

模糊信息粒化理论及应用研究综述*

汤雨晴 于福生†

(北京师范大学数学科学学院, 100875, 北京)

摘要 信息粒化因模拟人类分析处理复杂问题的方式而受到广泛关注. 基于模糊集理论、粗糙集理论及商空间理论等典型信息粒化理论及方法的研究已取得长足的发展, 其中基于模糊集的信息粒化理论及方法旨在应对普遍存在的具有模糊性的问题和现象. 本文综述模糊信息粒化理论与方法的建立及发展过程, 梳理模糊信息粒化研究成果的系统体系结构, 总结其在聚类、预测和关联规则挖掘等重要领域的应用, 以促进该领域的研究.

关键词 信息粒化; 模糊信息粒化; 聚类; 预测; 关联规则挖掘

中图分类号 TP18

DOI: 10.12202/j.0476-0301.2022177

0 引言

1979年 Zadeh^[1]首次提出并讨论了信息粒度的概念, 自此掀开了对信息粒化的关注和研究热潮. 信息粒化思想渗透于诸多领域, 例如: 自动机与系统论中的“分解”与“划分”, 区间分析中的“区间数的运算”, 经济学中“集聚”的概念, 以及 D - S 证据理论中的“证据”等. Hobbs^[2]于 1985 年给出了粒度理论的框架, 探讨粒的分解与合并, 并给出了构造不同大小粒的方法; 1988 年 Lin^[3]提出邻域系统, 并研究了邻域系统与关系数据库之间的关系; 后来 Lin^[4-5]又提出了“granular computing”的概念, 并就粒计算与邻域系统展开深入研究.

信息粒化自提出以来, 其在学术界所受的关注度持续稳步提升. 信息粒化理论为研究者提供了不同的思维视角和处理问题的方式, 其研究成果已被众多国际期刊或学术会议所报道(图 1).

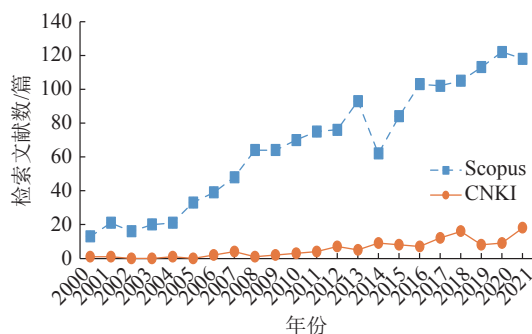


图 1 以“信息粒化”为关键词在 CNKI 与 Scopus 数据库所检索的文献结果显示

信息粒化作为一个新兴的研究热点应用广泛, 原因有 3 点: 1) 信息粒化的思想符合人们将问题建立在由一些通用且有概念意义的实体组成的框架之下的倾向; 2) 寻找对复杂问题较好的近似解而非最佳的精确解, 避免落入不必要的细节中, 降低计算成本; 3) 信息粒化在探究问题本质上是分层次的^[6].

信息粒化涵盖集合理论、区间分析、模糊集、粗糙集、概率论, 以及商空间理论等^[7]; 典型的是基于粗糙集理论的信息粒化^[3-6, 8-17]、基于商空间理论的信息粒化^[18-22]和基于模糊集理论的信息粒化^[1, 7, 23-38]这 3 种.

Pawlak^[8]于 1982 年创立粗糙集理论, 旨在根据不可分辨关系来划分论域, 以得到不区分的等价类, 并由此建立大小不同的概念粒所形成的近似空间, 其核心在于从近似空间导出表达集合概念的上、下近似集; 李道国等^[9]概述了粗糙集理论扩展, 包括等价关系泛化^[10]、基本知识粒度的构造、知识表示方法的推广, 以及粗糙集的代数方法^[11]; 在二元关系的基础上, Lin^[3-4]提出了粗糙集邻域系统, 并将粗糙集中的近似空间作为信息粒结构; Yao^[13-15]在知识挖掘等领域引入粒计算并进行深入研究, 为数据挖掘提供了新的视角和思路; 胡峰等^[16]给出了不完备信息系统在容差关系下的信息粒表示、分解和运算方法, 受粗糙集中属性约简的启发, 给出了信息粒框架下不完备信息系统属性必要性的判定; 刘清等^[17]基于粗糙集方法提出了粒逻辑, 并证实了其应用价值.

1990 年, 张钹等^[18]提出了研究和求解复杂问题的空间关系理论, 即商空间理论, 该理论力求综合来

* 国家自然科学基金资助项目(11971065, 11571001)

† 通信作者: 于福生(1965—), 男, 博士, 教授. 研究方向: 模糊数学与人工智能. E-mail: yufusheng@bnu.edu.cn

收稿日期: 2022-04-26

自于所有不同商空间对同一问题不同层次的分析,作为问题的最终解,符合人类智能的主要特征,因此基于商空间理论的信息粒化发展势头迅猛;刘仁金等^[19]深入分析并提出了图像分割问题的商空间粒度原理;Zhang 等^[20]设计了多分辨率信号分析的商空间近似模型,并探讨其性质和特点;徐峰等^[21-22]围绕商空间理论的推广和应用展开研究,如论域合成技术、模糊粒度与分层递阶结构的关系等,他们提出基于商空间的非均匀粒度聚类,旨在借助距离度量空间的性质,在不同粒度和不同层次的观察和理解下,研究模糊粒度聚类算法。

面对复杂且定义不明晰的现象,1965年 Zadeh^[24]创立了模糊集理论,他认为:信息粒、信息粒属性及其属性值的模糊性符合人类分析处理概念的偏好^[1, 7, 30]。在此基础上,围绕词计算理论^[31]、模糊数学和粒计算^[32-35]等人工智能领域的研究成果层出不穷;模糊信息粒化使我们能够在不确定的环境下作出合理决策。

40多年来,信息粒化在理论和应用研究方面均取得了长足发展,衍生出不同的信息粒化方法和种类^[40]。模糊信息粒化在推理中占据中心地位,因而对模糊信息粒化的研究具有深远的理论意义和应用前景。

1 模糊信息粒化的理论基础

非模糊的粒化方式(c -粒化)普遍存在于信息粒化中,如将问题求解空间变为划分空间。尽管非模糊信息在众多领域或方法中应用广泛,但在大部分推理与概念形成的过程中,粒都是模糊的(f -粒化)、不确切的。例如:就人的头部而言,鼻子、头发及两颊等粒子之间没有明确的边界,从而是模糊的;同时,模糊粒的属性,如头发的长度是模糊的;此外,属性值诸如长、短、特别长也是模糊的;粒、粒的属性及其属性值的模糊性反映了人类粒化及处理概念的方式。

受人类粒化信息的方式以及据此进行计算与推理的启发,Zadeh^[7, 23]提出模糊信息粒化理论。在人类的认识中,粒的模糊性源于无区别、相似、邻近以及功能相近等这些概念的模糊性;模糊信息粒化理论之所以能够渗透于不精确、部分知识、部分确定以及部分真实的领域,与其牢固的理论基础是分不开的。

自 Zadeh^[24]提出模糊集合论之后,他便着眼于模糊系统及其状态的定义,为模糊系统理论体系的构架奠定了基础;1973年语言变量、模糊条件命题和模糊算法基本特征的引入^[25]成为一项转折性工作,为模糊信息粒化理论打下了基础;Zadeh^[26-29]又进一步详述语言变量、模糊“IF-THEN”规则以及模糊图,于1979年^[1]提出并讨论了模糊信息粒化问题,推动了模糊逻辑

理论及应用的发展;广义约束与“词计算理论”^[30]的提出标志着模糊信息粒化理论的诞生。

模糊信息粒化中最关键的问题是模糊信息粒的刻画与描述。Zadeh 将模糊信息粒描述为

$$g \triangleq (x \text{ 是 } G) \text{ 是 } \lambda, \quad (1)$$

式中: x 是论域 U 中的变量,一般 U 为实直线(或 \mathbf{R}^n); G 是 U 的1个凸模糊子集,由隶属函数 μ_G 刻画; λ 表示模糊概率(或似然性),是单位区间 $[0, 1]$ 的模糊子集。例如: $g \triangleq (x \text{ 是小的})$ 是可能的, $g \triangleq (x \text{ 比 } y \text{ 大得多})$ 是不可能的。

模糊信息粒的另一种命题描述为

$$g \triangleq x \text{ 是 } G. \quad (2)$$

式(2)是式(1)的特例,不具概率特性。为区别二者,称形如式(1)的命题刻画的模糊信息粒为 πp -粒,对应于式(2)的模糊信息粒为 π -粒。

2 模糊信息粒化的方法论

2.1 模糊信息粒化的问题本源 对给定数据建立一个或一批较为合理的模糊信息粒,使之能够取代数据中所蕴含的信息,是模糊信息粒化的问题本源。

给定数据 $X = \{x_1, x_2, \dots, x_N\}$,以建立一个能够合理描述 X 的模糊概念 G (以 X 为论域的模糊集)的模糊信息粒 P 为例,一旦确定 G ,即可确定模糊信息粒

$$P \triangleq x \text{ 是 } G, x \in X.$$

建立模糊信息粒等价于确定模糊概念 G 的隶属函数,即 $A = \mu_G$ 。除特殊情况外,本文用模糊信息粒 P 代替模糊概念 G ,即

$$P = A(x), x \in X,$$

式中 P 的基本形式包括区间型 $A(x; a, b)$ 、三角型 $A(x; a, m, b)$ 、梯型 $A(x; a, m, n, b)$ 和高斯型 $A(x; \mu, \sigma)$ 等。

尽管模糊信息粒的基本形式已经确定,但是模糊信息粒化过程中仍存在2个不可忽略的问题:1)单个模糊信息粒与一批模糊信息粒的构造方法不同;2)数据类型不同时,模糊信息粒的构造方法不同。

2.2 模糊信息粒的建立原则与方法 模糊信息粒化旨在使问题易于求解并更好地接近现实,因此,从现实问题出发,对于不同数据类型,如实数、时序、区间时序、模糊时序,以及其他数据,建立单个或一批模糊信息粒时,所考虑的原则和方法应当是不同的。

2.2.1 实数数据模糊信息粒的建立原则与方法 给定数据 $X = \{x_1, x_2, \dots, x_N\}$, $x_i \in \mathbf{R}$,模糊信息粒化的目标是在 X 上建立单个的模糊信息粒。假设已确定模糊信

息粒的基本形式为含参数的隶属函数A, 目标是求解最优隶属函数A中的参数. 模糊信息粒的建立^[34-36]需遵从2条准则(统称合理粒度原则): 1) 数据覆盖性, 模糊信息粒应尽可能好地代表数据; 2) 语义清晰性, 模糊信息粒应具有好的语义特指性.

优化问题

$$\max Q(A) = \sum_{x \in X} A(x) / E_m(F_S(A)),$$

式中: 最大化 $\sum_{x \in X} A(x)$ 保证模糊信息粒的数据覆盖性; 最小化A的支集 $F_S(A)$ 的测度 $E_m(F_S(A))$ 保证模糊信息粒的语义清晰性. 以A为梯形模糊信息粒为例, 其参数m和n可通过对实数由小到大重排后取平均数、中位数或众数等来确定. 参数a和b一般通过求解优化问题

$$\max Q(a) = \sum_{x \in X, x < m} A(x) / (m - a), \quad \max Q(b) = \sum_{x \in X, x > n} A(x) / (b - n)$$

获得. 这样建立的模糊信息粒虽具有好的语义清晰性, 但牺牲了数据覆盖性, 造成原有信息丢失, 且计算复杂度较高. 董克强^[41]定义了取值范围为[-1, 1]的仿隶属函数, 取代取值范围为[0, 1]的隶属函数, 对于范围跨度大、变化趋势陡峭的实数仍可建立合理的模糊信息粒, 并减少参数寻优带来的计算量; 孙秋艳^[42]认为: 模糊信息粒化过程中将所有数据的贡献不加以区分是不合适的, 并据此提出了加权模糊信息粒化方法. Gacek^[37]、Pedrycz等^[38]探索了关于合理粒度原则的更一般化表达, 分别以 $f_1\left(\sum_{x \in X} A(x)\right)$ 和 $f_2(E_m(F_S(A)))$ 来保证模糊信息粒的数据覆盖性和语义清晰性, 不断完善合理粒度原则; $f_1(x)$ 和 $f_2(x)$ 分别是关于x的增函数和减函数. 仍以A为梯形模糊信息粒为例, 参数m和n的确定方法同第2章所介绍, 参数a和b则经求解

$$\max V(a) = f_1\left(\sum_{x \in X, x < m} A(x)\right) f_2(m - a),$$

$$\max V(b) = f_1\left(\sum_{x \in X, x > n} A(x)\right) f_2(b - n)$$

优化问题得到. 需要说明的是, 合理粒度原则的数学形式与模糊信息粒化方法^[36, 43-46]不限于这些, 受篇幅限制这里不再列举.

在X上构造一批模糊信息粒是研究者们不断追求的目标, 此时各模糊信息粒所描述的数据因难以区别, 如: 相似、邻近或某种功能相近等, 而结合在一

起. 在关注数据分布的基础上, 考虑模糊聚类方法是较为直观的思路. 模糊c-均值聚类^[47-49]是对 $X = \{x_1, x_2, \dots, x_N\}$ 进行模糊信息粒化的主要工具^[43, 50-54], 通过最小化目标

$$Q = \sum_{i=1}^c \sum_{k=1}^N u_{ik}^m \|x_k - v_i\|^2$$

来实现. 式中: $U = (u_{ik})$ 为c-模糊划分矩阵, 且 $\sum_{i=1}^c u_{ik} = 1, k = 1, 2, \dots, N; 0 < \sum_{k=1}^N u_{ik} < N, i = 1, 2, \dots, c$. U给出X的1个c-模糊划分, 即将X分为c个模糊子集, 对应的类中心为 v_1, v_2, \dots, v_c ; m是模糊化因子, $m \geq 1$; $\|\cdot\|$ 是R空间中的范数. 最理想的c-模糊划分即c个模糊信息粒是通过目标函数的极值条件实施交替优化得到. 此基础上, 出现了关于模糊c-均值聚类的不同扩展形式, 例如可能性和带噪声聚类^[55-56]等; 数值相近的数据在聚类过程中更可能被划分至同一类, Pedrycz^[48]在目标函数中融合数据关于数值空间的邻域信息, 以此指导聚类过程.

2.2.2 时序数据上模糊信息粒的建立原则与方法 时序数据 $X = \{(t_1, x_1), (t_2, x_2), \dots, (t_N, x_N)\}$, $t_i < t_{i+1}, x_i \in \mathbf{R}$, 是由一维随机变量在不同时刻采样结果组成的序列, 也称时间序列; x_i 为随机变量在时刻 t_i 的1次采样, $i = 1, 2, \dots, N$. 通常采样按照等时间间隔方式进行, 因此X可简记作 $X = \{x_1, x_2, \dots, x_N\}$, $x_i \in \mathbf{R}$. 时间序列同时具有时间和数值粒度, 这区别于仅具有数值粒度的一般实数.

首先考虑在时间序列上建立单个模糊信息粒. 将时间序列退化为实数数据, 即关注时序数据在数值空间的分布, 则实数数据上单个模糊信息粒的建立方法均适用. 事实上根据该方法所建立的模糊信息粒因忽略时序信息而难以描述数据的趋势变化. Yang等^[57]证实了Pedrycz等^[34]、Yu等^[36]所提出的模糊信息粒化方法的局限性, 并定义了线性模糊信息粒.

定义 线性模糊信息粒 $G_L(k, b, \sigma, T)$ 的隶属函数为

$$f(x; kt + b, \sigma) = \exp\left(-\frac{(x - (kt + b))^2}{2\sigma^2}\right), t \in [0, T],$$

式中: $\mu(t) = kt + b$, 是随时间变化的核心线, $k, b \in \mathbf{R}$, 表示时变核心线的斜率和截距; σ 描述数据偏离时变核心线的程度; 参数 k, b 和 σ 通过线性回归确定.

线性模糊信息粒在经典模糊信息粒的基础上增加了时间维度, 能够描述数据随时间变化的不确定性, 包括趋势特征和波动范围等. Luo等^[58]尝试在极坐标下建立模糊信息粒, 将时变核心线的斜率与截距映射为极角与极径; 以时变区间取代时变核心线, 此

时得到的带状时变模糊信息粒能够表征趋势特征的可信度^[59]; 这些模糊信息粒在描述呈现明显线性趋势的时序数据时行得通, 然而时序数据的趋势变化通常复杂多样; Luo 等^[60]考虑了非线性时变核心线问题, 构造了广义带状时变模糊信息粒, 实现了对时序数据中非线性趋势的合理描述。

为建立时间序列上的一批模糊信息粒, 将时间序列退化为实数数据, 便可借助模糊聚类来实现。在聚类过程中视时序数据为完全独立的个体, 而忽略它们在时序意义上的相近性, 在面对现实问题时可能受限。Pedrycz^[48]提出具有时序约束的模糊聚类方法, 即引入能够刻画数据在时序上的相近信息, 且以绝对二元特征表示的邻域函数。蔡瑞琼^[61]指出: 1) 邻域内外、远近不同的数据对目标函数值的贡献应当不同; 2) 目标函数中不同项的比例是难以把握的。针对 1), Cai 等^[62]选取不同形式的模糊隶属函数作为邻域函数; 针对 2), 将目标函数中不同项的比例之和取为 1, 某项比例的上升必然会导致其他项比例的下降。在极度复杂性问题中, 采用语言值成为评判数据的惯用思维, 在目标函数中进一步加入体现聚类区间均匀的信息, 这样可以保证语言值描述的合理性^[61]。此类方法因关注时序数据在数值空间的分布而被称作“基于数值粒度的时序数据模糊信息粒化”, 这符合数值相差很大的 2 个数据被划分至不同的模糊信息粒中是更合理的思路。

研究者们倾向于从普遍使用或具有现实意义的时间概念中分析和研究时序数据, 例如时间概念 s、min、h、d、周、月、季度等。基于这种思想, 在时序数据上建立一批模糊信息粒分为 2 步: 1) 按照确定的时间概念, 在时间层面上等长分割时间序列为一系列子段; 2) 在各子段上建立模糊信息粒。显然, 步骤 2) 更为关键。董克强^[41]将实数中单个模糊信息粒的建立方法应用于时序数据的子段; Dong 等^[64]对子段集合进行模糊 c-均值聚类^[47], 由类中心衍生出具有丰富语义的词汇, 建立各子段上的模糊信息粒; 相似的模糊信息粒化思路, 在 Ma 等^[65]的研究中有所体现; Lu 等^[54]将模糊时间序列建模方法由数值层面推广至粒层面, 给出子段上模糊信息粒的建立方法。近年来, 人们深刻认识到对趋势本质的追踪是有效分析时序数据的途径, 因此探索并建立能够大致描述数据中趋势变化的模糊信息粒十分必要。考虑到差分序列 ΔX 能够反映时序数据的动态变化, Liu 等^[66]分别在 X 和 ΔX 的各子段上建立区间 II 型模糊信息粒; Yang 等^[57]建立各子段上的线性模糊信息粒, 揭示了数据的线性趋势以及量级水平, 符合人类分析处理时序数据的特点;

Luo 等^[58]在极坐标下探讨趋势的表征方式, 并给出了时序数据的极模糊信息粒化方法。尽管时间序列的等长分割符合人类在固定时间概念下分析数据并给出语义描述的偏好, 但在现实背景下, 采用该方法不免会疏漏或打破时序数据中有趣的动态模式, 因此有必要采用动态方法灵活分割时间序列^[67-68]。通过推广合理粒度原则, 可给出时间序列的不等长分割与直观描述^[37, 49]。考虑将时序数据分割为长度分别为 T_1, T_2, \dots, T_p 的 p 个不重叠子段, 并分别在其上建立模糊信息粒, 这可视作优化问题

$$\min_{T_1, T_2, \dots, T_p} \sum_{i=1}^p V(A_i),$$

式中: $V(A_i) = T_i \int_0^1 A_i^\alpha d\alpha$ 是建立在第 i 个子段数据上的模糊信息粒 A_i 的体积, 对确定的 α , A_i^α 是经合理粒度原则建立的第 i 个子段数据上的区间型信息粒。该思想在一些工作中多有体现^[69-71], 其强调时序数据在数值空间的分布; Lu 等^[72]在乘积空间 $X \times \Delta X$ 利用合理粒度原则, 给出时间序列的不等长分割和模糊粒化描述; Guo 等^[73]遵循 Cramer 分解定理来推广合理粒度原则, 证实融合数据线性趋势特征后建立的模糊信息粒更具“信息性”; Luo 等^[59]将时序数据转化为 k -线图, 并提出逐步线性分割的方法, 追踪时序数据的动态变化以实现不等长分割, 建立了能够评估线性趋势置信度的带状时变模糊信息粒。这些方法的可行性均过分依赖于各子段内的数据需呈现较为明显的线性趋势这一前提, 然而现实数据的变化趋势往往相当复杂。在子段上建立广义带状时变模糊信息粒^[60]是将带状时变模糊信息粒进一步推向实用的关键。Li 等^[74]提出时序数据的 2 个阶段分割过程, 给出多线性趋势模糊信息粒的概念, 充分考虑数据变化趋势本质的同时, 保证粒化在时间层面上的高度可解释性。基于时序信息的模糊信息粒化方法得到了长足的发展, 此类方法统称为基于时间粒度的时序数据模糊信息粒化。

基于数值和时间粒度的时序数据模糊粒化各有侧重: 前者采用模糊聚类思想, 根据时序数据数值分布的密集程度建立一批模糊信息粒; 后者在时间层面上分割时序数据并在子段上建立模糊信息粒。为发挥它们二者的优势, 有学者兼顾数值与时间粒度的时序数据模糊信息粒化, 采用动态时间规整的思想, 寻求这 2 种方法得到的粒化结果间的相似性^[61]。

2.2.3 区间时序数据上模糊信息粒的建立原则与方法 模糊信息粒化在区间时序数据上的推广具有现实意义。例如: 通过跟踪各日气温的最高和最低值来

了解天气的变化状况;根据各交易日股票的最高和最低价格来做决策;通过观察患者每天的收缩压和舒张压来评估健康状况等.这些实例均以区间值的形式来描述变量的不确定性.为将模糊信息粒化推广至区间时序数据,需要考虑建立区间数数据上模糊信息粒的原则与方法.

设: $X = \{x_1, x_2, \dots, x_N\}$, $x_i \in I(\mathbf{R})$ 是区间数数据, $I(\mathbf{R}) = \{x | x = [x^-, x^+]; x^- \leq x^+; x^-, x^+ \in \mathbf{R}\}$ 是区间数空间. 建立区间数数据上的单个模糊信息粒的相关方法大致分为如下2类^[75]:

1) 将区间数数据转换为2个独立的实数数据, 此时建立区间数数据的单个模糊信息粒问题转化为整合分别建立在2个实数数据的单个模糊信息粒问题. 例如: 董克强^[41]在由 X 的上下界所获得的彼此独立的实数数据上分别建立单个模糊信息粒; 刘艳^[76]为得到目标模糊信息粒采用不同的整合策略; 此外, X 可转化为由区间中点和半径做成的2个独立实数数据, 它们给出 X 的中心趋势和分散程度.

2) 遵循在区间数数据直接建立单个模糊信息粒的思想. 例如: 董克强^[41]将区间数数据细化为一个关于关系“ $<$ ”(“ $>$ ”)的全序集, 并借助区间数的序关系自然地将合理粒度原则推广至区间数数据; Zhang等^[77]受实数数据模糊信息粒化方法的启发, 提出了基于二元对比的区间数排序方法, 这样区间数数据的单个模糊信息粒的建立便迎刃而解. 这类方法本质是在 \mathbf{R} 上定义1个模糊信息粒, 根据实数对该模糊信息粒的隶属度确定区间数数据的模糊信息粒参数; Hao等^[78]将 \mathbf{R} 上的模糊信息粒转化为区间数空间 $I(\mathbf{R})$ 上的静态模糊信息粒, 其基本形式有三角型、抛物型或高斯型等, 并将合理粒度原则推广至区间数的情形以便求解静态模糊信息粒的参数.

建立区间数数据上一批模糊信息粒的常用方法是模糊聚类法. De Carvalho^[79]将范数的定义空间^[47-49]由 \mathbf{R} 推广至 $I(\mathbf{R})$, 并给出区间数数据的模糊 c -均值聚类问题的数学描述, 他为不同的类灵活选取距离度量以期获得形状和大小不同的类, 并进一步提出基于自适应平方欧氏距离的模糊 c -均值聚类; 基于自适应二次距离的模糊 c -均值聚类在兼顾区间值上下界的同时优化距离度量, 能更充分地考虑区间数数据的语义^[80]. D'urso等^[81-83]从区间中点与半径出发, 定义不同的距离度量, 提出并改进了模糊 c -中值聚类, 用以中和对异常值的敏感性; 提出了模糊 c -有序中值聚类, 以在距离度量的构造中融入不同数据对不同类的典型性信息. 结合 L_1 -距离和 Hausdorff 距离, De Carvalho等^[84]分别提出了基于自适应 L_1 -距离和 Hausdorff 距离

的模糊 c -均值聚类, 对数据中异常值的存在具有更强的鲁棒性; Rodriguez等^[85-87]根据区间数数据的上下界对聚类过程的不同重要性, 给出了含正则项的区间数数据软空间聚类, 提高了聚类结果对初始和异常值的鲁棒性; 基于自适应欧氏距离和 L_1 -距离, 概述了区间数数据的模糊聚类.

设

$$X = \{(t_1, x_1), (t_2, x_2), \dots, (t_N, x_N)\}, t_i < t_{i+1}, x_i = [x_i^-, x_i^+] \in I(\mathbf{R})$$

是区间时序数据, 随机变量在不同时刻的不确定性由区间值给出; 考虑 $\{t_1, t_2, \dots, t_N\}$ 是等时间间隔的情况, 此时可将 X 简记为 $X = \{x_1, x_2, \dots, x_N\}$, $x_i \in I(\mathbf{R})$; 除了数值粒度以外, 区间时序数据增加了时间粒度, 有别于区间数数据.

建立区间时序数据的单个模糊信息粒可将区间时序数据简单地退化为区间数数据, 此类方法并未考虑区间时序数据的时间层面信息, 故难以描述区间时序数据的动态变化. Hao等^[78]将静态模糊信息粒中实数形式的核推广到随时间变化的核心线, 进而定义并建立了线性动态模糊信息粒.

建立区间时序数据的一批模糊信息粒时, 若忽略区间时序数据的时序信息, 仅关注其数值空间分布, 则可借助模糊聚类实现区间时序数据的模糊信息粒化. 基于区间时序数据在时间层面的相近性, 蔡瑞琼^[61]提出了具有时序约束的模糊 c -均值聚类; 基于相对语言值的模糊 c -均值聚类, 符合人类评判数据的惯用思维. 这类方法是基于数值粒度的区间时序数据的模糊信息粒化. 为方便研究者分析区间时序数据, 在时间层面将区间时序数据分割为不同子段, 并在各子段上建立模糊信息粒, 这是基于时间粒度的区间时序数据模糊信息粒化的思想. Lu等^[88]构造了一系列能够描述区间值数据的振幅特征的基本概念, 并以模糊集的形式为在语义层面表达区间时序数据提供参考; Hao等^[78]根据预期预测水平确定了时间概念的大小, 将区间时序数据在时间层面等长分割为多个子段, 并在各子段上分别建立静态和线性动态模糊信息粒, 使得对区间时序数据趋势变化的分析和描述成为可能.

2.2.4 建立模糊时序数据模糊信息粒的原则与方法 人们在进行与主观评价相关的描述、判断或感知时, 均涉及模糊数的概念. 此外, 所测数据有时是不准确的, 为克服这一缺点, 模糊化数据是常用且合理的做法. 因此, 研究模糊时序数据上模糊信息粒的建立方法十分必要^[89].

模糊数数据 $X = \{x_1, x_2, \dots, x_N\}$, $x_i \in \mathcal{F}(\mathbf{R})$, 其中 $\mathcal{F}(\mathbf{R})$ 为全体 \mathbf{R} 的模糊子集所组成的集合. 给定置信水平 α ,

将所有模糊数的 α -截集做成一个区间数数据,分别建立不同 α 对应的区间数数据上的单个模糊信息粒,并由模糊集值统计方法建立模糊数数据上的单个目标模糊信息粒^[41].针对该方法在应用时带来的不切实际的计算成本,张明欣^[90]修正模糊数的距离定义,并基于与 X 中所有模糊数距离最小的原则,建立了模糊数数据上的单个模糊信息粒,受区间数数据模糊信息粒化方法的启发,她将模糊数数据转换为有限的多个相互独立的实数数据,分别建立了实数数据上的单个模糊信息粒,并以此确定了 X 上的单个目标模糊信息粒;Zhang等^[91]提出了基于二元对比的模糊数排序方法,所给出的对模糊数总体排序是合理可行的.

学者们^[47]将模糊 c -均值聚类推广至模糊数数据,建立一批模糊信息粒.Hathaway等^[92]、Pedrycz等^[93]基于统一广义坐标将模糊数转换为数值向量,结合模糊聚类方法得到了对 X 的模糊划分;在模糊 c -均值聚类框架下,基于模糊数之间的Hausdorff距离,产生由不同类型的模糊数组成的模糊数数据的模糊聚类方法^[94];Yang等^[95]将模糊聚类的适用性扩展至高维模糊数数据;Yang等^[96]定义了混合特征变量之间的不相似度量,由此衍生的模糊聚类方法适用于具有不同特征的数据类型,如符号型数据和模糊数数据等;D'urso等^[97]给出了模糊数之间的“加权”不相似度量,提出模糊数数据的加权模糊 c -均值聚类;De Oliveira等^[98]详细总结了利用模糊聚类思想对模糊数数据进行模糊信息粒化的方法.

设 $X = \{(t_1, x_1), (t_2, x_2), \dots, (t_N, x_N)\}$, $t_i < t_{i+1}, x_i \in \mathcal{F}(\mathbf{R})$ 是模糊时序数据.当 $\{t_1, t_2, \dots, t_N\}$ 为等时间间隔时, X 可简记为 $X = \{x_1, x_2, \dots, x_N\}$, $x_i \in \mathcal{F}(\mathbf{R})$.在不考虑时序关系的情况下,模糊数数据上模糊信息粒的建立方法均可用于模糊时序数据的模糊信息粒化,统称为基于数值粒度的模糊时序数据模糊信息粒化;基于时间粒度对模糊时序数据进行模糊信息粒化,在时间层面将模糊时序数据分割为多个子段,然后在各子段上建立模糊信息粒.张明欣^[90]在模糊时序数据的各等长子段上建立了模糊信息粒,并提供了3种确定子段合适长度的方法,其本质在于设计并寻优关于子段长度的指标;她基于对模糊时序数据在数值空间分布和时间层面信息的权衡,提出了兼顾数值与时间粒度的模糊时序数据模糊信息粒化,以实现最佳模糊粒化.

2.2.5 建立他类数据模糊信息粒的原则与方法 当今大数据时代,来源广泛的数据其形式也日益复杂,这使得依靠单一原则或方法来实现对各类数据的模糊信息粒化难以完成.

时空数据是指同时具有时间和空间信息,且可能

附带其他属性的一类数据.Izakian等^[99-100]对时空 $X = \{x_1, x_2, \dots, x_N\}$ 的数学形式进行了描述, x_i 由其空间 $x_i(s)$ 和时间 $x_i(t)$ 组成:

$$x_i = [x_i(s)|x_i(t)]^T = [x_{i1}(s), \dots, x_{ir}(s)|x_{i1}(t), \dots, x_{iq}(t)]^T,$$

空间部分含 r 个特征,时间部分含 q 个特征.此时时空数据的模糊信息粒化,即在 X 之上建立单个或一批模糊信息粒.Coppi等^[101]通过比较 X 的一对数据的瞬时位置来衡量二者的距离,并基于数据的速度和加速度概念,提出了 X 中2数据间的不相似度量,并在此基础上给出了数据间的一种折中距离^[101-102].D'urso^[103-104]将模糊 c -均值聚类推广至时空数据,并在其上建立了一批模糊信息粒;他们还提出了交叉时滞模糊 c -均值聚类,以保证模糊信息粒化方法对时间滞后的调整能力;融合了空间轨迹的趋势信息并提出趋势模糊 c -均值聚类;改进了模糊聚类以应对可能存在的异常数据.Coppi等^[105]在模糊 c -均值聚类的目标函数中增加空间惩罚项,以确保模糊信息粒化在空间上的合理性.Izakian等^[99-100]统筹了时空数据的空间与时间表示对数据间距离的影响,并给出了相应的模糊聚类方法;他们将时空数据在时间层面分割为等长的多个时空子段,通过模糊聚类获得了各时空子段内的时空结构.作为时空数据的1种,轨迹数据的空间部分含经度和纬度2个特征.高旭^[106]考虑空间轨迹的趋势信息,给出了二维线性模糊信息粒的数学定义.在模糊 c -中值聚类框架下,D'urso等^[107-108]对距离度量进行了指数变换,克服了其对异常时空数据的敏感性;针对不同空间单元的邻近性,他们基于动态时间规整,将具有空间约束的模糊 c -中值聚类应用于时空数据的模糊划分.

人们所获得的数据一般具有各种各样的特征,因此研究多维数据 $X = \{x_1, x_2, \dots, x_N\}$, $x_i \in \mathbf{R}^n$ 的模糊信息粒化十分必要.利用较为成熟的一维数据模糊信息粒化方法,分别考虑各维度的数据特征,所建立的模糊信息粒作卡氏积,即为多维数据上的目标模糊信息粒^[41, 109].在合理粒度原则的指导下,Wang等^[45]考虑了多维数据的聚类结果可以为各维度数据的模糊信息粒化过程提供参考,并遵循卡氏积聚合的思想建立了一批模糊信息粒;黄也白^[109]将合理粒度原则推广至多维,给出了多维数据上球状模糊信息粒的构造方法;在此基础上,董克强^[41]改变了隶属函数的表现形式以及其参数的确定策略,在多维数据上建立了圆锥形模糊信息粒;Zhu等^[110]给出了一批椭球状模糊信息粒的构造方法,并将数据对椭球状模糊信息粒的隶属度由实数值进一步推广为区间值,探讨了多维数据

上区间Ⅱ型模糊信息粒的建立,提供了高型模糊信息粒化的新思路^[111-113];赵芳等^[114]结合区间Ⅱ型模糊c-均值聚类与合理粒度原则,充分考虑了多维数据模糊信息粒化过程的不确定性;Yang等^[115]提出基于特征的Ⅱ型模糊c-均值聚类方法,在多维数据上建立一批Ⅱ型模糊信息粒。越来越多的学者意识到不可忽略多维数据各维度之间的内在相关性,这直接影响了模糊信息粒化的质量。Wang等^[116]借助主成分分析与合理粒度原则,力求在不相关空间中构造一批模糊信息粒,并改进了自上而下的模糊信息粒化模型,保证模糊信息粒更具描述性和代表性;Wang等^[117]给出了数据覆盖性和语义清晰性在高维情形下的数学描述,并推广了合理粒度原则。针对具有时间层面的多维时序数据,Liu等^[118]根据合理粒度原则构造了局部模糊信息粒,对局部模糊信息粒的参数曲线进行了拟合,建立了全局模糊信息粒。

3 模糊信息粒化的应用

作为人工智能领域正在兴起的研究分支,模糊信息粒化已被成功应用于预测、聚类和关联规则挖掘等领域。

3.1 应用于预测 对具有时序特征的数据进行预测,旨在从历史数据中挖掘规律并外推将来。现今大数据时代的数据呈现高维、复杂和不确定性等特征,传统预测方法难以满足实际应用中的多元化需求。将复杂问题简单化为若干子问题,并寻求子问题有意义的模糊描述成为迫切需求,因此模糊信息粒化在预测领域广受青睐。一般对数据进行模糊信息粒化,以便实现高效降维;面向不同的预测水平,在有语义的模糊信息粒层面结合传统的预测方法,保证了预测过程的可解释性。许多学者从探究模糊信息粒之间的解释映射出发,构造了模糊“IF-THEN”规则,并基于模糊推理系统完成预测过程。在时间维度上等长分割数据,并在子段数据上尝试不同的模糊粒化方法,包括非趋势的模糊粒化方法^[64-65, 69, 75]与趋势的模糊粒化方法^[57, 74, 119],通过灵活设计不同形式的模糊“IF-THEN”规则与模糊推理方法,以实现不同水平的预测;给出数据在时间维度上的动态分割和模糊信息粒化方法,并在模糊信息粒层面建立模糊“IF-THEN”规则,得到不同水平的预测结果^[59-60, 120]。鉴于这类方法容易出现规则冗余,故在统计分析方法中引入模糊信息粒化思想是一种选择。Guo等^[70]证实了隐马尔可夫模型在模糊信息粒层面对于实现不同水平预测的有效性;借鉴了马尔可夫过程中状态转移的思想,黄也白^[109]在模糊信息粒化的基础上定义了模糊状态转

移规则,并提出了多维时序数据的预测方法。神经网络凭借非线性逼近和自学习能力等优势,在预测领域具有显著的应用价值。Hao等^[78]在区间时序数据的模糊信息粒化的基础上应用人工神经网络,研究了不同水平的预测性能;Lu等^[53-54]借助模糊信息粒化,将时序数据映射为模糊认知图以及高阶模糊认知图的结构;Froelich等^[63]对模糊信息粒序列进行了模糊聚类,以构造模糊认知图的节点与各时刻的状态值;Wang等^[121]在线性模糊信息粒层面构造了自适应模糊认知图,并证实了其在时间序列长期预测问题中的现实意义;Liu等^[66]结合不同的神经网络模型并设计高型模糊信息粒,完成了预测过程;Tang等^[122]将时序数据的趋势特征作为长短期记忆网络的输入输出,其中趋势特征由线性模糊信息粒的参数刻画;Wang等^[71]证实了反向传播网络在时序数据长期预测问题中的有效性。

模糊信息粒化思想在处理多步预测问题时避免了对一步预测结果的反复迭代,在提升预测效率和精度的同时增加预测结果的可解释性。

3.2 应用于聚类 聚类分析是有效挖掘数据信息的重要方式之一。对具有时序特征的数据族进行聚类时,一般面临数据规模大且聚类效率低的问题,因此借助传统的聚类方法挖掘数据结构并不容易。

受人脑分析处理机制的启发,借鉴模糊信息粒化的思想,将具有时序特征的数据族转化为解释性强且适当规模的模糊信息粒序列族,在粒层面探求灵活高效的聚类分析方法备受关注。Guo等^[123]考虑聚类的对象由时序数据族转换为低维表示序列族,而时序数据族的动态特征得以保留,保证数据结构挖掘的智能和效率;Wang等^[124-125]认为,模糊信息粒序列族的模糊聚类结果能够揭示原始等长时序数据族的数据结构,为体现专家知识和不同特征空间下聚类知识的指导作用,结合模糊信息粒化和知识指导的思想,他们提出了面向等长和不等长时序数据族的模糊聚类方法;Duan等^[126]进行时序数据族的聚类研究时强调趋势变化的重要性,将时序数据族低维模糊信息粒化表出并应用于层次聚类;Guo等^[73, 127]将时序数据族的聚类问题描述为蕴含趋势信息的模糊信息粒序列族的层次聚类问题,考虑金融时序数据族的波动特征,实现了信息粒化框架下的聚类;刘艳^[76]将模糊信息粒序列族的聚类结果视作对原始等长区间时序数据族整体结构的挖掘,进一步给出了基于模糊信息粒化和协作的等长区间时序数据族的模糊聚类方法,以融入来自不同特征空间的指导知识。

尽管模糊信息粒化在预测分析和聚类分析中应

用广泛,但是针对模糊信息粒的设计、向更复杂现实场景的推广,以及分析过程的高效性等仍值得深入探讨.总之,寻找更有效的模糊信息粒化方法是研究者追求的目标,以帮助人类探索隐藏于数据中的多层次与多视角的知识.

3.3 应用于关联规则挖掘 作为数据挖掘中的一个重要课题,关联规则挖掘是发现大量数据中潜在有用的关联的过程.

模糊信息粒化思想在关联规则挖掘领域的应用,对于现实决策有极大的参考价值,主要有 2 点:1)保证关联规则在语义表示上符合人脑的概念生成及推理方式;2)追求更高效的挖掘算法. Bettini 等^[128]首次在信息粒化框架下探讨时序数据挖掘问题,之后的研究大多沿袭该思想.在模糊信息粒化思想的指导下,湛力^[129]结合模糊聚类将数值时序关联规则转换为模糊时序关联规则;Li 等^[130]推广模糊关联规则的支持度定义,这是线性模糊信息粒化和模糊聚类下的直接结果,由此提出模糊时序关联规则挖掘算法,并证实其在时间效率上的优越性.

优化模糊信息粒化,彰显其在应对海量数据的关联规则挖掘任务时的优势是一个重要方向.此外,模糊信息粒化在关联规则挖掘中的应用,需要向更加多样化的关联规则形式做推广.关于这方面的研究成果尚不多,欢迎更多学者积极参与该问题的研究.

3.4 应用于其他领域 模糊信息粒化的思想渗透于诸多领域,例如模糊控制等.李征等^[131]以语词计算与信息粒化为基础,设计了一类可模拟人工智能的模糊控制器,其擅长分析复杂且不确定的大型数据系统;针对多级模糊控制固有的维数局限,Kacprzyk^[132]将模糊信息粒化方法应用于状态和控制空间,从而简化了多级控制.寻求模糊信息粒化思想在提高模糊控制系统的稳定性、精确性与快速性方面的潜力,有待于进一步深入研究和探讨.

模糊信息粒化思想在解决数据分类问题时发挥了重要作用.Kumar 等^[133]在模糊信息粒层面设计了神经网络,证实其在分类可解释性方面的进步;Balamash 等^[134]利用模糊信息粒初始化分类器;Fu 等^[135]建立了加权数据的模糊信息粒,以初始化基于规则的模糊分类模型.受人脑信息粒化机制的启发,Fu 等^[136]对多维数据进行分类建模.基于模糊信息粒化的数据分类研究尚存在一些问题,如模糊信息粒的设计、分类模型的可解释性研究,以及应用场景的推广等.

为了进一步提高机器智能,当务之急是研究和处理模糊性问题,用模糊信息粒语言解释不精确性的本质与机制,可能突破人工智能在可解释性上遇到的

“瓶颈”问题.同时,模糊信息粒化思想凭借多粒度知识表示以及问题的求解能力,在人工智能领域有极大优势,在模糊、不完整、不精确以及不确定性广泛存在的诸多领域,如机器学习、图像处理、医疗诊断、品种选育、小波分析、异常检测、交通轨迹、信息控制、地质地震、天气预报,以及股票交易等,研究模糊信息粒化理论与方法成为必然趋势.然而,模糊信息粒化理论和应用本身尚存在很多值得深入探究的问题,为促进其持久良好的发展态势,本研究提出如下 4 点思考:1)快速寻找合适的模糊粒度,实现问题处理效率和求解精度间的平衡是智能化的关键,可以考虑建立形式化的数学模型或开发通用且成熟的算法;2)不同粒层之间的关系和自由转换是一个极具潜力的研究方向;3)模糊信息粒化模型在动态数据系统或没有先验知识条件下的应用,值得进一步深入探讨;4)从模糊信息粒化理论研究与方法优化这 2 方面出发,促进其处理海量数据或复杂问题时的高效性等.

4 结语

模糊信息粒化是人工智能领域发展起来的一种全新的理念,用于应对普遍存在且具有模糊性的问题和现象.本文以信息粒化作为切入点,介绍了其研究进展与现状,引出模糊信息粒化在人们推理中的中心地位;聚焦模糊信息粒化,概述了其理论基础,梳理了其问题本源,探讨了不同数据模糊信息粒化的原则与方法;论述了模糊信息粒化的应用前景,并进一步提出了研究方向和问题.希望这些工作能够对该领域的研究同行提供参考和帮助,并促进相关研究的发展.

5 参考文献

- [1] ZADEH L A. Fuzzy sets and information granularity[C]//Advances in Fuzzy Set Theory and Applications, July 19, 1979, Holland. Amsterdam: North-Holland Publishing, 1979: 3
- [2] HOBBS J R. Granularity[C]//Proceedings of the Ninth International Joint Conference on Artificial Intelligence, August 18-23, 1985, California, United States. San Francisco: Morgan Kaufmann Publishers, 1985: 432
- [3] LIN T Y. Neighborhood systems and relational databases[C]//Proceedings of the 1988 ACM Sixteenth Annual Conference on Computer Science-CSC'88, February 1, 1988, United States. New York: Association for Computing Machinery, 1988: 725
- [4] LIN T Y. Granular computing on binary relations I: data mining and neighborhood systems, II: rough set representations and belief functions, rough sets in

- knowledge discovery[M]. Heidelberg: Physica-Verlag, 1998: 107
- [5] LIN T Y. Granular computing rough set perspective[J]. The Newsletter of the IEEE Computational Intelligence Society, 2005, 2(4): 1543
- [6] 王国胤. Rough集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001
- [7] ZADEH L A. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic[J]. Fuzzy Sets and Systems, 1997, 90(2): 111
- [8] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11: 341
- [9] LIU G L, ZHU W. The algebraic structures of generalized rough set theory[J]. Information Sciences, 2008, 178(21): 4105
- [10] KRYSZKIEWICZ M. Rough set approach to incomplete information systems[J]. Information Sciences, 1998, 112(1/2/3/4): 39
- [11] 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001
- [12] 王国胤, 姚一豫, 于洪. 粗糙集理论与应用研究综述[J]. 计算机学报, 2009, 32(7): 1229
- [13] YAO Y Y. Relational interpretations of neighborhood operators and rough set approximation operators[J]. Information Sciences, 1998, 111(1/2/3/4): 239
- [14] YAO Y Y. Information granulation and rough set approximation[J]. International Journal of Intelligent Systems, 2001, 16(1): 87
- [15] YAO Y Y. Three perspectives of granular computing[J]. Journal of Nanchang Institute of Technology, 2006, 25(2): 16
- [16] 胡峰, 黄海, 王国胤, 等. 不完备信息系统的粒计算方法[J]. 小型微型计算机系统, 2005(8): 1335
- [17] 刘清, 刘群. 粒及粒计算在逻辑推理中的应用[J]. 计算机研究与发展, 2004(4): 546
- [18] 张钹, 张玲. 问题求解理论及应用[M]. 北京: 清华大学出版社, 1990
- [19] 刘仁金, 黄贤武. 图像分割的商空间粒度原理[J]. 计算机学报, 2005(10): 1680
- [20] ZHANG L, ZHANG B. A quotient space approximation model of multiresolution signal analysis[J]. Journal of Computer Science and Technology, 2005(20): 90
- [21] 徐峰, 张铃, 王伦文. 基于商空间理论的模糊粒度计算方法[J]. 模式识别与人工智能, 2004, 17(4): 424
- [22] 徐峰, 张铃. 基于商空间的非均匀粒度聚类分析[J]. 计算机工程, 2005(3): 26
- [23] ZADEH L A. Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems[J]. Soft Computing, 1998, 2(1): 23
- [24] ZADEH L A. Fuzzy sets[J]. Information and Control, 1965, 8(3): 338
- [25] ZADEH L A. Outline of a new approach to the analysis of complex systems and decision processes[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1973, SMC/3(1): 28
- [26] ZADEH L A. The concept of a linguistic variable and its application to approximate reasoning: I[J]. Information Sciences, 1975, 8(3): 199
- [27] ZADEH L A. The concept of a linguistic variable and its application to approximate reasoning: II[J]. Information Sciences, 1975, 8(4): 301
- [28] ZADEH L A. The concept of a linguistic variable and its application to approximate reasoning: III[J]. Information Sciences, 1975, 9(1): 43
- [29] ZADEH L A. A fuzzy-algorithmic approach to the definition of complex or imprecise concepts[J]. International Journal of Man-Machine Studies, 1976, 8(3): 249
- [30] ZADEH L A. Fuzzy logic = computing with words[J]. IEEE Transactions on Fuzzy Systems, 1996, 4(2): 103
- [31] WANG P P. Computing with words[M]. Carrolton: Wiley-Interscience, 2001
- [32] PEDRYCZ W. Granular computing: an emerging paradigm[M]. New York: Physica-Verlag Heidelberg, 2001
- [33] PEDRYCZ W, VUKOVICH G. Abstraction and specialization of information granules[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2001, 31(1): 106
- [34] PEDRYCZ W, GACEK A. Temporal granulation and its application to signal analysis[J]. Information Sciences, 2002, 143(1/2/3/4): 47
- [35] BARGIELA A, PEDRYCZ W. Granular computing: an introduction[M]. Boston, Dordrecht, and London: Kluwer Academics Publishers, 2003
- [36] YU F S, PEDRYCZ W. The design of fuzzy information granules: tradeoffs between specificity and experimental evidence[J]. Applied Soft Computing, 2009, 9(1): 264
- [37] GACEK A. Granular modelling of signals: a framework of granular computing[J]. Information Sciences, 2013, 221: 1
- [38] PEDRYCZ W, HOMENDA W. Building the fundamentals of granular computing: a principle of justifiable granularity[J]. Applied Soft Computing, 2013, 13(10): 4209
- [39] 李道国, 苗夺谦, 张东星, 等. 粒度计算研究综述[J]. 计算机科学, 2005(9): 1

- [40] ZHANG Y Q. Constructive granular systems with universal approximation and fast knowledge discovery[J]. *IEEE Transactions on Fuzzy Systems*, 2005, 13(1): 48
- [41] 董克强. 时间序列的信息粒化及基于粒化的聚类分析[D]. 北京: 北京师范大学, 2005
- [42] 孙秋艳. 基于模糊信息粒化的时间序列相似性研究[D]. 北京: 北京师范大学, 2007
- [43] PEDRYCZ W, SUCCI G, SILLITTI A, et al. Data description: a general framework of information granules[J]. *Knowledge-Based Systems*, 2015, 80: 98
- [44] PEDRYCZ W, WANG X M. Designing fuzzy sets with the use of the parametric principle of justifiable granularity[J]. *IEEE Transactions on Fuzzy Systems*, 2016, 24(2): 489
- [45] WANG X, PEDRYCZ W, GACEK A, et al. From numeric data to information granules: a design through clustering and the principle of justifiable granularity[J]. *Knowledge-Based Systems*, 2016, 101: 100
- [46] PEDRYCZ W, BARGIELA A. An optimization of allocation of information granularity in the interpretation of data structures: toward granular fuzzy clustering[J]. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2012, 42(3): 582
- [47] BEZDEK J C. *Pattern recognition with fuzzy objective function algorithms*[M]. New York: Kluwer Academic Publishers, 1981
- [48] PEDRYCZ W. *Knowledge-based clustering: from data to information granules*[M]. Hoboken: John Wiley & Sons, Inc, 2005
- [49] PEDRYCZ W, GOMIDE F. *Fuzzy systems engineering: toward human-centric computing*[M]. Hoboken: John Wiley & Sons, Inc, 2007
- [50] BARGIELA A, PEDRYCZ W. A model of granular data: a design problem with the Tchebyshev FCM[J]. *Soft Computing*, 2005, 9(3): 155
- [51] 高新波. 模糊聚类分析及其应用[M]. 西安: 西安电子科技大学出版社, 2004
- [52] 范九伦. 模糊聚类新算法与聚类有效性问题研究[D]. 西安: 西安电子科技大学, 1998
- [53] LU W, YANG J H, LIU X D. Numerical prediction of time series based on FCMs with information granules[J]. *International Journal of Computers Communications & Control*, 2014, 9(3): 313
- [54] LU W, YANG J H, LIU X D, et al. The modeling and prediction of time series based on synergy of high-order fuzzy cognitive map and fuzzy *c*-means clustering[J]. *Knowledge-Based Systems*, 2014, 70: 242
- [55] KRISHNAPURAM R, KELLER J M. A possibilistic approach to clustering[J]. *IEEE Transactions on Fuzzy Systems*, 1993, 1(2): 98
- [56] DAVE R N. Characterization and detection of noise in clustering[J]. *Pattern Recognition Letters*, 1991, 12(11): 657
- [57] YANG X Y, YU F S, PEDRYCZ W. Long-term forecasting of time series based on linear fuzzy information granules and fuzzy inference system[J]. *International Journal of Approximate Reasoning*, 2016, 81: 1
- [58] LUO C, SONG X, ZHENG Y J. A novel forecasting model for the long-term fluctuation of time series based on polar fuzzy information granules[J]. *Information Sciences*, 2020, 512: 760
- [59] LUO C, TAN C H, ZHENG Y J. Long-term prediction of time series based on stepwise linear division algorithm and time-variant zonary fuzzy information granules[J]. *International Journal of Approximate Reasoning*, 2019, 108: 38
- [60] LUO C, WANG H Y. Fuzzy forecasting for long-term time series based on time-variant fuzzy information granules[EB/OL]. (2019-12-30)[2022-04-10]. <https://www.sciencedirect.com/science/article/pii/S1568494619308282>
- [61] 蔡瑞琼. 兼顾数值粒度与时间粒度的时序数据模糊粒化[D]. 北京: 北京师范大学, 2009
- [62] CAI R Q, YU F S. Fuzzy temporal constraints based fuzzy clustering algorithm for temporal dataset[C]//2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, August 14-16, 2009, China. Tianjin: IEEE, 2009: 480
- [63] FROELICH W, PEDRYCZ W. Fuzzy cognitive maps in the modeling of granular time series[J]. *Knowledge-Based Systems*, 2017, 115: 110
- [64] DONG R J, PEDRYCZ W. A granular time series approach to long-term forecasting and trend forecasting[J]. *Physica A: Statistical Mechanics and its Applications*, 2008, 387(13): 3253
- [65] MA C, ZHANG L Y, PEDRYCZ W, et al. The long-term prediction of time series: a granular computing based design approach[EB/OL]. (2022-01-31)[2022-04-10]. <https://ieeexplore.ieee.org/document/9697907>
- [66] LIU S, PEDRYCZ W, GACEK A, et al. Development of information granules of higher type and their applications to granular models of time series[J]. *Engineering Applications of Artificial Intelligence*, 2018, 71: 60
- [67] ABONYI J, FEIL B, NEMETH S, et al. Modified Gath-Geva clustering for fuzzy segmentation of multivariate time-series[J]. *Fuzzy Sets and Systems*, 2005, 149(1): 39
- [68] KIM S J, KOH K, BOYD S, et al. L_1 trend filtering[J].

- SIAM Review, 2009, 51(2): 339
- [69] WANG W N, PEDRYCZ W, LIU X D. Time series long-term forecasting model based on information granules and fuzzy clustering[J]. Engineering Applications of Artificial Intelligence, 2015, 41: 17
- [70] GUO H Y, PEDRYCZ W, LIU X D. Hidden Markov models based approaches to long-term prediction for granular time series[J]. IEEE Transactions on Fuzzy Systems, 2018, 26(5): 2807
- [71] WANG W N, LIU W Q, CHEN H. Information granules-based BP neural network for long-term prediction of time series[J]. IEEE Transactions on Fuzzy Systems, 2021, 29(10): 2975
- [72] LU W, CHEN X Y, PEDRYCZ W, et al. Using interval information granules to improve forecasting in fuzzy time series[J]. International Journal of Approximate Reasoning, 2015, 57: 1
- [73] GUO H Y, WANG L D, LIU X D, et al. Trend-based granular representation of time series and its application in clustering[EB/OL]. (2021-02-26)[2022-04-10]. <https://ieeexplore.ieee.org/document/9364925>
- [74] LI F, TANG Y Q, YU F S, et al. Multilinear-trend fuzzy information granule based short-term forecasting for time series[EB/OL]. (2021-09-20)[2022-04-10]. <https://ieeexplore.ieee.org/document/9540982/keywords#keywords>
- [75] MACIEL L, BALLINI R. A fuzzy inference system modeling approach for interval-valued symbolic data forecasting[J]. Knowledge-Based Systems, 2019, 164: 139
- [76] 刘艳. 基于协作与模糊粒化的大规模区间值时间序列族聚类[D]. 北京: 北京师范大学, 2017
- [77] ZHANG M X, DONG K Q, YU F S. Fuzzy granulation of interval numbers[C/OL]. China, Shanghai: IEEE, 2011. [2022-04-12]. <https://ieeexplore.ieee.org/document/6019540>
- [78] HAO Y D, JIANG S R, YU F S, et al. Linear dynamic fuzzy granule based long-term forecasting model of interval-valued time series[J]. Information Sciences, 2022, 586: 563
- [79] DE CARVALHO F DE A T. Fuzzy c -means clustering methods for symbolic interval data[J]. Pattern Recognition Letters, 2007, 28(4): 423
- [80] DE CARVALHO F DE A T, TENÓRIO C P. Fuzzy k -means clustering algorithms for interval-valued data based on adaptive quadratic distances[J]. Fuzzy Sets and Systems, 2010, 161(23): 2978
- [81] D'URSO P, DE GIOVANNI L, MASSARI R. Trimmed fuzzy clustering for interval-valued data[J]. Advances in Data Analysis and Classification, 2015, 9(1): 21
- [82] D'URSO P, MASSARI R, DE GIOVANNI L, et al. Exponential distance-based fuzzy clustering for interval-valued data[J]. Fuzzy Optimization and Decision Making, 2017, 16(1): 51
- [83] D'URSO P, LESKI J M. Fuzzy c -ordered medoids clustering for interval-valued data[J]. Pattern Recognition, 2016, 58: 49
- [84] DE CARVALHO F A T, SIMÕES E C. Fuzzy clustering of interval-valued data with City-Block and Hausdorff distances[J]. Neurocomputing, 2017, 266: 659
- [85] RODRÍGUEZ S I R, DE CARVALHO F A T. A new fuzzy clustering algorithm for interval-valued data based on City-Block distances[C/OL]. Los Angeles, United States, New Orleans: IEEE, 2019. [2022-04-12]. <https://ieeexplore.ieee.org/document/8859017>
- [86] RODRÍGUEZ S I R, DE CARVALHO F A T. Soft subspace clustering of interval-valued data with regularizations[EB/OL]. (2021-06-10)[2022-04-11]. <https://www.sciencedirect.com/science/article/pii/S0950705121004536>
- [87] RODRÍGUEZ S I R, DE CARVALHO F A T. Clustering interval-valued data with adaptive Euclidean and City-Block distances[EB/OL]. (2022-03-09)[2022-04-11]. <https://www.sciencedirect.com/science/article/pii/S0957417422002366>
- [88] LU W, ZHOU W, SHAN D, et al. The linguistic modeling of interval-valued time series: a perspective of granular computing[J]. Information Sciences, 2019, 478: 476
- [89] COPPI R, GIORDANI P, D'URSO P. Component models for fuzzy data[J]. Psychometrika, 2006, 71(4): 733
- [90] 张明欣. 模糊数时间序列的模糊粒化与预测[D]. 北京: 北京师范大学, 2011
- [91] ZHANG M X, YU F S. A new pairwise comparison based method of ranking LR-fuzzy numbers[C]// International Conference on Artificial Intelligence and Computational Intelligence, October 23-24, 2010, Berlin, Germany. Heidelberg: Springer-Verlag, 2010: 160
- [92] HATHAWAY R J, BEZDEK J C, PEDRYCZ W. A parametric model for fusing heterogeneous fuzzy data[J]. IEEE Transactions on Fuzzy Systems, 1996, 4(3): 270
- [93] PEDRYCZ W, BEZDEK J C, HATHAWAY R J, et al. Two nonparametric models for fusing heterogeneous fuzzy data[J]. IEEE Transactions on Fuzzy Systems, 1998, 6(3): 411
- [94] YANG M S, KO C H. On a class of fuzzy c -numbers clustering procedures for fuzzy data[J]. Fuzzy Sets and Systems, 1996, 84(1): 49
- [95] YANG M S, LIU H H. Fuzzy clustering procedures for conical fuzzy vector data[J]. Fuzzy Sets and Systems,

- 1999, 106(2): 189
- [96] YANG M S, HWANG P Y, CHEN D H. Fuzzy clustering algorithms for mixed feature variables[J]. *Fuzzy Sets and Systems*, 2004, 141(2): 301
- [97] D'URSO P, GIORDANI P. A weighted fuzzy *c*-means clustering model for fuzzy data[J]. *Computational Statistics & Data Analysis*, 2006, 50(6): 1496
- [98] DE OLIVEIRA J V, PEDRYCZ W. *Advances in fuzzy clustering and its applications*[M]. England: John Wiley & Sons, Ltd, 2007
- [99] IZAKIAN H, PEDRYCZ W, JAMAL I. Clustering spatiotemporal data: an augmented fuzzy *c*-means[J]. *IEEE Transactions on Fuzzy Systems*, 2013, 21(5): 855
- [100] IZAKIAN H, PEDRYCZ W. Anomaly detection and characterization in spatial time series data: a cluster-centric approach[J]. *IEEE Transactions on Fuzzy Systems*, 2014, 22(6): 1612
- [101] COPPI R, D'URSO P. Fuzzy *k*-means clustering models for triangular fuzzy time trajectories[J]. *Statistical Methods & Applications*, 2002, 11: 21
- [102] D'URSO P. Dissimilarity measures for time trajectories[J]. *Journal of the Italian Statistical Society*, 2000, 9(1/2/3): 53
- [103] D'URSO P. Fuzzy *c*-means clustering models for multivariate time-varying data: different approaches[J]. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2004, 12(3): 287
- [104] D'URSO P. Fuzzy clustering for data time arrays with inlier and outlier time trajectories[J]. *IEEE Transactions on Fuzzy Systems*, 2005, 13(5): 583
- [105] COPPI R, D'URSO P, GIORDANI P. A fuzzy clustering model for multivariate spatial time series[J]. *Journal of Classification*, 2010, 27(1): 54
- [106] 高旭. 基于最小凸包体积距离和模糊信息粒化的轨迹聚类[D]. 北京: 北京师范大学, 2017
- [107] D'URSO P, DE GIOVANNI L, MASSARI R. Robust fuzzy clustering of multivariate time trajectories[J]. *International Journal of Approximate Reasoning*, 2018, 99: 12
- [108] D'URSO P, DE GIOVANNI L, DISEGNA M, et al. Fuzzy clustering with spatial-temporal information[J]. *Spatial Statistics*, 2019, 30: 71
- [109] 黄也白. 基于模糊信息粒化的多维时间序列预测[D]. 北京: 北京师范大学, 2016
- [110] ZHU X B, PEDRYCZ W, LI Z W. Granular data description: designing ellipsoidal information granules[J]. *IEEE Transactions on Cybernetics*, 2017, 47(12): 4475
- [111] ZHU X B, PEDRYCZ W, LI Z W. A two-stage approach for constructing type-2 information granules[J]. *IEEE Transactions on Cybernetics*, 2022, 52(4): 2214
- [112] PEDRYCZ W, AL-HMOUZ R, BALAMASH A S, et al. Hierarchical granular clustering: an emergence of information granules of higher type and higher order[J]. *IEEE Transactions on Fuzzy Systems*, 2015, 23(6): 2270
- [113] GACEK A, PEDRYCZ W. Clustering granular data and their characterization with information granules of higher type[J]. *IEEE Transactions on Fuzzy Systems*, 2015, 23(4): 850
- [114] 赵芳, 郭红月, 王利东. 基于区间二型FCM和合理粒度原则的信息粒生成方法及应用[J]. *模糊系统与数学*, 2021, 35(1): 101
- [115] YANG X Y, YU F S, PEDRYCZ W. Typical characteristic-based type-2 fuzzy *c*-means algorithm[J]. *IEEE Transactions on Fuzzy Systems*, 2021, 29(5): 1173
- [116] WANG L D, ZHAO F, GUO H Y, et al. Top-down granulation modeling based on the principle of justifiable granularity[J]. *IEEE Transactions on Fuzzy Systems*, 2022, 30(3): 701
- [117] WANG D G, LIU H, PEDRYCZ W, et al. Design Gaussian information granule based on the principle of justifiable granularity: a multi-dimensional perspective [EB/OL]. (2022-03-02)[2022-04-12]. <https://www.sciencedirect.com/science/article/pii/S0957417422002287>
- [118] LIU S, PEDRYCZ W, GACEK A, et al. A two-phase method of forming a granular representation of signals[J]. *Signal Processing*, 2017, 141: 1
- [119] LI F, YANG H L, YU F S, et al. A one-factor granular fuzzy Logical relationship based multi-point ahead prediction model[C]//2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), November 14-16, 2019, China, Dalian; IEEE, 2019: 1133
- [120] ZHAO Y Y, LI T T, LUO C. Spatial-temporal fuzzy information granules for time series forecasting[J]. *Soft Computing*, 2021, 25(3): 1963
- [121] WANG Y H, YU F S, HOMENDA W, et al. The trend-fuzzy-granulation-based adaptive fuzzy cognitive map for long-term time series forecasting[EB/OL]. (2022-04-26) [2022-04-11]. <https://ieeexplore.ieee.org/document/9763011>
- [122] TANG Y Q, YU F S, PEDRYCZ W, et al. Building trend fuzzy granulation based LSTM recurrent neural network for long-term time series forecasting[J]. *IEEE Transactions on Fuzzy Systems*, 2022, 30(6): 1599
- [123] GUO H Y, WANG L D, LIU X D, et al. Information granulation-based fuzzy clustering of time series[J]. *IEEE Transactions on Cybernetics*, 2021, 51(12): 6253
- [124] WANG X, YU F S, ZHANG H X, et al. Large-scale time

- series clustering based on fuzzy granulation and collaboration[J]. *International Journal of Intelligent Systems*, 2015, 30(6): 763
- [125] 王晓. 大规模时序数据的智能聚类 and 预测研究[D]. 北京: 北京师范大学, 2013
- [126] DUAN L Z, YU F S, PEDRYCZ W, et al. Time-series clustering based on linear fuzzy information granules[J]. *Applied Soft Computing*, 2018, 73: 1053
- [127] GUO H Y, KUANG H B, WANG L D, et al. Hierarchical axiomatic fuzzy set granulation for financial time series clustering[J]. *IEEE Transactions on Fuzzy Systems*, 2022, 30(3): 755
- [128] BETTINI C, WANG X S, JAJODIA S, et al. Discovering frequent event patterns with multiple granularities in time sequences[J]. *IEEE Transactions on Knowledge and Data Engineering*, 1998, 10(2): 222
- [129] 湛力. 时序关联规则的挖掘[D]. 北京: 北京师范大学, 2018
- [130] LI Z B, BU F, YU F S. Temporal fuzzy association rules mining based on fuzzy information granulation[C]//2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), July 29-31, 2017, China. Guilin: IEEE, 2017: 1168
- [131] 李征, 王维工. 基于信息粒化和语词计算的模糊控制器[J]. *辽宁工程技术大学学报(自然科学版)*, 2001(5): 636
- [132] KACPRZYK J. Optimal policies in multistage fuzzy control information granulation and interpolative reasoning[C]//2005 IEEE International Conference on Granular Computing, July 25-27, 2005, China. Beijing: IEEE, 2005: 1
- [133] KUMAR D A, MEHER S K, KUMARI K P. Fusion of progressive granular neural networks for pattern classification[J]. *Soft Computing*, 2019, 23(12): 4051
- [134] BALAMASH A, PEDRYCZ W, AL-HMOUZ R, et al. Granular classifiers and their design through refinement of information granules[J]. *Soft Computing*, 2017, 21(10): 2745
- [135] FU C, LU W, PEDRYCZ W, et al. Fuzzy granular classification based on the principle of justifiable granularity[J]. *Knowledge-Based Systems*, 2019, 170: 89
- [136] FU C, LU W, PEDRYCZ W, et al. Rule-based granular classification: a hypersphere information granule-based method[EB/OL]. (2020-01-15)[2022-04-15]. <https://www.sciencedirect.com/science/article/pii/S0950705120300198>

Fuzzy information granulation: review of theory and applications

TANG Yuqing YU Fusheng

(School of Mathematical Sciences, Beijing Normal University, 100875, Beijing, China)

Abstract Information granulation has attracted much attention for simulating the way humans analyze and deal with complex problems. Considerable strides have been made on theory and method of information granulation, typically based on fuzzy set theory, rough set theory, quotient space theory and so on, where fuzzy information granulation aims at handling problems and phenomena with fuzziness. In this paper we review establishment and development of fuzzy information granulation, sort out systematic structure of fuzzy information granulation. We summarize the extensive applications of fuzzy information granulation in fields such as clustering, forecasting and association rules mining.

Keywords information granulation; fuzzy information granulation; clustering; forecasting; association rules mining

【责任编辑: 陆有忠】