>\overline{V}_{a_k}(G_C)$
27. $G_C\leftarrow G_C-\{x\}$，$F_{x_i}=0$；//x_i 从超区间粒删除，并标记 x_i 未粒化
28. EDN FOR
29. $X_j\leftarrow X_j-G_C$，$\mathcal{G}_B\leftarrow\mathcal{G}_B\cup G_C$；
30. END WHILE
31. RETURN $\mathcal{G}_C$

DPCIG 的步骤 1 对应算法 1 的 1～5 行，主要是按照如下定义计算决策类簇 X_j' 的密度中心 $\rho(X_j')$.

定义 10. 决策类的密度中心. 给定一个决策系统 $DS=(U,C,d,V,f)$，$\mathcal{R}_d$ 诱导 U 的划分为 $\mathcal{F}=\{X_1,X_2,\cdots,X_z\}$，$X_j'\subseteq X_j(j=1,2,\cdots,z)$. 对于 $\forall X_j'$，X_j' 的密度中心为

$$\rho(X_j')=\arg\max_{x_i\in X_j'}(\rho(x_i)) \tag{9}$$

其中 $\rho(x_i)=\sum\limits_{x_i,x_k\in X_j'}\chi(\Delta(x_i,x_k)-\Delta_c)$，$\chi$ 和 Δ_c 如式(3)所示. 在归一化数据应用场景中，Δ_c 通常取值 0.18，本文与常规取值一致. 注意 $\rho(X_j')$ 是 X_j' 中密度最大的对象并非虚拟点.

DPCIG 算法基于 DPC 算法计算超区间粒的中心主要是为了在粒计算的框架下能充分结合“大范围首先”处理机制来建立粒化模型，实现算法输出的同质超区间粒在构成论域划分的同时尽可能使粒的个数少，避免了粒球计算在粒化过程中随机选择粒化中心，以及多次迭代修正中心的问题. 假如随机选择密度中心，若是要保证算法输出的同质超区间粒能构成论域的划分，则算法输出的超区间粒的个数会增加，算法的性能也会随之降低.

DPCIG 算法的步骤 2 是将决策系统映射到欧式空间，以 $\rho(X_j')$ 为初始粒化中心按照不同维度不同方向分别确定各个属性的区间划分，进而构造纯度为 1 的超区间粒 G_C，其主要流程图如图 3 所示，对应算法 1 的 6～28 行.

算法 1 第 2 行是初始化对象的粒化标志符，$F_{x_i}=0$ 表示对象 x_i 未粒化，即 x_i 不属于任何一个超区间粒，$F_{x_i}=1$ 表示对象 x_i 已被粒化. 第 6 行是对 G_C 中包含的对象和 G_C 在各属性下的上下界初始化，其值分别为当前决策类的密度中心和其的属性值. 第 8 行 $\overline{F}_{a_k}=0$，$\underline{F}_{a_k}=0$ 分别表示 G_C 在 a_k 维度上以 $f(x_v,a_k)$ 为初始中心的两个上下界都还未确定. 反之，当 $\overline{F}_{a_k}=1$，$\underline{F}_{a_k}=1$ 表示 G_C 在 a_k 维度上的上下界已确定. 注意 G_C 在各个属性的上界与下界并非同时确定，且最终形成的区间不再围绕初始粒化中心点对称. 第 10～23 行是按与中心点的距离从小到大遍历论域上的对象，以此确定 G_C 在每个属性上的上界和下界.

第 11～14 行表示如果 x_i 与中心点的标签相同且 $F_{x_i}=0$，则更新 G_C 中所包含的对象；同时计算 x_i 的属性值与 G_C 在对应属性的上下界的大小关系，进而判断是否更新 G_C 的上界或下界. 第 13 行是按属性分别判断 x_i 的属性值是否小于当前 G_C 的下界 $\underline{V}_{a_k}(G_C)$，若成立则更新 $\underline{V}_{a_k}(G_C)$；第 14 行是更新 G_C 的上界. 显然更新的过程是 G_C 粒化的过程，即 G_C 所包含的对象逐渐增多，区间逐渐变大.

第 16～22 行与 11～14 行构成二分支选择，表示

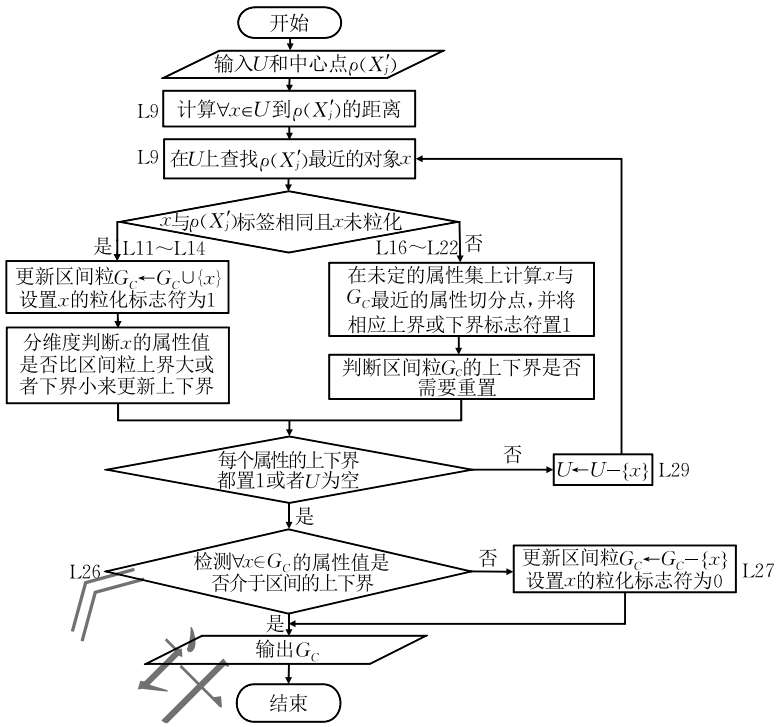


图 3　已知密度中心计算超区间粒的核心流程

当 G_C 在粒化时遍历到已粒化的 x_i 或者与粒化初始中心异类的对象 x_i 时，则计算 x_i 与 G_C 最近的某个属性的上界或下界，并将停止 G_C 在该属性上界或下界的更新. 16 行首先计算 G_C 的上下界还未确定的属性索引集 $\overline{ind}$ 和 $\underline{ind}$；然后在 $\overline{ind}$ 和 $\underline{ind}$ 上查找 x_i 与 G_C 最近的上下界，并将相应属性上界或下界标志符置为 1. 第 17～19 行表示当 $f(x_i,a_k)$ 离 $\overline{V}_{a_s}(G_C)$ 距离最近，则停止 G_C 在 a_s 区间上界的更新即 $\overline{F}_{a_s}$ 置 1，同时判断 $f(x_i,a_s)$ 是否小 $\overline{V}_{a_s}(G_C)$，若是则将 $\overline{V}_{a_s}(G_C)$ 重置. 第 20～22 行与第 17～19 行构成二分支选择，即当 $f(x_i,a_k)$ 离 $\underline{V}_{a_s}(G_C)$ 近，则停止 G_C 在 a_s 区间下界的更新，同理判断是否需要重置 $\underline{V}_{a_s}(G_C)$.

第 24～28 行检测已合并到 G_C 的对象 x_i 的属性值是否介于 G_C 的区间范围，若 x_i 不满足条件则将 x_i 从 G_C 中删除并将 x_i 粒化标志符重置 0. 重检是因为 19 行和 22 行可能对区间的上下界重置.

为了理解 DPCIG 算法，举例重点说明超区间粒的形成过程. 假设一个数据集有两个 2 条件属性 a_1 和 a_2、37 个对象，将其映射到欧式空间下如图 4 所示，圆形和正方形分别表示 l_1 和 l_2 两个类的对象. DPCIG 算法首先根据定义 10 计算出 x_1^1 和 x_2^1 分别为 l_1 和 l_2 的密度中心. 接着，按照 DPCIG 算法 6～28 行以 x_1^1 为粒化中心开始生成超区间粒 G_C^1. 在二维欧式空间，超区间粒是由两个属性的区间段构成的矩形. 初始状态下，G_C^1 在 a_1 和 a_2 上的上下界相等，分别等于 x_1^1 在 a_1 和 a_2 上的值. 接着更新 G_C^1 的上下界，最终 G_C^1 的输出为图 4 实线矩形，其内部最小虚线矩形为第一次的更新，第二个矩形为第二次的更新.

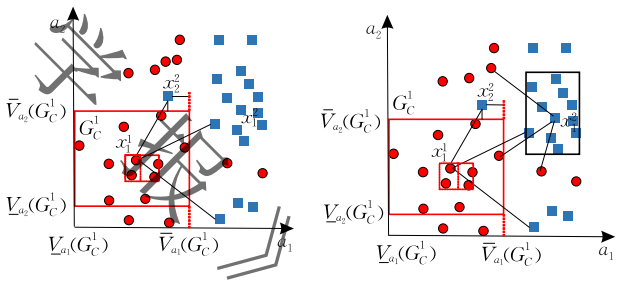


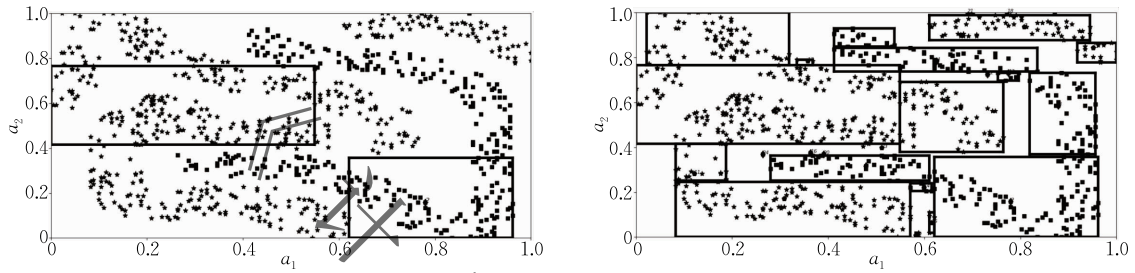
图 4　超区间粒的形成过程

生成 G_C^1 的具体方法如下：(1) 计算剩余 36 个对象与 x_1^1 的距离并按距离从小到排序；(2) 按照排序依此判断对象是否与 x_1^1 标签相同，若对象标签相同且未被粒化，则分维度判断对象的属性值是否比区间粒上界大或者下界小来更新上下界，使得 G_C^1 逐渐变大；当遇到异类对象或者已经粒化后的对象时，按属性找到该对象与 G_C^1 最近未确定的区间端点，并停止该端点的更新. 如当 G_C^1 遍历到异类点 x_2^2 时，由于 x_2^2 在 a_2 上的值与 G_C^1 在 a_2 的上界(表示为 $\overline{V}_{a_2}(G_C^1)$)最近，则停止 $\overline{V}_{a_2}(G_C^1)$ 的更新，即 G_C^1 在 a_2 的上界确定，同时只要对象在 a_2 上的取值大于 $\overline{V}_{a_2}(G_C^1)$ 的对象(即 x_2^2 上方的点)则不用再判断是否与中心点同类. 同理，依此确定 $\overline{V}_{a_1}(G_C^1)$、$\underline{V}_{a_2}(G_C^1)$、$\underline{V}_{a_1}(G_C^1)$ 形成 G_C^1.

假设决策系统 DS 有 n 个对象，标签最多的对象

个数为 n_1($n_1<n$). 给定一个 DS,DPCIG 算法 4～30 行通过迭代上述两个步骤输出超区间粒集合 $\mathcal{G}_C$, $\mathcal{G}_C$ 的势 $|\mathcal{G}_C|<n_1$. 算法的迭代次数 $t=|\mathcal{G}_C|$,即 $t<n_1$. 由算法 29 行知已经粒化的对象不再参与下一次迭代的计算,即随着迭代次数的增加,两个步骤的时间复杂度一定逐渐减少. 第一次迭代过程中,步骤 1 是计算 n_1 个对象之间的欧式距离得出 1 个密度中心点,时间复杂度为 $O(n_1^2)$;步骤 2 是计算 1 个粒化中心到 n 个对象的距离,时间复杂度为 $O(n)$. 综上所述,DPCIG 算法的时间复杂为 $O(n_1^2)$. 从超区粒的表示来看,t 与数据集的分布相关,更利于快速实现方形、点簇形以及月牙形数据集的粒化,面对环形的数据集,算法的迭代次数可能会增加.

为了直观说明 DPCIG 算法的迭代次数远小于数据集中对象的个数. 算法将 fourclass 数据集($n=863$, $n_1=556$)的输出结果可视化,如图 5 所示. 图中的矩形代表超区间粒,超区间粒的个数随着迭代次数逐个增加. 图 5(b)是 DPCIG 算法在经过 24 次迭代后生成的覆盖论域的 24 个超区间粒,显然 $t<n_1$. 相比粒球粒化算法的输出如图 1(b),可以看出 DPCIG 算法输出的超区间粒没有损失对象信息,满足性质 2.



(a) 第2次迭代 (b) 第24次迭代

图 5 DPCIG 算法粒化 fourclass 数据集

本文针对监督学习任务,在粒计算的框架下,通过区间划分的粒化策略,将数据集合转换为超区间粒集合,从属性值和对象两个角度实现了数据约简. 图 5 直观地说明了粒化算法 DPCIG 将 863 个样本简化为 24 个超区间粒的过程.

3.3 基于超区间粒的自适应近邻分类模型

规则学习是粗糙集中经典的分类模型. 为了验证 DPCIG 算法的有效性,本节将其用于分类任务. 首先给出超区间粒转换为分类规则的定义,然后为了解决新的测试对象不能与规则完全匹配的问题,基于超区间粒提出自适应近邻分类模型.

定义 11. 给定一个决策系统 $DS=(U,C,d,V,f)$, $C=\{a_1,a_2,\cdots,a_m\}$. 设 $\mathcal{G}_C=\{G_C^1,G_C^2,\cdots,G_C^{p_C}\}$ 是 $\mathcal{I}_C$ 诱导 U 的划分. 对于 $\forall x\in U$,若 x 在任意 a_k 下的取值满足 $\underline{V}_{a_k}(G_C^i)\leq f(x,a_k)\leq\overline{V}_{a_k}(G_C^i)$ ($i=1,2,\cdots,p_C$; $k=1,2,\cdots,m$),则 x 的标签与 G_C^i 相同,即 $f(x,d)=L(G_C^i)$.

从定义 11 可知,超区间粒的集合等价于一个规则集,一个超区间粒对应一条规则. 基于超区间粒构建的规则能覆盖论域,但面对新的测试对象,完全可能存在测试对象不能与规则匹配的情况. 鉴于此,本文受 kNN 算法的启发,提出基于超区间粒的自适应近邻分类模型(IGANN),如算法 2 所示.

算法 2. IGANN.

输入: $DS=(U',C,d,V,f)$,其中 $U'=\{x_1^*,x_2^*,\cdots,x_{n'}^*\}$ 为测试集;算法 1 的输出 $\mathcal{G}_C=\{G_C^1,G_C^2,\cdots,G_C^p\}$

输出: 测试集的标签集 $L(U')$

1. 初始化 $L(U')\leftarrow\varnothing$;
2. FOR i from 1 to n'
3. 按照定义 12 计算 x_i^* 到 G_C^j ($j=1,2,\cdots,p$) 的距离,并按距离从小到大对 $\mathcal{G}_C$ 中的粒排序,设 G_C^1, G_C^2,…, G_C^p 为排序结果;
4. IF $L(G_C^1)=L(G_C^2)$
5. $L(x_i^*)=L(G_C^1)$; //$L(x_i^*)$表示 x_i^* 的标签,即 $f(x_i^*,d)=L(x_i^*)$
6. ELSE
7. $T\leftarrow\{L(G_C^1),L(G_C^2)\}$, $a\leftarrow0.5$, $t\leftarrow2$;
8. WHILE $a\leq1/|T|$
9. $t\leftarrow t+1$; $T\leftarrow T\cup L(G_C^t)$;
10. $l'=\arg\max_{l\in T}|\{L(G_C^j)=l\}|$,其中 $j=1,2,\cdots,t$;
11. $a=|\{L(G_C^j)\in T|L(G_C^j)=l'\}|/|T|$;
12. $L(x_i^*)=l'$;
13. $L(U')\leftarrow L(U')\cup L(x_i^*)$;
14. END FOR
15. RETURN $L(U')$

算法 2 第 3 行首先按照定义 12 计算测试对象与超区间粒之间的距离. 若最近的两个超区间粒标签相同,则测试对象的标签确定,见 4～5 行;否则依次从 $\mathcal{G}_C$ 取出近邻超区间粒的标签,直到某个标签的

票数大于其它标签的票数为止，见 6～13 行.

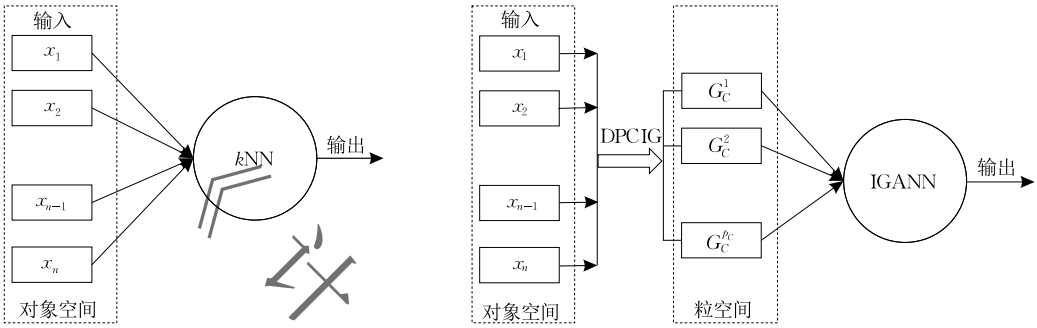
定义 12. 给定一个决策系统 $DS=(U,C,d,V,f)$，$C=\{a_1,a_2,\cdots,a_m\}$，设 G^C 是 U 上的一个超区间粒. 对于 $\forall x\in U$，x 到 G_C 的距离定义为

$$d(x,G_C)=\sum_{k=1}^{m}d_{a_k}(x,G_C) \tag{10}$$

其中 $d_{a_k}(x,G_C)=\begin{cases}0, & \underline{V}_{a_k}(G_C)\le f(x,a_k)\le\overline{V}_{a_k}(G_C)\\ \delta, & \text{其他}\end{cases}$，且 $\delta=\min((f(x,a_k)-\overline{V}_{a_k}(G_C))^2,(f(x,a_k)-\underline{V}_{a_k}(G_C))^2)$，$\min(\delta_1,\delta_2)$ 表示在 δ_1 和 δ_2 之间取小. 由式(10)知，属于超区间粒的对象到对应超区间粒的距离为 0.

假设 DPCIG 算法输出的超区间粒个数为 $|\mathcal{G}_C|$，且 $|\mathcal{G}_C|<n_1$，则 IGANN 分类算法的时间复杂度为 $O(|\mathcal{G}_C|)$.

相比 kNN，IGANN 有两个区别：(1) IGANN 是在超区间粒空间上建模，如图 6. 简单说，在查找测试对象的近邻时，IGANN 计算对象与超区间粒的距离，而 kNN 计算对象间的距离；(2) IGANN 是无参自适应的分类模型.



(a) kNN分类器　　(b) IGANN分类器

图 6　IGANN 与 kNN 的比较

4　实验对比与分析

本文实验在 21 个标准 UCI 数据集上进行，其中包含了对比文献[35]中的 13 个数据集. 实验数据的详细信息如表 2 所示，来源于公开的 UCI 数据集学习库(https://archive.ics.uci.edu/ml/index.php). 实验运行环境为 Windows 10 64 位操作系统，8GB 内存和 Intel(R) Core(TM) i7-8565U CPU@1.80 GHz 处理器，编程语言为 Python.

表 2　21 个 UCI 数据集的描述

编号	数据集	对象数目	属性数目	类别数目	属性类型
No. 1	iris	151	4	3	实数
No. 2	hepatitis(hep)	155	19	2	整数，实数
No. 3	heart1	293	13	2	整数
No. 4	arrhythmia(arr)	453	279	13	整数，实数
No. 5	band	539	19	2	整数，实数
No. 6	energy-y1(y1)	768	8	3	整数，实数
No. 7	fourclass(four)	863	2	2	实数
No. 8	abalone(aba)	4177	8	3	实数，名义
No. 9	pen	10992	16	10	整数
No. 10	drybean(dry)	13611	17	7	实数
No. 11	htru2	17898	8	2	实数
No. 12	occupation(occ)	20560	7	2	实数
No. 13	credit	690	15	2	整数，实数
No. 14	anneal	798	38	5	整数，实数
No. 15	german(ger)	1000	19	2	整数
No. 16	heart2	303	13	5	整数，实数
No. 17	wdbc	569	30	2	实数
No. 18	hourse	368	23	2	整数，实数
No. 19	iono	351	34	2	整数，实数
No. 20	wine	178	13	3	整数，实数
No. 21	elect	10000	13	2	实数

为了验证 IGANN 分类模型的有效性，采用评估分类性能的精确率(*Precision*)、准确率(*Accuracy*)、召回率(*Recall*)以及调和平均 *F*1 四个通用指标[31]测试其性能，同时采用标准差(STD)来度量分类器的稳定性，并与基于粒计算的四个先进分类模型：GBNRS[35]、NCRDPP[42]、NCR[43] 和 RSC[44] 以及四个经典的机器学习分类模型：高斯朴素贝叶斯(GNBC)、kNN($k=3$)、支持向量机(SVM)、决策树(CART)进行分类性能对比.

本文分类模型 IGANN 和对比模型的参数说明如下. IGANN 的是一个无参模型，IGANN 的输入涉及超区间粒化算法 DPCIG，DPCIG 算法在按定义 10 选取粒化中心时所涉及的截距采用常用值 0.18. 另外，由于 DPCIG 算法是对定量属性的数据集实现信息粒化，若数据集包含二值定性属性则将其转换为 0 和 1，比如 abalone 中包含 1 个二值定性取值为“是”和“否”，我们将其转换为 0 和 1 后与归一化的定性属性一起映射到欧式空间作为 DPCIG 算法的输入.

GBNRS 算法中生成粒球的纯度阈值设置 1，分类准则采用最近邻分类机制. NCRDPP 和 NCR 的相关参数与文献一致. 由于 RSC 只适用于定性数据类型，实验中数据集的属性若是整数或者实数类型则采用等宽离散策略对属性进行离散化且宽度设置为

计算机学报

5,分类准则是多数投票机制.在 CART 中,属性重要度采用基尼指数衡量,树的深度设置为 3.GNBC 和 SVM 两个模型的所有参数信息利用 scikit-learn 学习库提供的参数值 default setting.

对于9个分类器,在21个数据集上关于 *Accuracy*、*Precision* 和 *Recall* 的实验结果如表 3～表 5 所示,*F*1 的实验结果如图 7 所示,在四个指标下的平均性能如图 8 所示.表 3～表 5 中第 1 行是分类器,第 1 列是数据集,单元格中的数据代表不同分类器在不同数据集上的分类性能.性能指标取值范围为 0 到 1,值越大分类器性能越好.表中的 STD 为标准差,STD 越小说明分类器性能越稳定.

表 3 分类器在准确率(*Accuracy*)指标下的性能对比

数据集	IGANN	基于粒计算的经典分类模型				基于机器学习的经典分类模型			
		GBNRS	NCRDPP	NCR	RSC	*k*NN	CART	GNBC	SVM
iris	0.95	0.94	0.93	0.94	0.79	0.94	0.95	0.95	0.95
hep	0.88	0.93	0.87	0.84	0.88	0.89	0.86	0.31	0.90
heart1	0.82	0.73	0.80	0.80	0.70	0.78	0.75	0.82	0.66
arr	0.63	0.54	0.62	0.62	0.61	0.59	0.65	0.23	0.63
band	0.70	0.72	0.70	0.72	0.71	0.65	0.66	0.67	0.67
y1	0.94	0.85	0.82	0.86	0.94	0.78	0.97	0.81	0.81
four	0.97	0.99	0.98	0.99	0.95	1.00	0.76	0.75	0.80
aba	0.62	0.60	0.50	0.68	0.50	0.59	0.61	0.59	0.63
pen	0.99	0.97	0.75	0.98	0.70	0.99	0.57	0.86	0.99
dry	0.91	0.82	0.88	0.57	0.53	0.91	0.91	0.90	0.62
htru2	0.98	0.97	0.96	0.95	0.95	0.98	0.97	0.95	0.91
occ	0.99	0.98	0.98	0.98	0.97	0.99	0.99	0.97	0.99
credit	0.99	0.88	0.99	0.82	0.87	0.96	0.99	0.99	0.96
anneal	0.89	0.82	0.86	0.84	0.84	0.89	0.87	0.59	0.83
ger	0.65	0.62	0.69	0.59	0.60	0.64	0.66	0.69	0.60
heart2	0.58	0.56	0.58	0.41	0.54	0.56	0.57	0.55	0.54
wdbc	0.92	0.88	0.95	0.91	0.90	0.97	0.93	0.94	0.91
hourse	0.79	0.77	0.69	0.61	0.64	0.81	0.83	0.76	0.66
iono	0.91	0.86	0.91	0.91	0.85	0.85	0.90	0.89	0.93
wine	0.90	0.82	0.94	0.76	0.80	0.96	0.96	0.97	0.46
elect	0.84	0.92	0.86	0.84	0.62	0.92	1.00	0.98	0.92
平均	**0.85**	0.82	0.82	0.79	0.76	0.84	0.83	0.77	0.78
STD	**0.13**	0.14	0.14	0.16	0.15	0.15	0.15	0.21	0.16

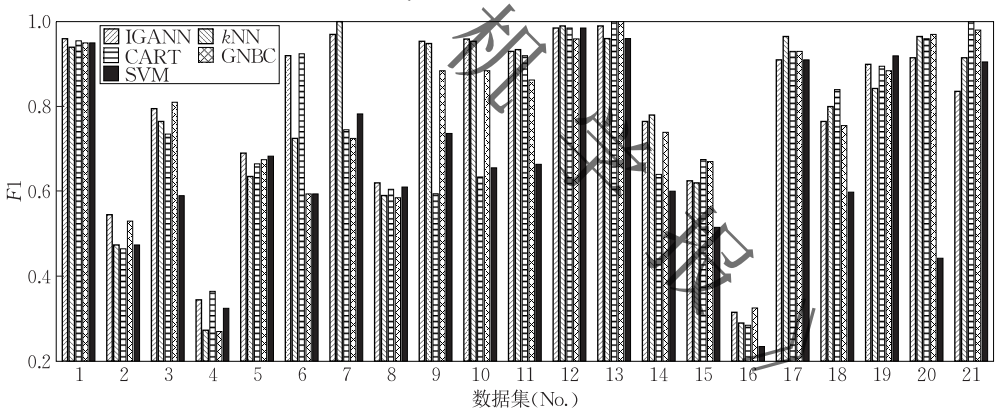
表 4 分类器在精确率(*Precision*)指标下的性能对比

数据集	IGANN	基于粒计算的经典分类模型				基于机器学习的经典分类模型			
		GBNRS	NCRDPP	NCR	RSC	*k*NN	CART	GNBC	SVM
iris	0.96	0.94	0.95	0.93	0.79	0.93	0.95	0.95	0.95
hep	0.55	0.47	0.60	0.51	0.45	0.45	0.45	0.52	0.45
heart1	0.80	0.73	0.70	0.82	0.80	0.77	0.74	0.81	0.65
arr	0.36	0.36	0.35	0.34	0.34	0.30	0.35	0.26	0.34
band	0.69	0.72	0.70	0.73	0.77	0.64	0.67	0.74	0.76
y1	0.93	0.78	0.73	0.80	0.94	0.74	0.95	0.54	0.54
four	0.97	0.99	0.98	0.99	0.95	1.00	0.81	0.73	0.83
aba	0.62	0.60	0.49	0.69	0.65	0.59	0.60	0.58	0.61
pen	0.99	0.97	0.75	0.98	0.80	0.99	0.48	0.86	0.98
dry	0.93	0.84	0.88	0.55	0.53	0.92	0.73	0.91	0.49
htru2	0.94	0.98	0.96	0.95	0.94	0.96	0.93	0.82	0.95
occ	0.98	0.97	0.98	0.98	0.97	0.99	0.99	0.94	0.98
credit	0.99	0.88	0.99	0.82	0.87	0.96	1.00	1.00	0.96
anneal	0.78	0.72	0.69	0.71	0.40	0.79	0.64	0.66	0.73
ger	0.63	0.61	0.68	0.59	0.62	0.63	0.68	0.67	0.51
heart2	0.32	0.32	0.29	0.26	0.30	0.28	0.27	0.32	0.25
wdbc	0.90	0.87	0.96	0.91	0.91	0.97	0.93	0.93	0.91
hourse	0.78	0.69	0.69	0.58	0.63	0.80	0.82	0.75	0.67
iono	0.92	0.88	0.93	0.92	0.84	0.89	0.90	0.91	0.94
wine	0.93	0.89	0.94	0.75	0.83	0.96	0.97	0.97	0.48
elect	0.90	0.92	0.85	0.83	0.59	0.92	1.00	0.98	0.91
平均	**0.80**	0.77	0.77	0.74	0.71	0.78	0.76	0.75	0.71
STD	**0.20**	**0.20**	**0.20**	0.21	0.21	0.22	0.22	0.21	0.23

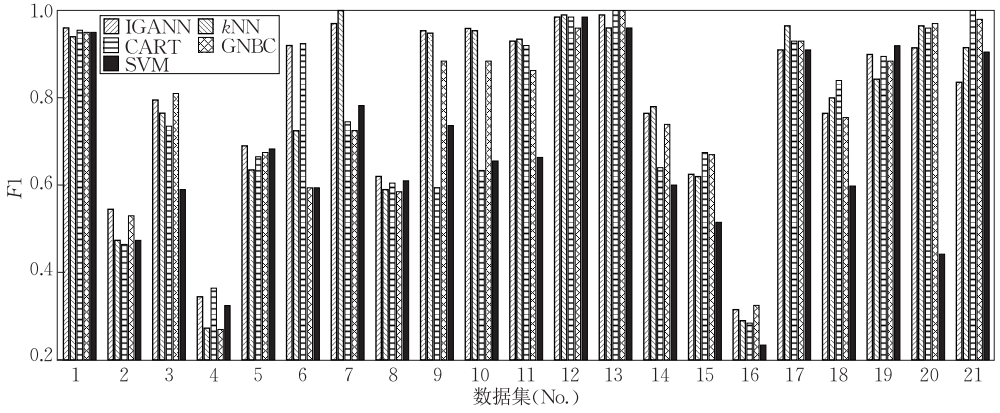
《计算机学报》

表 5　分类器在召回率(*Recall*)指标下的性能对比

数据集	IGANN	基于粒计算的经典分类模型				基于机器学习的经典分类模型			
		GBNRS	NCRDPP	NCR	RSC	*k*NN	CART	GNBC	SVM
iris	0.96	0.84	0.95	0.95	0.8	0.95	0.96	0.95	0.95
hep	0.54	0.50	0.64	0.53	0.49	0.50	0.48	0.54	0.50
heart1	0.79	0.73	0.75	0.77	0.77	0.76	0.73	0.81	0.54
arr	0.33	0.35	0.32	0.30	0.29	0.25	0.38	0.28	0.31
band	0.69	0.72	0.70	0.72	0.65	0.63	0.66	0.62	0.62
y1	0.91	0.82	0.92	0.80	0.91	0.71	0.90	0.66	0.66
four	0.97	0.99	0.98	0.99	0.95	1.00	0.69	0.72	0.74
aba	0.62	0.60	0.50	0.68	0.53	0.59	0.61	0.59	0.61
pen	0.92	0.84	0.88	0.57	0.60	0.91	0.78	0.91	0.59
dry	0.99	0.96	0.75	0.98	0.73	0.99	0.56	0.86	0.99
htru2	0.92	0.98	0.96	0.95	0.91	0.91	0.91	0.91	0.51
occ	0.99	0.98	0.98	0.99	0.97	0.99	0.98	0.98	0.99
credit	0.99	0.88	0.99	0.82	0.87	0.96	1.00	1.00	0.96
anneal	0.75	0.71	0.68	0.65	0.44	0.77	0.64	0.84	0.51
ger	0.62	0.59	0.68	0.57	0.59	0.61	0.67	0.67	0.52
heart2	0.31	0.29	0.23	0.23	0.29	0.30	0.30	0.33	0.22
wdbc	0.92	0.87	0.95	0.90	0.87	0.96	0.93	0.93	0.91
hourse	0.75	0.59	0.68	0.58	0.63	0.80	0.81	0.76	0.54
iono	0.88	0.82	0.87	0.88	0.82	0.80	0.89	0.86	0.90
wine	0.90	0.87	0.96	0.77	0.81	0.97	0.95	0.97	0.41
elect	0.78	0.92	0.83	0.82	0.59	0.91	1.00	0.98	0.90
平均	**0.80**	0.75	0.77	0.73	0.70	0.79	0.77	0.79	0.67
STD	**0.20**	**0.20**	0.21	0.21	**0.20**	0.22	**0.20**	**0.20**	0.23

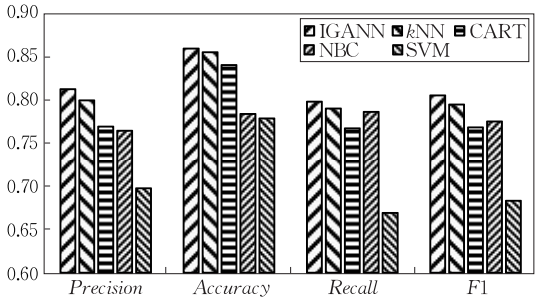


(a) IGANN与基于粒计算的分类器性能对比

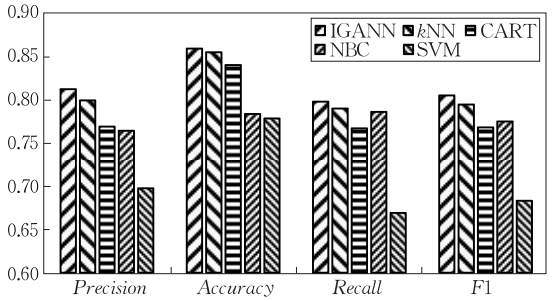


(b) IGANN与基于机器学习的经典分类器性能对比

图 7　分类器在 *F*1 指标下的性能对比



(a) IGANN与基于粒计算的分类模型对比



(b) IGANN与基于机器学习的分类模型对比

图 8 在 4 个指标下模型性能的均值对比

从表 3～表 5 和图 7 可以看出在大部分数据集上,IGANN 相比其它 8 个分类器在性能上有着明显优势.从平均性能来看 IGANN 的 4 项指标都高于其它 8 个分类器,如图 8 所示.从表 3～表 5 可看出,IGANN 的 STD 在 3 指标下都不高于其它 8 个模型.实验结果表明,IGANN 在保证平均性能高于其它分类器的情况下,同时保证了模型的稳定性.

DPCIG 主要是针对实数属性实现数据粒化.因此,从属性取值角度看,本文的分类模型 IGANN 更适合包含实数属性的数据集.从表 3 可看出,在 iris、dry、htru2、occ 实数数据集中,IGANN 相比其它模型占有优势,尤其是相比只能处理定性属性的分类模型 RSC 具有显著优势.但在混合数据集比如 band、german 数据集中则不具有明显的优势.

另外,DPCIG 算法输出的信息粒是由超区粒表示.比如,在二维欧式空间,超区间粒是由两个属性的区间段构成的矩形;在三维空间,超区间粒是立方体.因此,从信息粒的表示来看,IGANN 更利于学习条状、簇状以及月牙形数据分布的分类任务.

在基于机器学习的 4 个经典分类模型中,kNN 在 4 项指标下的分类性能相对最优;在基于粒计算的 4 个经典分类模型中,GBNRS 在 4 项指标下的分类性能相对最优.相比 GBNRS,IGANN 在 *Accuracy* 和 *Precision* 上平均比其高 3%,在 *Recall* 上平均比其高 5%.相比 kNN,IGANN 在 *Accuracy* 和 *Recall* 上平均性能比其高 1%,在 *Precision* 上平均性能比其高 2%.另外,IGANN 在每个指标下的稳定性都比 kNN 高.

从分类准则来看,IGANN、kNN 和 GBNRS 均是基于距离的分类模型,都用欧氏距离度量距离.在平均性能和多数数据集上,IGANN 性能达到最优的原因主要有以下 3 个:

(1) DPCIG 输出的超区间粒即是论域上的覆盖也划分,而 IGANN 是在 DPCIG 输出的超区间粒集合上建立分类规则.GBNRS 是在粒球集合上建立分类规则.粒计算输出的粒球不能构成论域的覆盖.相比 GBNRS,IGANN 显著减少了信息损失,更利于规则学习.从表 3～表 5 可看出,IGANN 在多数数据集上优于 GBNRS.另外,GBNRS 中的信息粒是用超球表示,而 IGANN 中的信息粒是用超区间表示.从信息粒的表示来看,GBNRS 相比 IGANN 更利于学习球状的数据分布,因此存在 IGANN 比 GBNRS 性能低的情况,如 elect 数据集.

(2) IGANN 和 GBNRS 都是在信息粒的空间上建模,而 kNN 是在对象空间上建模.因此,相比 kNN,IGANN 和 GBNRS 具有更强的稳定性.如表 3～表 5 中的 STD 所示,IGANN 和 GBNRS 在三个指标下的 STD 都低于 kNN.另外,由于信息粒是粒化的结果,是数据约简的表示形式.因此,相比 kNN,IGANN 和 GBNRS 都存在一定的信息损失,即不能绝对保证 IGANN 和 GBNRS 的性能一定优于 kNN.

(3) 为了增强 IGANN 在多分类数据集上的分类性能,IGANN 在粒化过程中基于“大范围首先”处理机制按决策类依次产生信息粒.这样保证了各个类的信息粒交替产生,从而使得信息粒在各个类之间保持均衡.因此,相比 kNN 和 GBNRS,IGANN 在 iris、arrhythmia、pen 以及 drybean 多分类数据集上的分类准确率和精确率具有优势.

5 结 论

作为解决大规模复杂问题的新范式,粒计算已成功用于数据挖掘.粒化是粒计算的核心任务,信息粒是数据粒化后的基本单元.粒球计算虽然效率高,但存在如下三个问题:(1) 粒球计算通过粒球表示信息粒会损失一定的边界信息;(2) 粒化过程中,类簇中心点的确定不够准确;(3) 粒球在每个属性维度下的表示都严格围绕中心对称,易受边界对象的

影响.因此,本文首先定义多维属性下的区间等价关系来诱导论域的划分,每个划分单元用超区间粒表示,克服了用粒球表示信息粒存在信息损失的问题.然后,基于密度峰值算法提出无参的超区间粒化算法,算法不仅在粒化过程中能准确计算类簇的密度中心,而且输出的同质超区间粒不再围绕初始中心点对称,并构成了论域的划分.最后,实现了超区间粒到分类规则的转换;基于超区间粒提出了自适应近邻分类模型,通过对比实验在 4 个分类指标下验证了分类模型的性能.在未来研究中,我们将研究如何在超区间粒化算法的框架下实现属性约简,期望同时实现数据的纵向约简和横向约简;另外,在保证粒化性能不下降的基础上我们还将进一步研究如何降低超区间粒化算法的时间复杂度.

参 考 文 献

[1] Liang Ji-Ye, Qian Yu-Hua, Li De-Yu, et al. Theory and method of granular computing for big data mining. Scientia Sinica Informationis, 2015, 45(11): 1355-1369(in Chinese)
(梁吉业，钱宇华，李德玉等. 大数据挖掘的粒计算理论与方法. 中国科学：信息科学，2015，45(11)：1355-1369)

[2] Zadeh L. Fuzzy sets and information granularity. Advances in Fuzzy Set Theory and Applications, 1979, 11: 3-18

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

[4] Hobbs J. Granularity//Proceedings of the Ninth International Joint Conference on Artificial Intelligence. Los Angeles, USA, 1985: 432-435

[5] Zhang B, Zhang L. Theory and Applications of Problem Solving. Amsterdam: North-Holland, 1992

[6] Li De-Yi, Meng Hai-Jun, Shi Xue-Mei. Membership clouds and membership cloud generators. Journal of Computer Research and Development, 1995, 32(6): 15-20(in Chinese)
(李德毅，孟海军，史雪梅. 隶属云和隶属云发生器. 计算机研究与发展，1995，32(6)：15-20)

[7] Wang Guo-Yin, Fu Shun, Yang Jie, et al. A review of research on multi-granularity cognition based intelligent computing. Chinese Journal of Computers, 2022, 45(6): 1161-1175(in Chinese)
(王国胤，傅顺，杨洁等. 基于多粒度认知的智能计算研究. 计算机学报，2022，45(6)：1161-1175)

[8] Huang Z, Li J, Qian Y. Noise-tolerant fuzzy-β-covering-based multigranulation rough sets and feature subset selection. IEEE Transactions on Fuzzy Systems, 2022, 30(7): 2721-2735

[9] Chen H, Li T, Ruan D, et al. A rough-set-based incremental approach for updating approximations under dynamic maintenance environments. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(2): 274-284

[10] Zhang Q, Yang C, Wang G. A sequential three-way decision model with intuitionistic fuzzy numbers. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2019, 51(5): 2640-2652

[11] Dai J, Zou X, Wu W. Novel fuzzy β-covering rough set models and their applications. Information Sciences, 2022, 608: 286-312

[12] Zhou S, Li D, Zhang Z, et al. A new membership scaling fuzzy c-means clustering algorithm. IEEE Transactions on Fuzzy Systems, 2020, 29(9): 2810-2818

[13] Pedrycz W, Bargiela A. An optimization of allocation of information granularity in the interpretation of data structures: Toward granular fuzzy clustering. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 2012, 42(3): 582-590

[14] Wang Guo-Yin, Zhang Qing-Hua, Hu Jun. An overview of granular computing. CAAI Transactions on Intelligent Systems, 2007, 2(6): 8-26(in Chinese)
(王国胤，张清华，胡军. 粒计算研究综述. 智能系统学报，2007，2(6)：8-26)

[15] Wang W, Liu W, Chen H. Information granules-based BP neural network for long-term prediction of time series. IEEE Transactions on Fuzzy Systems, 2021, 29(10): 2975-2987

[16] Zhu X, Pedrycz W, Li Z. Construction and evaluation of information granules: From the perspective of clustering. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(3): 2024-2037

[17] Hu Sheng-Dan, Miao Duo-Qian, Yao Yi-Yu. Three-way label propagation based semi-supervised attribute reduction. Chinese Journal of Computers, 2021, 44(11): 2332-2343(in Chinese)
(胡声丹，苗夺谦，姚一豫. 基于三支标签传播的半监督属性约简. 计算机学报，2021，44(11)：2332-2343)

[18] Zhang Qing-Hua, Zhi Xue-Chao, Wang Guo-Yin, et al. Multi-granularity ensemble classification algorithm based on attribute representation. Chinese Journal of Computers, 2022, 45(8): 1712-1729(in Chinese)
(张清华，支学超，王国胤等. 基于属性代表的多粒度集成分类算法. 计算机学报，2022，45(8)：1712-1729)

[19] Li Jin-Hai, Mi Yun-Long, Liu Wen-Qi. Incremental cognition of concepts: Theories and methods. Chinese Journal of Computers, 2019, 42(10): 2233-2250(in Chinese)
(李金海，米允龙，刘文奇. 概念的渐进式认知理论与方法. 计算机学报，2019，42(10)：2233-2250)

[20] Wang W, Zhan J, Zhang C, et al. A regret-theory-based three-way decision method with a priori probability tolerance dominance relation in fuzzy incomplete information systems. Information Fusion, 2023, 89: 382-396

[21] Chen Y, Wang P, Yang X, et al. Granular ball guided selector for attribute reduction. Knowledge-Based Systems, 2021, 229: 107326

计算机学报

[22] Dai J, Hu Q, Hu H, et al. Neighbor inconsistent pair selection for attribute reduction by rough set approach. IEEE Transactions on Fuzzy Systems, 2018, 26(2): 937-950

[23] Zhang X, Yao Y. Tri-level attribute reduction in rough set theory. Expert Systems with Applications, 2022, 190: 116187

[24] Sang B, Chen H, Wan J, et al. Self-adaptive weighted interaction feature selection based on robust fuzzy dominance rough sets for monotonic classification. Knowledge-Based Systems, 2022, 253: 109523

[25] Zhang Q, Wu C, Xia S, et al. Incremental learning based on granular ball rough sets for classification in dynamic mixed-type decision system. IEEE Transactions on Knowledge and Data Engineering, 2023, doi: 10.1109/TKDE.2023.3237833

[26] Naouali S, Salem S, Chtourou Z. Uncertainty mode selection in categorical clustering using the rough set theory. Expert Systems with Applications, 2020, 158: 113555

[27] Zhang Q, Cheng Y, Zhao F, et al. Optimal scale combination selection integrating three-way decision with hasse diagram. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(8): 3675-3689

[28] Xing J, Gao C, Zhou J. Weighted fuzzy rough sets-based tri-training and its application to medical diagnosis. Applied Soft Computing, 2022, 124: 109025

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

[30] Zhang Q, Ai Z, Zhang J, et al. A novel fast constructing neighborhood covering algorithm for efficient classification. Knowledge-Based Systems, 2021, 225: 107104

[31] Wu C, Zhang Q, Zhao F, et al. Three-way recommendation model based on shadowed set with uncertainty invariance. International Journal of Approximate Reasoning, 2021, 135: 53-70

[32] Sun L, Wang T, Ding W, et al. Feature selection using fisher score and multilabel neighborhood rough sets for multilabel classification. Information Sciences, 2021, 578: 887-912

[33] Liu J, Lin Y, Du J, et al. ASFS: A novel streaming feature selection for multi-label data based on neighborhood rough set. Applied Intelligence, 2023, 53(2): 1707-1724

[34] Bai J, Sun B, Chu X, et al. Neighborhood rough set-based multi-attribute prediction approach and its application of gout patients. Applied Soft Computing, 2022, 114: 108127

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

[36] Xia S, Peng D, Meng D, et al. Ball k-means: Fast adaptive clustering with no bounds. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(1): 87-99

[37] Xia S, Liu Y, Ding X, et al. Granular ball computing classifiers for efficient, scalable and robust learning. Information Sciences, 2019, 483: 136-152

[38] Xia S, Dai X, Wang G, et al. An efficient and adaptive granular-ball generation method in classification problem. IEEE Transactions on Neural Networks and Learning Systems, 2022, doi: 10.1109/TNNLS.2022.3203381

[39] Xia S, Zheng S, Wang G, et al. Granular ball sampling for noisy label classification or imbalanced classification. IEEE Transactions on Neural Networks and Learning Systems, 2021, doi:org/10.1109/TNN LS.2021.3105984

[40] Chen L. Topological structure in visual perception. Science, 1982, 218(4573): 699-700

[41] Rodriguez A, Laio A. Clustering by fast search and find of density peaks. Science, 2014, 344(6191): 1492-1496

[42] Yue X, Xiao X, Chen Y, et al. Robust neighborhood covering reduction with determinantal point process sampling. Knowledge-Based Systems, 2020, 188: 105063

[43] Du Y, Hu Q, Zhu P, et al. Rule learning for classification based on neighborhood covering reduction. Information Sciences, 2011, 181(24): 5457-5467

[44] Liu D, Li T, Li H. A multiple-category classification approach with decision-theoretic rough sets. Fundamenta Informaticae, 2012, 115(2): 173-188

WU Cheng-Ying, Ph. D. candidate. Her main research interests include knowledge discovery, and granular computing.

ZHANG Qing-Hua, Ph. D., professor, Ph. D. supervisor. His main research interests include rough sets, fuzzy sets, granular computing, and uncertain information processing.

ZHAO Fan, Ph. D. candidate. Her research interests include uncertain information processing, fuzzy sets, and rough sets.

CHENG Yun-Long, Ph. D. candidate. His research interests include granular computing, three-way decisions, and rough sets.

XIE Qin, Ph. D. candidate. Her research interests include data mining, knowledge discovery, and granular computing.

XIA Shu-Yin, Ph. D., professor, Ph. D. supervisor. His research interests include data mining, and granular computing.

WANG Guo-Yin, Ph. D., professor, Ph. D. supervisor, CCF Fellow, Yangtze River Scholar. His main research interests include granular computing, artificial intelligence, and cognitive computing.

Background

Granular computing (GrC) combined with human cognitive mechanism is an efficient way to reveal the descriptions of data, and successfully applied to knowledge discovery. In the era of big data mining, the contradiction between rich data and poor knowledge becomes more and more prominent. GrC, as a new paradigm for solving large-scale and complex problems, provides a new direction for data mining.

The core task of GrC is granulation that is the process of generating information granules. Granulation starting from the whole problem space first partitions the space into several subspaces, then constructs the desirable information granules on these subspaces based on equivalence, similarity, or proximity relations. An information granule is a group of objects with common properties and can be regarded as a community. Information granules are considered as a general outcome of data reduction. Because of this, an information granule can be regarded as an embedded computing unit to replace samples and improve the efficiency and fault tolerance of the model.

However, the existing granulation methods still have some shortcomings. For example, the granulation approach based on an equivalence relation is too strict, which makes granulation invalid when granulating quantitative data. The granulation approach based on neighborhood relation is not only time-consuming but also prone to form some redundant information granules. To the best of my knowledge, granular ball computing is the latest granulation method, but it may lose information.

Therefore, in this paper, an adaptive super interval granulation algorithm based on density peak clustering is proposed under considering the correlation between condition attributes from the perspective of decision attributes. The output of the algorithm is not only the partition of the universe, i.e., there is no information loss, but also all objects in partition block have the same label. Based on the super interval granule, a novel rough set model is proposed that can effectively unify quantitative data and qualitative data into one framework. Finally, to verify the effectiveness of proposed super interval granulation algorithm, we used it in classification task.

This work is supported by the National Natural Science Foundation of China (Nos. 62276038, 62221005), the National Key Research and Development Program of China (No. 2020YFC-2003502), the Foundation for Innovative Research Groups of Natural Science Foundation of Chongqing (No. cstc2019jcyj-cxttX0002), the Graduate Scientific Research Innovation Project of Chongqing under Grant No. CYB21207, and the Doctoral Talent Training Program of Chongqing University of Posts and Telecommunications (No. BYJS202012).

《计算机学报》