计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(4): 973−986, 2026

# 基于相对相关属性子集的最优尺度组合选择

王蕾晰[1] 吴伟志[1] 王熙照[2,3]

[1](浙江海洋大学信息工程学院 浙江舟山 316022)

[2](深圳大学计算机与软件学院 广东深圳 518060)

[3](广东省智能信息处理重点实验室(深圳大学) 广东深圳 518060)

(wangleixi22@163.com)

# Optimal Scale Combination Selection Based on Relatively Relevant Attribute Subsets

Wang Leixi[1], Wu Weizhi[1], and Wang Xizhao[2,3]

[1](*School of Information Engineering*, *Zhejiang Ocean University*, *Zhoushan*, *Zhejiang* 316022)

[2](*College of Computer Science and Software Engineering*, *Shenzhen University*, *Shenzhen*, *Guangdong* 518060)

[3](*Guangdong Key Laboratory of Intelligent Information Processing* (*Shenzhen University*), *Shenzhen*, *Guangdong* 518060)

**Abstract** Knowledge acquisition of multi-scale hybrid data is an important research direction in data modeling under a multi-granularity environment, and the selection of an optimal scale is a key step in the knowledge acquisition of multi-scale data.However, in the calculation of an optimal scale, most of multi-scale granular computing analysis models are only based on all attributes without considering the influence of the order of the selected attributes on the performance of the obtained optimal scale, which affects the robustness and effectiveness of the models.Based on this observation, a new method to search for an optimal scale combination based on relatively relevant attribute subsets is proposed to deal with the optimal scale selection in multi-scale hybrid data.Based on conditional entropies, quantitative measures of relatively relevance of conditional attributes with respect to the decision are first introduced in a generalized multi-scale hybrid decision system (GMHDS).The concept of positive region optimal scale combinations based on relatively relevant subsets of attributes is then defined to ensure that the optimal scale combinations are selected in highly relevant subsets of attributes.Furthermore, a new stepwise search algorithm for positive region optimal scale combinations with a directed order is designed in a GMHDS, which is used to select the scales corresponding to conditional attributes with high relevance to the decision during the search process.The experimental results demonstrate that, in most cases, the results of this model have better classification performances than those of other comparative models.

**Key words** granular computing; generalized multi-scale hybrid decision systems; optimal scale combinations; relatively relevant attribute subsets; rough sets

**摘 要** 多尺度混合数据的知识获取是多粒度环境下数据建模的重要研究方向,而最优尺度的选择是多尺度数据知识获取的一个关键步骤。然而,在计算最优尺度时,大多数多尺度粒计算分析模型都只是基于所有属性,并没有考虑所选择属性的顺序对所获最优尺度的作用,影响了模型的鲁棒性和有效性。为此,提出一种定向搜索基于相对相关属性子集的最优尺度组合的新方法,用以处理多尺度混合数据中的

**收稿日期**: 2025-06-09; **修回日期**: 2025-09-16

**基金项目**: 国家自然科学基金项目(12371466, 62376161, U24A20322); 深圳市稳定支持计划项目(20231122124602001)

This work was supported by the National Natural Science Foundation of China (12371466, 62376161, U24A20322) and the Shenzhen Stable Support Program (20231122124602001).

**通信作者**: 吴伟志(wuwz@zjou.edu.cn)

最优尺度组合选择问题。首先,基于条件熵,在一个广义多尺度混合决策系统(generalized multi-scale hybrid decision system, GMHDS)中给出条件属性关于决策的相对相关性的度量;其次,给出基于相对相关属性子集的正域最优尺度组合的定义,以确保最优尺度组合在相关性较高的属性子集中搜索;进一步,在一个 GMHDS 中设计一种新的定向搜索最优尺度组合的逐步搜索算法,在搜索过程中尽可能选择与决策相关性较高的条件属性所对应的尺度。实验结果表明,在大多数情况下,该模型的结果比其他对比模型的结果具有更好的分类表现。

**关键词** 粒计算;广义多尺度混合决策系统;最优尺度组合;相对相关属性子集;粗糙集

**中图法分类号** TP18

**DOI:** 10.7544/issn1000-1239.202550464 **CSTR:** 32373.14.issn1000-1239.202550464

粒计算(granular computing, GrC)是模拟人类问题思考一个新的计算范式,其思想最早来源于 Zadeh[1]提出的信息粒度(information granularity)概念,它以粒(granule)代替传统数据分析中的样本作为基本计算单元,利用数据集中的特征集或属性集构造论域上的二元关系来粒化论域获得信息粒,并进一步通过特征提取等方法获得蕴含在数据集中的决策知识。自 1997 年正式提出粒计算概念以来,粒计算在智能信息处理、模式识别、决策分析、大数据知识发现等领域的研究呈现出独特优势并获得了广泛应用[2-8]。

粗糙集数据分析[9]是粒计算的重要模型之一,经典 Pawlak 粗糙集数据分析模型主要针对符号型数据,通过属性子集构造论域上的等价关系,从而粒化论域获得每个对象的等价类(信息粒),进一步通过属性约简挖掘出蕴含在数据集中的决策规则集。由于 Pawlak 粗糙集数据分析模型的局限性,很多研究者发展了各种非等价关系下推广的粗糙集数据分析模型用于处理各种不同数据类型的知识获取问题,包括数值型、集合值型、混合值型等。

传统粗糙集数据分析模型针对的是单一尺度的数据集,即所处理数据的对象在每个属性下只取唯一值[10-15]。但在实际应用中,对象在同一属性下可能呈现多个尺度的属性值。例如,学生的考试成绩可以根据数据的用途用百分制、5 级记分制(优、良、中、及格、不及格)、2 级记分制(合格、不合格)等来记录,对于评奖评优往往采用百分制,而对于是否授予学位或准予毕业则可以采用 2 级计分制。针对类似数据集的知识表示与知识获取问题,Wu 和 Leung[16]提出了多尺度数据的粒计算分析模型,也称为 Wu-Leung 模型。近年来,多尺度数据分析已成为多粒度数据分析的重要模型之一[17-30]。在 Wu-Leung 模型中,有一个基本假设是每个属性都具有相同的尺度个数,但在实际应用中,不同的属性可能具有不同的尺度个数。为了突破 Wu-Leung 模型的假设存在的局限性,Li 等人[19]提出了广义多尺度决策系统(generalized multi-scale decision system, GMDS)的概念,数据集中不同的属性具有不同的尺度个数,该数据分析的核心思想是引入尺度组合的概念,一个尺度组合对应于一个单尺度决策系统,数据集中所有的尺度组合构成一个完备格,从系统中得到一个符合约束条件的最优尺度组合等价于在全体尺度组合所构成的完备格中获得一个符合约束条件的极大元,数据建模的目的就是在所选最优尺度组合所对应的决策系统中得到保持相同目标约束条件下的属性约简(特征选择),并进一步得到蕴含在数据集中的决策知识以及对知识进行不确定性分析等。因此,如何在格中快速搜索最优尺度组合是多尺度数据分析的重要研究内容[24-26]。通过定义新的尺度组合概念,Cheng 等人[24]设计了一种新的逐步搜索算法,能够同时进行最优尺度组合选择和属性约简的任务。由于格中的最优尺度组合不是唯一的,因此不同的属性筛选顺序会影响最后搜索的结果。但现有的算法在搜索尺度组合时没有考虑属性的顺序问题。而在决策系统中,属性并不是同等重要的,甚至存在属性是冗余的。在逐步搜索算法[24]中,若先考虑重要性较高的属性,当其所含的信息可能与多数重要性较低的属性所含的信息相似时,则重要性较高的属性容易被删去。而最优尺度组合对应的属性集中会保留较多重要性低的属性,这会导致存储资源的浪费,并且影响决策的正确性和简洁性。因此,本文将根据属性与决策的相关性来度量属性的重要性,在搜索最优尺度组合时,按照属性与决策的相关性由低到高的顺序来搜索属性的尺度,从而设计一种定向搜索的逐步搜索算法来获得一个最优尺度组合,以提高算法的搜索效率。

众所周知,大量现实生活中的同一批数据的不同

属性有可能具有不同类型特征，即一部分属性是符号型的，而另外一部分属性是数值型或者集值型等，类似的数据集称为混合信息系统或混合决策系统[31-33]。针对数值型数据，Sun 等人[34]引入了邻域多粒度粗糙集模型来研究不完备邻域决策系统中的属性约简问题。针对含有缺省值的符号型和数值型数据，Zhang 等人[35]利用了模糊证据理论来解决混合决策系统中的属性约简问题，但迄今还没有关于刻画多尺度混合数据与决策的相关性的研究。本文利用条件熵刻画最强尺度下属性与决策的相关性，用于筛选与决策相关性较高的属性子集进行最优尺度组合选择。

综上所述，亟需在最优尺度组合选择中找到一个合适的筛选顺序，以获得分类效果好的最优尺度组合，从而提升算法的搜索效率和模型的分类性能。该问题面临以下 3 个挑战：

1）现有的研究中没有涉及刻画多尺度混合数据与决策相关性的方法；

2）大多数的多尺度数据分析模型考虑的最优尺度组合都是基于所有属性的，而一些属性只区分个别对象，与整个系统的决策相关性较低；

3）现有的算法或是搜索一个最优尺度组合，或是搜索所有尺度组合，在搜索过程中没有考虑属性的顺序问题，算法的计算效率低。

针对上述挑战，本文提出一种基于相对相关性的定向搜索最优尺度组合的方法，我们研究的主要工作有以下 3 个方面：

1）基于条件熵，在最强尺度下定义属性与决策的 3 种相对相关性，以衡量各个属性对决策的重要性；

2）提出广义多尺度混合决策系统中基于相对相关属性子集的最优尺度组合的概念，减少与决策相关性较低的属性对最优尺度组合选择的影响；

3）设计一种新的定向逐步搜索算法，考虑属性与决策的相关性，提升算法搜索最优尺度组合的效率，并与其他 5 种属性约简算法进行比较。

## 1 相关工作

考虑多尺度混合数据中的尺度选择问题，Wang 等人[36]提出了广义多尺度混合决策系统（generalized multi-scale hybrid decision system, GMHDS）的概念，并基于正域和条件熵，研究了 GMHDS 中最优尺度组合之间的关系。在他们的讨论中，最优尺度组合考虑了所有属性。然而，有些特殊属性可能会影响搜索结果的有效性和搜索效率。尽管这些属性与决策的相关性不高，但在选择最优尺度组合时仍被保留。因此，我们引入一种基于相对相关属性子集来搜索最优尺度组合的选择方法。

由于现有算法在选择最优尺度组合时很少考虑属性顺序[37-38]，但属性顺序对搜索尺度组合的结果影响较大。尤其当算法搜索所有最优尺度组合的时间复杂度较高时，选择合适的属性顺序来确定分类性能更好的最优尺度组合更为实用。因此，本文在多尺度混合决策系统中提供一种新的定向逐步搜索算法。

在衡量属性集与决策的相关性方面，目前的研究方法主要包括以下 4 种：模糊集[39-40]、代价敏感[20,41]、证据理论[42]、信息熵理论[43-45]，其中信息熵理论由于具有相对成熟的理论体系和高效的计算方法，应用较为广泛。因此，本文根据 GMHDS 中不同的条件熵，构建属性与决策之间的 3 种相对相关性，并根据这些相关性对属性进行排序后比较定向搜索的 3 种正域最优尺度组合的分类效果。

## 2 预备知识

本节介绍决策系统、多尺度决策系统、广义多尺度混合决策系统、条件熵的相关概念。

### 2.1 决策系统和多尺度决策系统

设$U$是非空有限集合，记$\mathcal{P}(U)$为$U$上的幂集。若$C=\{C_1,C_2,\cdots,C_o\}\subseteq\mathcal{P}(U)$，满足$\varnothing\notin C$，$\bigcup_{i=1}^{o}C_i=U$，则称$C$是$U$上的一个覆盖。进一步，若$C_i\cap C_j=\varnothing$，$i\neq j$，$\forall i,\forall j\in\{1,2,\cdots,o\}$，则称$C$是$U$上的一个划分。

**定义 1.** 设$U$是一个非空有限集合，$\mathcal{X}=\{X_1,X_2,\cdots,X_t\}$和$\mathcal{Y}=\{Y_1,Y_2,\cdots,Y_u\}$是$U$上的 2 个覆盖，若对于任意的$X_i\in\mathcal{X}$，都存在$Y_j\in\mathcal{Y}$使得$X_i\subseteq Y_j$，则称$\mathcal{X}$细于$\mathcal{Y}$，也称$\mathcal{Y}$粗于$\mathcal{X}$，记为$\mathcal{X}\sqsubseteq\mathcal{Y}$。进一步，若存在$X_{i_0}\in\mathcal{X}$和$Y_{j_0}\in\mathcal{Y}$使得$X_{i_0}\subset Y_{j_0}$，则称$\mathcal{X}$严格细于$\mathcal{Y}$，记为$\mathcal{X}\sqsubset\mathcal{Y}$。

**定义 2**[9]. 称$S=(U,A\cup\{d\})$为一个决策系统（decision system, DS），其中$U=\{x_1,x_2,\cdots,x_n\}$为一个非空有限的对象集，称为论域，$A=\{a_1,a_2,\cdots,a_m\}$为一个非空有限的条件属性集，使得对于每个属性$a_j\in A$都有$a_j:U\to V_j$，其中$V_j=\{a_j(x)|x\in U\}$是属性$a_j$的值域，$d$为决策属性使得$d:U\to V_d$，$V_d=\{d(x)|x\in U\}$是$d$的值域。

对于任意$a_j\in A$，$j\in\{1,2,\cdots,m\}$，可导出$U$上一个等价关系$R_j$，其被称为不可分辨关系，定义如下：

$$R_j=\left\{(x,y)\in U\times U|a_j(x)=a_j(y)\right\}。$$

而对于任意$B\subseteq A$，等价关系$R_B$表示为

$$R_B=\left\{(x,y)\in U\times U|a_j(x)=a_j(y),\forall a_j\in B\right\}，\tag{1}$$

显然$R_B=\bigcap\limits_{a_j\in B}R_j$。对于对象$x\in U$，$x$关于$B$的等价类为

$$[x]_B=\{y\in U|(x,y)\in R_B\}，\tag{2}$$

也称为$x$由$B$导出的信息粒。

对于决策$d$，也可以导出$U$上一个等价关系：

$$R_d=\{(x,y)\in U\times U|d(x)=d(y)\}。\tag{3}$$

对象$x\in U$关于$d$的等价类为

$$[x]_d=\{y\in U|(x,y)\in R_d\}。\tag{4}$$

不失一般性，记$V_d=\{1,2,\cdots,\gamma\}$，则决策$d$导出的等价关系$R_d$将$U$划分（粒化）为互不相交的决策类

$$\mathcal{D}=U/R_d=\left\{D_1,D_2,\cdots,D_\gamma\right\}，$$

其中$D_e=\{x\in U|d(x)=e\}$，$e\in\{1,2,\cdots,\gamma\}$。

针对实际应用中数据集中同一个对象在同一个属性下根据不同尺度或粒度取不同属性值的知识获取问题，Wu 和 Leung[16] 首次提出了多尺度粗糙集数据分析模型，文献 [19] 中称为 Wu-Leung 模型。

**定义 3**[16]. 称$S=(U,A\cup\{d\})$为一个多尺度决策系统（multi-scale decision system, MDS），其中$U=\{x_1,x_2,\cdots,x_n\}$为一个论域，$A=\{a_1,a_2,\cdots,a_m\}$为一个非空有限的条件属性集，每个$a_j\in A$都有$I$个尺度等级，$d$是决策属性，这样的一个 MDS 可以表示为

$$\left(U,\left\{a_j^k\middle|k=1,2,\cdots,I,j=1,2,\cdots,m\right\}\cup\{d\}\right)。\tag{5}$$

对于$a_j^k\in A$和$k\in\{1,2,\cdots,I\}$，有$a_j^k:U\to V_j^k$，其中$V_j^k$是$a_j^k$的值域。对于$a_j^k\in A$和$k\in\{2,3,\cdots,I\}$，存在一个满射$g_j^{k,k-1}:V_j^k\to V_j^{k-1}$，使得$a_j^{k-1}=g_j^{k,k-1}\circ a_j^k$，即

$$a_j^{k-1}(x)=g_j^{k,k-1}\left(a_j^k(x)\right)，\ \forall x\in U，$$

其中$g_j^{k,k-1}$称为信息粒度变换。

这样的一个多尺度决策系统可以分解成$I$个单尺度决策系统$S^k=(U,A^k\cup\{d\})$，$k\in\{1,2,\cdots,I\}$。由于在 Wu-Leung 模型中有一个基本假设：所有属性具有相同的尺度个数。然而，在实际数据集中不同的属性可能会有不同的尺度个数，为此，Li 等人[19] 提出了广义多尺度决策系统的概念。

**定义 4**[19]. 称$S=(U,A\cup\{d\})$为一个广义多尺度决策系统（generalized multi-scale decision system, GMDS），其中$U=\{x_1,x_2,\cdots,x_n\}$为论域，$A=\{a_1,a_2,\cdots,a_m\}$为一个非空有限的条件属性集，每个$a_j\in A$对应$I_j$个尺度等级，$d:U\to V_d$是决策属性。这样的一个 GMDS 可以表示为

$$\left(U,\left\{a_j^k\middle|k=1,2,\cdots,I_j,j=1,2,\cdots,m\right\}\cup\{d\}\right)，\tag{6}$$

其中$a_j^k\in A$和$k\in\{1,2,\cdots,I_j\}$，$a_j^k:U\to V_j^k$，对于$k\in\{2,3,\cdots,I_j\}$，存在一个满射$g_j^{k,k-1}:V_j^k\to V_j^{k-1}$，使得$a_j^{k-1}=g_j^{k,k-1}\circ a_j^k$，$g_j^{k,k-1}$称为信息粒度变换。

在一个 GMDS$S=(U,A\cup\{d\})$中，对于每一个条件属性$a_j\in A(j\in\{1,2,\cdots,m\})$，若将$a_j$的尺度限制在第$l_j(0\leqslant l_j\leqslant I_j)$个尺度标记上，则所有条件属性的尺度所组成的指标集$L=(l_1,l_2,\cdots,l_m)$称为$S$的一个尺度组合，这里当$l_j=0$时，意味着属性$a_j$被删除。记$S$中全体尺度组合所构成的集合为$\mathcal{L}$，则每个尺度组合$L=(l_1,l_2,\cdots,l_m)\in\mathcal{L}$对应于一个单尺度决策系统$S^L=(U,A^L\cup\{d\})$，其中$A^L=\{a_1^{l_1},a_2^{l_2},\cdots,a_m^{l_m}\}$。

**定义 5**[19]. 设$S=(U,A\cup\{d\})$为一个 GMDS，$L=(l_1,l_2,\cdots,l_m)\in\mathcal{L}$，$P=(p_1,p_2,\cdots,p_m)\in\mathcal{L}$。若对于任意的$j\in\{1,2,\cdots,m\}$，有$l_j\geqslant p_j$，则称$L$强于$P$或$P$弱于$L$，记为$L\succeq P$。若进一步存在$j_0\in\{1,2,\cdots,m\}$使得$l_{j_0}>p_{j_0}$，则称$L$严格强于$P$，记为$L\succ P$。

易证，关系$\succeq$在全体尺度组合集$\mathcal{L}$上确定了一个偏序集$(\mathcal{L},\succeq)$。令

$$\begin{aligned}&L\wedge P=(l_1\wedge p_1,l_2\wedge p_2,\cdots,l_m\wedge p_m)，\\&L\vee P=(l_1\vee p_1,l_2\vee p_2,\cdots,l_m\vee p_m)，\end{aligned}\tag{7}$$

其中$l_j\wedge p_j=\min\{l_j,p_j\}$，$l_j\vee p_j=\max\{l_j,p_j\}$，$j\in\{1,2,\cdots,m\}$，则$(\mathcal{L},\succeq,\wedge,\vee)$构成了一个完备格，其最小元为$(0,0,\cdots,0)$，最大元为$\boldsymbol{I}=(I_1,I_2,\cdots,I_m)$。

### 2.2 广义多尺度混合决策系统

**定义 6**[36]. 称$S=(U,A\cup\{d\})$为一个 GMHDS，其中$U=\{x_1,x_2,\cdots,x_n\}$为论域，$A=A^{\text{num}}\cup A^{\text{nom}}\cup A^{\text{set}}=\{a_1,a_2,\cdots,a_m\}$为非空有限条件属性集，其中$A^{\text{num}}$，$A^{\text{nom}}$，$A^{\text{set}}$分别是数值型、符号型、集合值型属性子集，且两两不相交。每个属性$a_j\in A$都有$I_j$个尺度等级，$d$是决策属性。这样一个 GMHDS 可以表示为

$$\begin{aligned}&\left(U,A^{\text{num}}\cup A^{\text{nom}}\cup A^{\text{set}}\cup\{d\}\right)=\\&\left(U,\left\{a_j^k|k=1,2,\cdots,I_j,j=1,2,\cdots,m\right\}\cup\{d\}\right)。\end{aligned}\tag{8}$$

对于$a_j^k\in A^{\text{num}}\cup A^{\text{nom}}$和$k\in\{1,2,\cdots,I_j\}$，有$a_j^k:U\to V_j^k$，同时对于$a_j^k\in A^{\text{set}}$和$k\in\{1,2,\cdots,I_j\}$，有$a_j^k:U\to\mathcal{P}(V_j^k)$，其中$V_j^k$是$a_j^k$的值域。对于任意$a_j^k\in A$和$k\in\{2,3,\cdots,I_j\}$，存在一个满射$g_j^{k,k-1}:V_j^k\to V_j^{k-1}$，使得$a_j^{k-1}=g_j^{k,k-1}\circ a_j^k$，即

$$a_j^{k-1}(x)=g_j^{k,k-1}\left(a_j^k(x)\right)，\ \forall x\in U，$$

其中$g_j^{k,k-1}$称为信息粒度变换。其中，当$a_j^k\in A^{\text{set}}$和$k\in\{2,3,\cdots,I_j\}$时，有

$$g_j^{k,k-1}\left(a_j^k(x)\right)=\left\{g_j^{k,k-1}(v)\middle|v\in a_j^k(x)\right\}。$$

**注释 1.** 本文所研究的 GMHDS 是指同时含有数值型、符号型、集合值型的多尺度混合数据，即$A=$

$A^{\text{num}} \cup A^{\text{nom}} \cup A^{\text{set}}$，满足$A^o \cap A^u = \varnothing$，$o \neq u$，$o,u \in \{\text{num},\text{nom},\text{set}\}$。每个$L=(l_1,l_2,\cdots,l_m) \in \mathcal{L}$对应于一个单尺度的混合决策系统$S^L=(U,A^L \cup \{d\})$，其中$A^L = A^{\text{num},L} \cup A^{\text{nom},L} \cup A^{\text{set},L} = \{a_1^{l_1}, a_2^{l_2}, \cdots, a_m^{l_m}\}$，$A^{o,L} = A^o \cap A^L = \{a_j^{l_j} | a_j \in A, 0 \leqslant l_j \leqslant I_j\}$，$o \in \{\text{num},\text{nom},\text{set}\}$。而若$B \subseteq A$，则

$$B^L = B^{\text{num},L} \cup B^{\text{nom},L} \cup B^{\text{set},L} = \left\{ a_j^{l_j} | a_j \in B, 0 \leqslant l_j \leqslant I_j \right\},$$

这里$B^L$是尺度组合$L$限制在条件属性子集$B$上所得到的一个条件属性子集，它是$A^L$的一个子集。

在一个GMHDS $S=(U,A \cup \{d\})$中，对于给定尺度组合$L=(l_1,l_2,\cdots,l_m) \in \mathcal{L}$，恒假设在数值型属性$a_j \in A^{\text{num}}$的尺度$l_j$上存在一个邻域半径$\delta_j^{l_j}$满足$0 \leqslant \delta_j^{l_j} \leqslant \delta_j^{l_j-1}$。对于$L$，记$\delta^L = \left\{ \delta_j^{l_j} | a_j^{l_j} \in A^{\text{num},L} \right\}$为关于$L$的邻域半径集。

**定义 7**[36]. 设$S=(U,A \cup \{d\})$是一个GMHDS，$B \subseteq A$，对于$L=(l_1,l_2,\cdots,l_m) \in \mathcal{L}$和$\delta^L$，$U$上关于$B$的二元关系$R_B^{L,\delta^L}$定义如下：

$$R_B^{L,\delta^L} = \left\{ (x,y) \in U \times U | (x,y) \in R_j^{l_j,\delta_j^{l_j}}, \forall a_j^{l_j} \in B \right\}, \tag{9}$$

其中$R_j^{l_j,\delta_j^{l_j}}$是$U$上关于$a_j^{l_j} \in B^L$的二元关系，则

$$R_j^{l_j,\delta_j^{l_j}} = \begin{cases} R_{j,\text{num}}^{l_j,\delta_j^{l_j}}, \text{若} a_j^{l_j} \in B^{\text{num},L}, \\ R_{j,\text{nom}}^{l_j}, \text{若} a_j^{l_j} \in B^{\text{nom},L}, \\ R_{j,\text{set}}^{l_j}, \text{若} a_j^{l_j} \in B^{\text{set},L}, \end{cases} \tag{10}$$

而

$$R_{j,\text{num}}^{l_j,\delta_j^{l_j}} = \left\{ (x,y) \in U \times U | |a_j^{l_j}(x) - a_j^{l_j}(y)| \leqslant \delta_j^{l_j} \right\},$$
$$R_{j,\text{nom}}^{l_j} = \left\{ (x,y) \in U \times U | a_j^{l_j}(x) = a_j^{l_j}(y) \right\},$$
$$R_{j,\text{set}}^{l_j} = \left\{ (x,y) \in U \times U | a_j^{l_j}(x) \cap a_j^{l_j}(y) \neq \varnothing \right\}.$$

显然，$R_B^{L,\delta^L}$满足自反性和对称性，即$R_B^{L,\delta^L}$是相似关系，它将$U$粒化为若干相似类

$$U/R_B^{L,\delta^L} = \left\{ [x]_B^{L,\delta^L} | x \in U \right\},$$

其中$[x]_B^{L,\delta^L} = \left\{ y \in U | (x,y) \in R_B^{L,\delta^L} \right\}$为对象$x$由$B^L$和$\delta^L$所生成的信息粒。

**性质 1**[36].设$S=(U,A \cup \{d\})$是一个GMHDS，对于$L=(l_1,l_2,\cdots,l_m) \in \mathcal{L}$，$P=(p_1,p_2,\cdots,p_m) \in \mathcal{L}$，$B,C \subseteq A$，若$L \succeq P$，$B \subseteq C$，则：

1）$R_B^{L,\delta^L} \subseteq R_B^{P,\delta^P}$；

2）$[x]_B^{L,\delta^L} \subseteq [x]_B^{P,\delta^P}$，$\forall x \in U$；

3）$R_C^{L,\delta^L} \subseteq R_B^{L,\delta^L}$；

4）$[x]_C^{L,\delta^L} \subseteq [x]_B^{L,\delta^L}$，$\forall x \in U$。

**定义 8**[36]. 设$S=(U,A \cup \{d\})$是一个GMHDS，$\mathcal{D}=U/R_d=\{D_1,D_2,\cdots,D_\gamma\}$，对于$L \in \mathcal{L}$和$B \subseteq A$，决策类簇$\mathcal{D}$关于属性子集$B$的下、上近似定义为

$$\underline{app}_B^{L,\delta^L}(\mathcal{D}) = \bigcup_{e=1}^{\gamma} \underline{app}_B^{L,\delta^L}(D_e),$$
$$\overline{app}_B^{L,\delta^L}(\mathcal{D}) = \bigcup_{e=1}^{\gamma} \overline{app}_B^{L,\delta^L}(D_e), \tag{11}$$

其中

$$\underline{app}_B^{L,\delta^L}(D_e) = \left\{ x \in U | [x]_B^{L,\delta^L} \subseteq D_e \right\},$$
$$\overline{app}_B^{L,\delta^L}(D_e) = \left\{ x \in U | [x]_B^{L,\delta^L} \cap D_e \neq \varnothing \right\}, \tag{12}$$

$e \in \{1,2,\cdots,\gamma\}$。因此，决策类簇$\mathcal{D}$在$B^L$下的正域、负域、边界域分别为

$$POS_B^{L,\delta^L}(\mathcal{D}) = \underline{app}_B^{L,\delta^L}(\mathcal{D}),$$
$$NEG_B^{L,\delta^L}(\mathcal{D}) = U - \overline{app}_B^{L,\delta^L}(\mathcal{D}), \tag{13}$$
$$BN_B^{L,\delta^L}(\mathcal{D}) = \overline{app}_B^{L,\delta^L}(\mathcal{D}) - \underline{app}_B^{L,\delta^L}(\mathcal{D}).$$

**定义 9**[36]. 设$S=(U,A \cup \{d\})$是一个GMHDS，$\boldsymbol{I}=(I_1,I_2,\cdots,I_m)$，$\mathcal{D}=U/R_d$，对于$L \in \mathcal{L}$，若$POS_A^{L,\delta^L}(\mathcal{D}) = POS_A^{I,\delta^I}(\mathcal{D})$，则称$S^L$关于$S$是正域协调的，否则称$S^L$关于$S$不是正域协调的。若$S^L$关于$S$是正域协调的，且对于满足$L^* \prec L$的任意$L^* \in \mathcal{L}$，$S^{L^*}$关于$S$不是正域协调的，则称$L$是$S$的一个正域最优尺度组合（positive region optimal scale combination, POSC）。

### 2.3 条件熵

当一个决策系统$S=(U,A \cup \{d\})$中具有缺失的属性值，称$S$为一个不完备决策系统（incomplete decision system, IDS）。若系统中存在邻域半径$\delta$，则称$S=(U,A \cup \{d\},\delta)$为一个邻域决策系统（neighborhood decision system, NDS）。

为了刻画相似关系下系统的不确定性，Dai等人[46]引入了一种条件熵。

**定义 10**[46]. 设$S=(U,A \cup \{d\})$是一个IDS，$\mathcal{D}=U/R_d$，对于$B \subseteq A$，决策类簇$\mathcal{D}$关于$B$的条件熵定义为

$$H_1(\mathcal{D}|B) = -\frac{1}{|U|} \sum_{x \in U} \text{lb} \frac{\left| [x]_B^* \cap [x]_d \right|}{\left| [x]_B^* \right|}, \tag{14}$$

其中$|\cdot|$表示集合的基数，$[x]_B^*$是不完备决策系统上对象$x$的相似类，$[x]_d$用式(4)计算。

由于定义10中的条件熵在属性约简时关于条件属性增减并不单调，故Dai等人[47]给出了一个关于条件属性增减具有单调性的新条件熵。

**定义 11**[47]. 设$S=(U,A \cup \{d\})$是一个IDS，$\mathcal{D}=U/R_d=\{D_1,D_2,\cdots,D_\gamma\}$，对于$B \subseteq A$，决策类簇$\mathcal{D}$关于$B$的条件熵定义为

$$H_2(\mathcal{D}|B) = -\sum_{x \in U} \sum_{e=1}^{\gamma} \frac{\left| [x]_B^* \cap D_e \right|}{|U|} \text{lb} \frac{\left| [x]_B^* \cap D_e \right|}{\left| [x]_B^* \right|}, \tag{15}$$

其中若$[x]_B^* \cap D_e = \varnothing$，则默认为

$$\mathrm{lb}\left(\left|[x]_B^*\cap D_e\right|/\left|[x]_B^*\right|\right)=0\text{。}$$

**性质 2**[47]. 设 $S=(U,A\cup\{d\})$是一个 IDS, $B,C\subseteq A$, 若 $B\subseteq C$, 则 $H_2(\mathcal{D}|C)\leqslant H_2(\mathcal{D}|B)$。

在 NDS 中基于对象之间的相似关系, Sun 等人[48]引入了新的条件熵。

**定义 12**[48]. 设 $S=(U,A\cup\{d\},\delta)$是一个 NDS, $\mathcal{D}=U/R_d$, 对于 $B\subseteq A$, 决策类簇 $\mathcal{D}$ 关于 $B$ 的条件熵定义为

$$H_3^{\delta}(\mathcal{D}|B)=-\frac{1}{|U|}\sum_{x\in U}\mathrm{lb}\frac{\left|[x]_B^{\delta}\cap[x]_d\right|^2}{\left|[x]_B^{\delta}\right|\left|[x]_d\right|}\text{,}\qquad(16)$$

其中 $\delta$ 是邻域半径, $[x]_B^{\delta}$ 是对象 $x$ 关于 $B$ 和 $\delta$ 的相似类。

虽然 $H_3^{\delta}(\mathcal{D}|B)$ 关于条件属性集不具有单调性, 但其从属性和决策 2 个方面衡量属性与决策的相关性。

## 3 基于属性子集的最优尺度组合

本节根据式(14)~(16)所定义的 3 种条件熵, 给出属性与决策的相关性定义。

**定义 13.** 设 $S=(U,A\cup\{d\})$是一个 GMHDS, $\boldsymbol{I}=(I_1,I_2,\cdots,I_m)$, $\mathcal{D}=U/R_d=\{D_1,D_2,\cdots,D_\gamma\}$。对于 $a_j^k\in A$, $k\in\{1,2,\cdots,I_j\}$和 $\delta_j^k$, 决策 $d$ 的 3 种条件熵分别定义如下:

$$H_1^{\delta_j^k}\left(\mathcal{D}|\left\{a_j^k\right\}\right)=-\frac{1}{|U|}\sum_{x\in U}\mathrm{lb}\frac{\left|[x]_j^{k,\delta_j^k}\cap[x]_d\right|}{\left|[x]_j^{k,\delta_j^k}\right|}\text{,}$$

$$H_2^{\delta_j^k}\left(\mathcal{D}|\left\{a_j^k\right\}\right)=-\sum_{x\in U}\sum_{e=1}^{\gamma}\frac{\left|[x]_j^{k,\delta_j^k}\cap D_e\right|}{|U|}\mathrm{lb}\frac{\left|[x]_j^{k,\delta_j^k}\cap D_e\right|}{\left|[x]_j^{k,\delta_j^k}\right|}\text{,}$$

$$H_3^{\delta_j^k}\left(\mathcal{D}|\left\{a_j^k\right\}\right)=-\frac{1}{|U|}\sum_{x\in U}\mathrm{lb}\frac{\left|[x]_j^{k,\delta_j^k}\cap[x]_d\right|^2}{\left|[x]_j^{k,\delta_j^k}\right|\left|[x]_d\right|}\text{。}$$

**例 1.** 表 1 是一个 GMHDS, 其中: $a_1$ 是数值型属性, 有 3 个尺度等级; $a_2$ 是符号型属性, 有 2 个尺度等级; $a_3$ 是集合值型属性, 有 2 个尺度等级。H, K, M, L 是属性 $a_2$ 在第 2 个尺度等级下的属性值; T, S, R 是属性 $a_2$ 在第 1 个尺度等级下的属性值。表 1 中属性尺度之间的信息粒度变换表示如下:

$$g_1^{3,2}(z)=r,\ \text{若}z\in[r-0.005,r+0.005)\text{;}$$

$$g_1^{2,1}(t)=s,\ \text{若}t\in[s-0.05,s+0.05)\text{;}$$

$$g_2^{2,1}(u)=\begin{cases}\mathrm{T},\ \text{若}u\in\{\mathrm{H},\mathrm{K}\}\text{;}\\ \mathrm{S},\ \text{若}u\in\{\mathrm{M}\}\text{;}\\ \mathrm{R},\ \text{若}u\in\{\mathrm{L}\}\text{;}\end{cases}$$

$$g_3^{2,1}(v)=\begin{cases}1,\ \text{若}v\in\{1\}\text{;}\\ 2,\ \text{若}v\in\{2\}\text{;}\\ 3,\ \text{若}v\in\{3,4\}\text{;}\\ 4,\ \text{若}v\in\{5,6\}\text{。}\end{cases}$$

在表 1 中, 设 $\delta_1^3=0.1$, 则 $\delta^{\boldsymbol{I}}=\{\delta_1^3\}(\boldsymbol{I}=(3,2,2))$。令 $\delta_1^{k-1}=h_1^{k,k-1}$, 其中 $k=\{3,2\}$。故可得

$$[x_1]_A^{I,\delta^{\boldsymbol{I}}}=\{x_1\},\ [x_2]_A^{I,\delta^{\boldsymbol{I}}}=\{x_2\},\ [x_3]_A^{I,\delta^{\boldsymbol{I}}}=\{x_3\}\text{,}$$

$$[x_4]_A^{I,\delta^{\boldsymbol{I}}}=[x_8]_A^{I,\delta^{\boldsymbol{I}}}=\{x_4,x_8\},\ [x_5]_A^{I,\delta^{\boldsymbol{I}}}=\{x_5\}\text{,}$$

$$[x_6]_A^{I,\delta^{\boldsymbol{I}}}=\{x_6\},\ [x_7]_A^{I,\delta^{\boldsymbol{I}}}=\{x_7\},\ [x_9]_A^{I,\delta^{\boldsymbol{I}}}=\{x_9\}\text{,}$$

$$[x_{10}]_A^{I,\delta^{\boldsymbol{I}}}=\{x_{10}\},\ NEG_A^{I,\delta^{\boldsymbol{I}}}(\mathcal{D})=\varnothing\text{,}$$

$$BN_A^{I,\delta^{\boldsymbol{I}}}(\mathcal{D})=\{x_4,x_8\}\text{,}$$

$$POS_A^{I,\delta^{\boldsymbol{I}}}(\mathcal{D})=\{x_1,x_2,x_3,x_5,x_6,x_7,x_9,x_{10}\}\text{。}$$

**Table 1 A GMHDS**

**表 1 1 个 GMHDS**

| 对象集 $U$ | $a_1$ | | | $a_2$ | | $a_3$ | | 决策属性 $d$ |
|---|---|---|---|---|---|---|---|---|
| | $a_1^3$ | $\xrightarrow{g_1^{3,2}}$ $a_1^2$ | $\xrightarrow{g_1^{2,1}}$ $a_1^1$ | $a_2^2$ | $\xrightarrow{g_2^{2,1}}$ $a_2^1$ | $a_3^2$ | $\xrightarrow{g_3^{2,1}}$ $a_3^1$ | |
| $x_1$ | 0.379 | 0.38 | 0.4 | H | T | {5} | {4} | 1 |
| $x_2$ | 0.768 | 0.77 | 0.8 | M | S | {6} | {4} | 1 |
| $x_3$ | 0.173 | 0.17 | 0.2 | M | S | {2} | {2} | 2 |
| $x_4$ | 0.176 | 0.18 | 0.2 | L | R | {1,2} | {1,2} | 1 |
| $x_5$ | 0.486 | 0.49 | 0.5 | M | S | {2,4} | {2,3} | 2 |
| $x_6$ | 0.643 | 0.64 | 0.6 | K | T | {1} | {1} | 1 |
| $x_7$ | 0.243 | 0.24 | 0.2 | H | T | {1} | {1} | 1 |
| $x_8$ | 0.227 | 0.23 | 0.2 | L | R | {2} | {2} | 2 |
| $x_9$ | 0.334 | 0.33 | 0.3 | M | S | {1,5,6} | {1,4} | 2 |
| $x_{10}$ | 0.576 | 0.58 | 0.6 | H | T | {6} | {4} | 1 |

根据文献 [36], 计算表 1 的一个 POSC, 即按照表 1 中属性原本的顺序, 可找到一个 POSC, $P=(1,1,0)$。同时, 计算决策属性关于每个条件属性的所有尺度的条件熵如表 2 所示。

**Table 2 Conditional Entropies of Decision $d$ with Each Attribute Under Each Scale Level**

**表 2 决策 $d$ 关于每个属性在每个尺度下的条件熵**

| $U$ | $a_1$ | | | $a_2$ | | $a_3$ | |
|---|---|---|---|---|---|---|---|
| | $a_1^3$ | $a_1^2$ | $a_1^1$ | $a_2^2$ | $a_2^1$ | $a_3^2$ | $a_3^1$ |
| $H_1^{\delta_j^k}$ | 0.849 2 | 0.764 4 | 0.902 5 | 0.524 5 | 0.524 5 | 0.827 5 | 0.735 0 |
| $H_2^{\delta_j^k}$ | 2.636 5 | 2.846 4 | 3.502 9 | 1.698 0 | 1.698 0 | 3.477 4 | 3.699 9 |
| $H_3^{\delta_j^k}$ | 2.600 2 | 2.356 9 | 2.378 0 | 1.924 5 | 1.600 0 | 2.027 5 | 1.659 5 |

从表 2 中发现, 决策属性关于 $a_1^1$ 的条件熵在 $H_1^{\delta_j^k}$ 和 $H_3^{\delta_j^k}$ 都要大于关于 $a_2^1$ 和 $a_3^1$ 的。这表明在条件熵 $H_1^{\delta_j^k}$ 和 $H_3^{\delta_j^k}$ 下, 最优尺度组合所对应的属性子集中的属性 $a_1$ 与决策属性的相关性要低于被删去的属性 $a_3$。

### 3.1 基于相对相关属性子集的最优尺度组合

基于 3 种条件熵, 本节度量条件属性与决策属性的相对相关性, 从而定义相对相关属性子集, 并讨论其相关性质, 然后给出基于相对相关属性子集的最

优尺度组合的定义。

通过定义 13 中的 3 种条件熵，在一个 GMHDS $S=(U,A\cup\{d\})$中，给出条件属性$a_j^k\in A$与决策属性$d$的相关性的定义。

**定义 14.** 设$S=(U,A\cup\{d\})$是一个 GMHDS，$\boldsymbol{I}=(I_1,I_2,\cdots,I_m)$，$\mathcal{D}=U/R_d$。对于$a_j^k\in A$，$k\in\{1,2,\cdots,I_j\}$和$\delta_j^k$，$a_j^k$与决策$d$的相关性定义为

$$Rele_\beta^{\delta_j^k,\mathcal{D}}\left(a_j^k\right)=\frac{1}{1+H_\beta^{\delta_j^k}\left(\mathcal{D}|\left\{a_j^k\right\}\right)},\tag{17}$$

其中$\beta=1,2,3$。

定义 14 所给出的相关性定义确保了相关性的数值$Rele_\beta^{\delta_j^k,\mathcal{D}}\left(a_j^k\right)\in(0,1]$，其中$a_j^k\in A$，$k\in\{1,2,\cdots,I_j\}$，$\beta\in\{1,2,3\}$。但是由于不同数据中条件属性与决策属性的相关性的分布情况不同，有些数据中条件属性与决策属性的相关性普遍较高，有些则普遍较低。为了使阈值的确定更合理，同时提高模型的泛化性，再对相关性进行标准化。易知当$a_j^k\in A$的尺度越强时，尺度中所包含的信息越详细。因此，考虑标准化在最强尺度下条件属性与决策属性的相关性。

**定义 15.** 设$S=(U,A\cup\{d\})$是一个 GMHDS，$\boldsymbol{I}=(I_1,I_2,\cdots,I_m)$，$\mathcal{D}=U/R_d$。$a_j^{I_j}\in A^{\boldsymbol{I}}$与$d$的相对相关性定义为：

$$\overline{Rele}_\beta^{\delta^{\boldsymbol{I}},\mathcal{D}}\left(a_j^{I_j}\right)=\frac{Rele_\beta^{\delta_j^{I_j},\mathcal{D}}\left(a_j^{I_j}\right)-\min\limits_{a_j^{I_j}\in A^{\boldsymbol{I}}}Rele_\beta^{\delta_j^{I_j},\mathcal{D}}\left(a_j^{I_j}\right)}{\max\limits_{a_j^{I_j}\in A^{\boldsymbol{I}}}Rele_\beta^{\delta_j^{I_j},\mathcal{D}}\left(a_j^{I_j}\right)-\min\limits_{a_j^{I_j}\in A^{\boldsymbol{I}}}Rele_\beta^{\delta_j^{I_j},\mathcal{D}}\left(a_j^{I_j}\right)},\tag{18}$$

其中$\beta=1,2,3$。

显然$\overline{Rele}_\beta^{\delta^{\boldsymbol{I}},\mathcal{D}}\left(a_j^{I_j}\right)\in[0,1]$，其中$a_j^{I_j}\in A^{\boldsymbol{I}}$，$\beta\in\{1,2,3\}$。对于$a_j^{I_j}\in A^{\boldsymbol{I}}$，当$\overline{Rele}_\beta^{\delta^{\boldsymbol{I}},\mathcal{D}}\left(a_j^{I_j}\right)=0$时，表明$a_j^{I_j}$是$A^{\boldsymbol{I}}$中与$d$相关性最低的属性，而当$\overline{Rele}_\beta^{\delta^{\boldsymbol{I}},\mathcal{D}}\left(a_j^{I_j}\right)=1$时，表明$a_j^{I_j}$是$A^{\boldsymbol{I}}$中与$d$相关性最高的属性。

然后，引入一个阈值$\theta\in[0,0.5]$，当且仅当条件属性与决策属性的相对相关性小于$\theta$时，删去该条件属性。

**定义 16.** 设$S=(U,A\cup\{d\})$是一个 GMHDS，$\boldsymbol{I}=(I_1,I_2,\cdots,I_m)$，$\mathcal{D}=U/R_d$。对于$\theta\in[0,0.5]$和$\beta$，关于决策属性$d$的相对相关属性子集$A_{\beta,\theta}^{\delta^{\boldsymbol{I}},\mathcal{D}}$定义为

$$A_{\beta,\theta}^{\delta^{\boldsymbol{I}},\mathcal{D}}=\left\{a_j\in A\left|\overline{Rele}_\beta^{\delta^{\boldsymbol{I}},\mathcal{D}}\left(a_j^{I_j}\right)\geqslant\theta\right.\right\},\tag{19}$$

其中$\beta=1,2,3$。需要注意的是，$A_{\beta,\theta}^{\delta^{\boldsymbol{I}},\mathcal{D}}$是不包含尺度信息的条件属性集合。当$\theta=0$时，$A_{\beta,\theta}^{\delta^{\boldsymbol{I}},\mathcal{D}}=A$。

通过式(19)的相对相关属性子集，给出关于相对相关属性子集的正域最优尺度组合的定义。

**定义 17.** 设$S=(U,A\cup\{d\})$是一个 GMHDS，$\boldsymbol{I}=(I_1,I_2,\cdots,I_m)$，$\mathcal{D}=U/R_d$。对于$L\in\mathcal{L}$，$\theta\in[0,0.5]$和$\beta\in\{1,2,3\}$，记$A_{\beta,\theta}^{\delta^{\boldsymbol{I}},\mathcal{D}}=A'$，若$POS_{A'}^{L,\delta^L}(\mathcal{D})=POS_{A'}^{\boldsymbol{I},\delta^{\boldsymbol{I}}}(\mathcal{D})$，则称$S^L$关于$S$和$A'$是正域协调的，否则$S^L$关于$S$和$A'$不是正域协调的。若$S^L$关于$S$和$A'$是正域协调的，且对于满足$L^*\prec L$的任意$L^*\in\mathcal{L}$，$S^{L^*}$关于$S$和$A'$不是正域协调的，则称$L$是$S$的一个关于相对相关属性子集$A'$的正域最优尺度组合(relatively relevant attribute subset positive region optimal scale combination, rras-POSC)。

### 3.2 定向逐步搜索算法

本节给出关于相对相关属性子集的正域最优尺度组合的定向搜索算法。

根据定义 15 所定义的相对相关性，在筛选最优尺度组合时，优先考虑删除与决策的相对相关性较低的属性，从而尽可能保留与决策相对相关性更高的属性。

在一个 GMHDS $S=(U,A\cup\{d\})$中，$\boldsymbol{I}=(I_1,I_2,\cdots,I_m)$，$\mathcal{D}=U/R_d$，对于$\theta\in[0,0.5]$和$\beta\in\{1,2,3\}$，记$\left|A_{\beta,\theta}^{\delta^{\boldsymbol{I}},\mathcal{D}}\right|=\kappa$，那么$A_{\beta,\theta}^{\delta^{\boldsymbol{I}},\mathcal{D}}$中属性的筛选顺序为$order_{\beta,\theta}^{\delta^{\boldsymbol{I}},\mathcal{D}}=(a_{o_1},a_{o_2},\cdots,a_{o_\kappa})$，其中每个元素$a_{o_r}$定义为

$$a_{o_r}=\arg\min_{a_j\in C}\left\{\overline{Rele}_\beta^{\delta^{\boldsymbol{I}},\mathcal{D}}\left(a_j^{I_j}\right)\middle|C=A_{\beta,\theta}^{\delta^{\boldsymbol{I}},\mathcal{D}}-\{a_{o_1},a_{o_2},\cdots,a_{o_{r-1}}\}\right\},$$

其中$r\in\{1,2,\cdots,\kappa\}$，$\arg\min\limits_x f(x)$表示使$f(x)$取得最小值的参数$x$。

首先根据文献 [36] 的算法 1，基于属性子集$B\subseteq A$，给出在 GMHDS 的一个尺度组合下计算正域的算法。与文献 [36] 中算法 1 的时间复杂度类似，本文算法 1 的时间复杂度为$O(n^2)$，其中$n$是对象个数。

**算法 1.** 基于属性子集计算一个 GMHDS 在一个尺度组合下的正域。

输入：一个 GMHDS $S=(U,A\cup\{d\})$，属性子集$B$，$L=(l_1,l_2,\cdots,l_m)$，在尺度组合下的邻域半径集$\delta^L=\left\{\delta_j^{l_j}\middle|a_j^{l_j}\in A^{\mathrm{num},L}\right\}$；

输出：正域$POS_B^{L,\delta^L}(\mathcal{D})$。

① 初始化：$POS_B^{L,\delta^L}(\mathcal{D})\leftarrow\varnothing$，$n\leftarrow|U|$；

② for $i\leftarrow1:n$ do

③ 计算$x_i$对应的$[x_i]_B^{L,\delta^L}$和$[x_i]_d$；

④ if $[x_i]_B^{L,\delta^L}\subseteq[x_i]_d$ then

⑤ $POS_B^{L,\delta^L}(\mathcal{D})\leftarrow POS_B^{L,\delta^L}(\mathcal{D})\cup\{x_i\}$；

⑥ end if

⑦ end for

⑧ return $POS_B^{L,\delta^L}(\mathcal{D})$。

然后，在一个 GMHDS 中给出搜索一个关于$A_{\beta,\theta}^{\delta^{\boldsymbol{I}},\mathcal{D}}$的 POSC 的定向逐步搜索算法，即算法 2。在算法 2 中，行②~⑦计算数值型属性下的邻域半径集和属

性在最强尺度下与决策的相关性，当$\beta=1,3$时其时间复杂度为$O\left(\sum_{j=1}^{m}I_j n^2\right)$，当$\beta=2$时其时间复杂度为$O\left(\sum_{j=1}^{m}I_j n^2+mn\gamma\right)$，其中$\gamma$是决策类的基数。而行⑩的时间复杂度为$O(n^2)$。行⑫~㉒的时间复杂度为$O\left(\sum_{j=1}^{\kappa}I_j n^2\right)$，其中在$\theta=0$时参数$\kappa$的值最大，即$\kappa=m$。综上所示，当$\beta=1,3$时，整个算法 2 的时间复杂度为$O\left(\sum_{j=1}^{m}I_j n^2\right)$，当$\beta=2$时，整个算法 2 的时间复杂度为$O\left(\sum_{j=1}^{m}I_j n^2+mn\gamma\right)$。若记$\tau=\max_j\{I_j\}$，则算法 2 的时间复杂度上界为$O(mn^2\tau+mn\gamma)$。

**算法 2.** 定向逐步搜索 GMHDS 的一个 rras-POSC。

输入：一个 GMHDS $S=(U,A\cup\{d\})$，$\boldsymbol{I}=(I_1,I_2,\cdots,I_m)$，在最强尺度组合下的邻域半径集$\delta^{\boldsymbol{I}}=\left\{\delta_j^{I_j}|a_j^{I_j}\in A^{\text{num},\boldsymbol{I}}\right\}$，$\beta\in\{1,2,3\}$，$\theta\in[0,0.5]$；

输出：一个 rras-POSC $AP=(ap_1,ap_2,\cdots,ap_m)$。

① 初始化：$AP\leftarrow\boldsymbol{I}$，$m\leftarrow|A|$，$n\leftarrow|U|$；
② for $j\leftarrow1:m$ do
③ if $a_j\in A^{\text{num}}$ then
④ 计算$\delta_j^k(k\in\{1,2,\cdots,I_j-1\})$；
⑤ end if
⑥ 计算$Rele_\beta^{\delta_j^{I_j},\mathcal{D}}\left(a_j^{I_j}\right)$；
⑦ end for
⑧ 计算$\overline{Rele}_\beta^{\delta^{\boldsymbol{I}},\mathcal{D}}\left(a_j^{I_j}\right)$；
⑨ 计算$A'=A_{\beta,\theta}^{\delta^{\boldsymbol{I}},\mathcal{D}}$和$order_{\beta,\theta}^{\delta^{\boldsymbol{I}},\mathcal{D}}=(a_{o_1},a_{o_2},\cdots,a_{o_\kappa})$，其中$\kappa=|A'|$；
⑩ 根据算法 1 计算$POS_{A'}^{\boldsymbol{I},\delta^{\boldsymbol{I}}}(\mathcal{D})$；
⑪ $ap_j\leftarrow0$，对于所有$A-A'$；
⑫ for $j\leftarrow1:\kappa$ do
⑬ $ap_{o_j}\leftarrow0$；
⑭ while $ap_{o_j}<I_{o_j}$ do
⑮ 根据算法 1 计算$POS_{A'}^{AP,\delta^{AP}}(\mathcal{D})$；
⑯ if $\left|POS_{A'}^{AP,\delta^{AP}}(\mathcal{D})\right|=\left|POS_{A'}^{\boldsymbol{I},\delta^{\boldsymbol{I}}}(\mathcal{D})\right|$ then
⑰ break;
⑱ else
⑲ $ap_{o_j}\leftarrow ap_{o_j}+1$；
⑳ end if
㉑ end while
㉒ end for
㉓ return $AP$。

**例 2.** 当$\beta=1$和$\theta=0.2$时，记定向搜索的基于相对相关属性子集的正域最优尺度组合为$AP$，则算法 2 搜索的步骤为：

1)初始化$AP\leftarrow(3,2,2)$，计算得

$$\overline{Rele}_1^{\delta^{\boldsymbol{I}},\mathcal{D}}\left(a_1^3\right)=0,\ \overline{Rele}_1^{\delta^{\boldsymbol{I}},\mathcal{D}}\left(a_2^2\right)=1,$$

$$\overline{Rele}_1^{\delta^{\boldsymbol{I}},\mathcal{D}}\left(a_3^2\right)\approx0.944\,4,\ A_{1,0.2}^{\delta^{\boldsymbol{I}},\mathcal{D}}=\{a_2,a_3\},$$

记$A_{1,0.2}^{\delta^{\boldsymbol{I}},\mathcal{D}}=A'$，则$AP\leftarrow(0,2,2)$，$order_{1,0.2}^{\delta^{\boldsymbol{I}},\mathcal{D}}=(a_3,a_2)$，而$|POS_{A'}^{AP,\delta^{AP}}(\mathcal{D})|=6$。

2)对于属性$a_3$，限制其尺度等级为 0，即$AP\leftarrow(0,2,0)$。由于$|POS_{A'}^{AP,\delta^{AP}}(\mathcal{D})|=4\neq6$。因此，$a_3$的尺度等级增加 1，即$AP\leftarrow(0,2,1)$。此时$|POS_{A'}^{AP,\delta^{AP}}(\mathcal{D})|=6$，则$AP$是第 2 步骤的结果。

3)对于属性$a_2$，限制其尺度等级为 0，即$AP\leftarrow(0,0,1)$。由于$|POS_{A'}^{AP,\delta^{AP}}(\mathcal{D})|=0\neq6$，则$a_2$的尺度等级增加 1，即$AP\leftarrow(0,1,1)$。而此时$|POS_{A'}^{AP,\delta^{AP}}(\mathcal{D})|=6$，即$AP$是第 3 步骤的结果。

4)找到一个 rras-POSC$(AP)$，$AP=(0,1,1)$。定向搜索的基于相对相关属性子集的正域最优尺度组合$AP=(0,1,1)$对应的属性集为$\left\{a_2^1,a_3^1\right\}$。根据表 2，与例 1 中的正域最优尺度组合$P=(1,1,0)$对应的属性子集相比，$AP$对应的属性子集中的属性在条件熵$H_1^{\delta_j^k}$和$H_3^{\delta_j^k}$下与决策$d$的相关性更高。

## 4 数值实验与参数分析

在本节中，我们将本文算法应用于 UCI 数据库上的数据集，采用 10 折交叉验证法来评估模型的性能，并与其他现有属性约简算法进行比较分析。首先，本文在 UCI 中获得了 9 个数据集，详情如表 3 所

**Table 3 The Details of the Selected Datasets**

**表 3 选定数据集的详细信息**

| 编号 | 数据 | 对象个数 | 属性个数 | | 不完备个数 | 决策类数 |
|---|---|---|---|---|---|---|
| | | | 数值 | 符号 | | |
| D1 | Breast | 569 | 30 | 0 | 0 | 2 |
| D2 | Crx | 690 | 6 | 9 | 7 | 2 |
| D3 | German | 1 000 | 7 | 13 | 0 | 2 |
| D4 | Hcvdat0 | 615 | 11 | 1 | 5 | 5 |
| D5 | Hepatitis | 155 | 7 | 12 | 15 | 2 |
| D6 | Ionosphere | 351 | 33 | 0 | 0 | 2 |
| D7 | Iris | 150 | 4 | 0 | 0 | 3 |
| D8 | Musk | 476 | 166 | 0 | 0 | 2 |
| D9 | Tcga | 839 | 1 | 22 | 0 | 2 |

注：“不完备个数”表示所有属性中存在未知的属性值的属性个数。

示。其次，根据 Wang 等人[36]的方法将 UCI 上的数据集转化为 GMHDS。然后，从 GMHDS 中提取基于 3 个相对相关属性子集的 POSC，并选择分类效果最好的一个正域最优尺度组合作为模型的结果。因此，本文提出的模型称为基于相对相关属性子集的 POSC 的定向逐步搜索模型，简称 da-POSC。

实验的 PC 硬件配置为第 13 代 Intel® Core™ i7-13700H CPU@2.40 GHz, 16 GB RAM)，实验在 Windows 11 操作系统下进行，算法使用 MATLAB R2023a 实现。

为了进一步分析本文提出模型 da-POSC 的有效性，选取了 5 种属性约简的模型，将 da-POSC 与 5 种模型进行对比，分析模型在所选数据集上的分类性能。对比的模型分别为：FSDNE[48], BARA[42], FBCDS[11], POSC[36], CEOSC[36]，其中 FSDNE 是指 Sun 等人[48]提出的基于邻域熵的属性约简模型，BARA 是指 Zhang 等人[42]提出的基于证据理论处理混合信息系统的属性约简模型，FBCDS 是指 Huang 等人[11]提出的基于多粒度粗糙模糊集的属性约简模型，而 POSC 和 CEOSC 是 Wang 等人[36]给出基于正域和条件熵的最优尺度组合。前 3 个模型的参数取值范围为 $[0,1]$，其步长为 0.1；而后 2 个模型的参数 $\delta=[0.1,1.8]$，步长也是 0.1。da-POSC 模型中有 3 个参数，分别是 $\delta$, $\theta$, $\beta$，其中根据 Wang 等人[36]的实验，本文中 $\delta$ 的取值在每个数据集上固定，选取在 KNN 和 NB 分类器上都合适的取值，如表 4 所示。而其余参数的取值范围是 $\theta\in[0,0.5]$，步长为 0.05，$\beta\in\{1,2,3\}$。

**Table 4 Classification Performance of Six Algorithms Under KNN and NB Classifiers**

**表 4 6 个算法在 KNN 和 NB 分类器下的分类性能**

| 编号 | $\delta$ | 分类器 | 准确率/% | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | FSDNE | BARA | FBCDS | POSC | CEOSC | da-POSC |
| D1 | 1.0 | KNN | 91.07±3.95 | 95.89±2.67 | 96.61±1.97 | 96.43±2.38 | 96.25±2.14 | **97.32±2.10** |
| | | NB | 91.61±5.12 | **94.82±1.56** | 92.86±3.76 | 93.57±2.10 | 93.57±2.10 | 94.11±2.92 |
| D2 | 0.7 | KNN | 53.97±8.06 | 85.59±16.98 | 83.24±16.05 | **87.21±11.58** | 86.62±12.77 | 86.91±12.58 |
| | | NB | 67.94±11.90 | 78.82±10.29 | 75.88±5.33 | 82.94±11.58 | 81.62±11.89 | **83.24±15.70** |
| D3 | 1.3 | KNN | 69.50±2.17 | **75.70±3.33** | 72.10±2.13 | 72.30±3.40 | 72.30±3.40 | 73.80±2.44 |
| | | NB | 70.50±2.76 | 74.40±6.28 | 74.70±5.52 | 74.60±3.53 | 74.80±3.55 | **75.10±4.65** |
| D4 | 0.4 | KNN | 90.50±1.58 | 93.67±2.92 | 93.00±3.02 | 93.17±2.99 | 93.17±2.66 | **93.83±1.58** |
| | | NB | 90.83±1.18 | 90.67±2.51 | 90.83±2.39 | 91.00±2.51 | 89.83±3.80 | **92.00±2.58** |
| D5 | 0.1 | KNN | 89.33±7.83 | 90.67±7.83 | **91.33±7.73** | **91.33±6.32** | **91.33±6.32** | **91.33±6.32** |
| | | NB | 82.67±6.44 | 88.00±5.26 | 89.33±10.04 | 89.33±7.82 | 89.33±7.82 | **90.00±6.48** |
| D6 | 0.6 | KNN | 78.82±10.81 | 90.88±7.13 | 87.35±7.97 | 90.59±4.56 | 90.88±5.96 | **92.06±4.81** |
| | | NB | 81.76±7.44 | 84.12±8.68 | 76.76±7.40 | 86.18±6.66 | 83.82±10.21 | **87.94±7.26** |
| D7 | 0.4 | KNN | 96.00±4.66 | 94.67±6.13 | 96.00±6.44 | 94.67±4.22 | 94.67±4.22 | **97.33±3.44** |
| | | NB | 96.00±4.66 | **96.67±3.51** | 96.00±3.44 | 95.33±3.22 | 95.33±4.50 | **96.67±3.51** |
| D8 | 1.8 | KNN | 59.78±7.69 | 73.26±10.40 | 66.62±13.09 | 79.35±11.51 | 77.61±10.50 | **80.65±12.01** |
| | | NB | 63.70±12.08 | 71.74±13.79 | 58.91±10.17 | 70.22±10.55 | 70.22±10.45 | **75.65±13.04** |
| D9 | 0.4 | KNN | 70.36±4.69 | **85.78±2.77** | 71.93±4.10 | 84.34±1.97 | 83.98±1.71 | 84.94±2.85 |
| | | NB | 73.61±5.16 | 86.99±2.46 | 75.66±4.32 | 86.63±2.23 | 86.63±2.23 | **87.47±2.74** |

注：“±”前的数值表示准确率的均值，“±”后的数值表示准确率的标准差，黑体数值表示最优结果。

表 4 中呈现了在 KNN 和 NB 分类器下，6 个算法在 9 个数据集上计算得到的准确率。根据表 4，在 KNN 分类器下，da-POSC 在 6 个数据集上取得了最好的分类效果；而在 NB 分类器下，da-POSC 在 8 个数据集上取得了最好的分类效果。综上所述，发现 da-POSC 的分类性能在大多数情况下要比其他 5 种对比模型的分类性能更好。

下面对模型的参数 $\theta$ 和 $\beta$ 进行分析。当 $\theta$ 和 $\beta$ 变化时，da-POSC 的准确率在大多数的参数组合上表现了一定的稳定性，但在图 1(b)中，当 $\beta=1$ 且 $\theta$ 较大时，da-POSC 在 KNN 分类器下有明显的下降。这是由于 $\theta$ 是控制相对相关属性子集大小的阈值，相对相关性低于 $\theta$ 的条件属性将被删去。因此，当 $\theta$ 较大时，相对相关属性子集内的属性个数相对较少。而 D2 中含

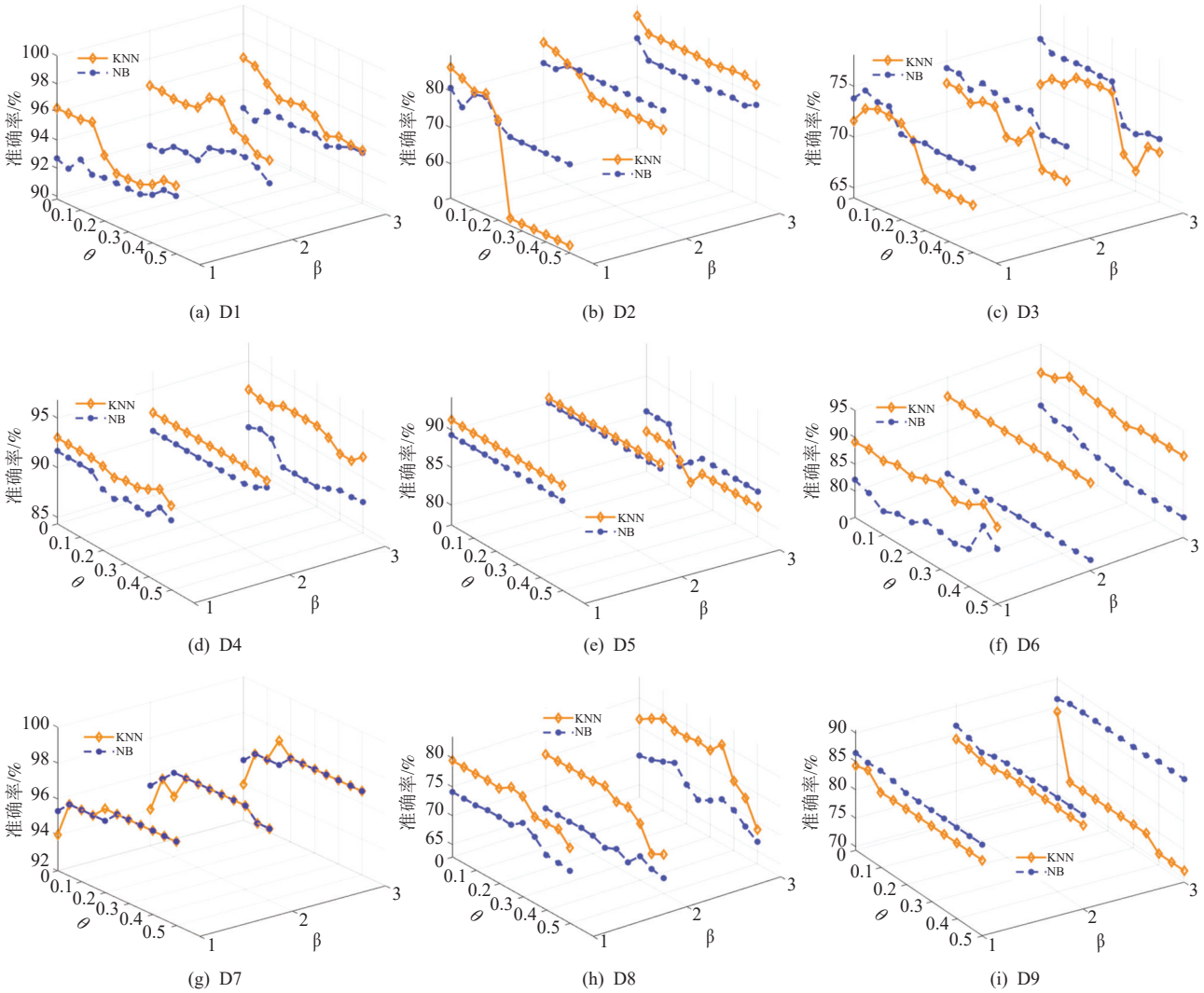



Fig. 1 Classification performance of da-POSC under KNN and NB classifiers when the parameters $\theta$ and $\beta$ vary

图 1 在分类器 KNN 和 NB 下 da-POSC 在参数 $\theta$ 和 $\beta$ 变化时的分类性能

有较多的缺省值, 且缺省值集中在少数属性中, 故数据的不确定性高。因此, 该数据集对参数$\theta$取值的变动更为敏感。

根据图 2, 在大多数情况下, da-POSC 的搜索时间在参数变化时呈现持续下降的趋势, 而在$\beta=2$时, 其搜索时间有先升高后降低的趋势(图 2(d))。

结合图 1 和图 2 的内容, 还可以判断哪个相关性不适合度量数据集上条件属性与决策属性的相关性。例如, 在图 1(i)和图 2(i)中, 当$\delta=3$时, da-POSC 的准确率在 KNN 分类器下突然下降, 但计算最优尺度组合的时间同时也快速下降, 说明相关性的度量并不合理。这是由于 D9 中绝大多数的属性都是符号型, 每个对象在符号属性下的相似类和在决策下的等价类的基数都较大, 从而造成在各个属性上计算得到的$H_3^{\delta t}$相差不大, 导致模型容易删去与决策的相关性较高的属性。同时符号型属性中距离要么为 0, 要么为 1, 与 KNN 分类器的分类原理契合度不高。因此, 该相对相关性不能和 KNN 分类器结合在 D9 中使用。

接下来, 通过准确率(accuracy, ACC)、精确率(precision, PRE)、召回率(recall rate, RE)、假阳性率(false positive rate, FPR)、宏-$F1$(macro- $F1$, MAR- $F1$)和曲线下面积(area under the curve, AUC)这 6 种分类指标, 利用 Friedman 检验[49]和 Nemenyi 后续检验[50]来统计 6 个算法的分类性能。若实验在$T$个数据集上分析$p$个算法的分类性能, Friedman 检验公式如式(20)所示。

$$\begin{aligned} \chi_{\mathrm{F}}^2 &= \frac{12T}{p(p+1)}\left(\sum_{i=1}^{p} r_i^2 - \frac{p(p+1)^2}{4}\right), \\ F_{\mathrm{F}} &= \frac{(T-1)\chi_{\mathrm{F}}^2}{T(p-1)-\chi_{\mathrm{F}}^2}, \end{aligned} \tag{20}$$

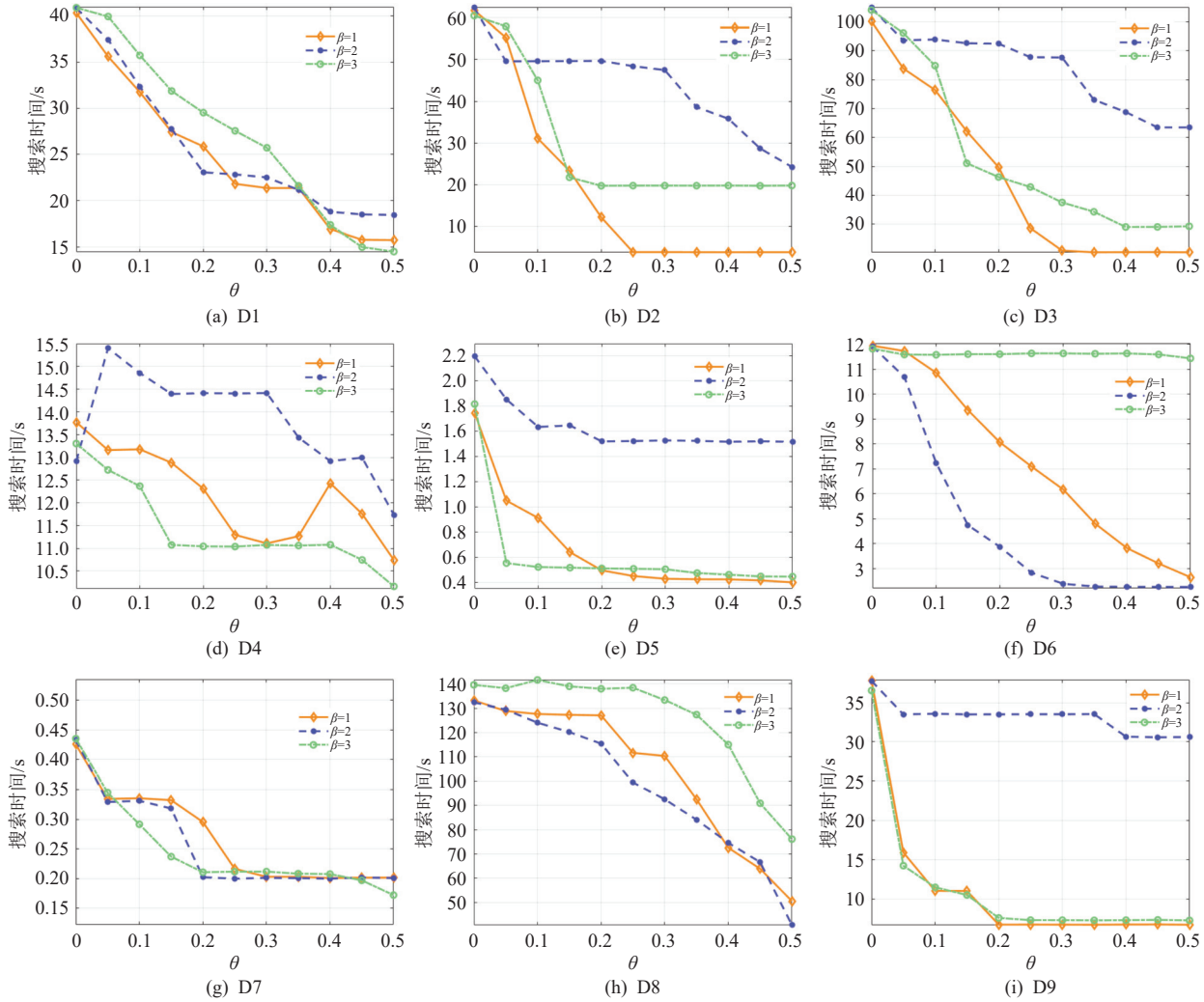



Fig. 2 Searching time of da-POSC when the parameters $\theta$ and $\beta$ vary

图 2 da-POSC 在参数 $\theta$ 和 $\beta$ 变化时的搜索时间

其中$r_i$是第$i$个算法分类指标的平均排名，$i=1,2,\cdots,p$，$\chi_F^2$服从自由度为$p-1$和$(p-1)(T-1)$的F-分布。在Nemenyi后续检验中，$CD_\alpha$是用来检验任意2个算法之间存在差异的临界值，在显著性水平为$\alpha$时，由式(21)计算。

$$CD_\alpha = c_\alpha \sqrt{\frac{p(p+1)}{6T}}, \tag{21}$$

其中$c_\alpha$是临界列表值。

对于$p$个算法，假设算法的性能值都具有相同的分布，若$F_F$超过了临界值$F_\alpha(p-1,(p-1)(T-1))$，则原假设不成立。若其中2个算法的平均序值之差超出了临界值$CD_\alpha$，则说明这2个算法的性能具有显著性差异。设$\alpha=0.05$，故6种分类指标的临界差异图如图3和图4所示，其中$c_\alpha=2.85$。因此，可以计算$CD_\alpha=2.513\,5$。根据图3和图4，da-POSC都位于临界差异图的最右侧。故在6种分类指标下，本文提出的da-POSC的分类效果优于比较的5个模型。

## 5 总结与展望

在多尺度数据分析中，由于传统的算法在搜索最优尺度组合时没有考虑属性的筛选顺序，同时遍历搜索全部属性的尺度，存在搜索效率不高且浪费计算资源的问题。本文提出了一种新的基于相对相关属性子集的最优尺度组合定向搜索算法da-POSC。首先，利用了3种条件熵在GMHDS中定义属性与决策的相对相关性，刻画了多尺度数据与决策的相关性，进而筛选了相对相关属性子集。然后，在GMHDS中，给出了基于相对相关属性子集的正域最优尺度组合的概念。最后，利用属性与决策的相对相关性，设计了一种新的定向搜索最优尺度组合的逐步搜索算法da-POSC。数据实验表明，在大多数情况下，da-POSC在6种分类指标上要优于比较的5种模型。同时，选取合适的相对相关性参数后，可以在保持分类

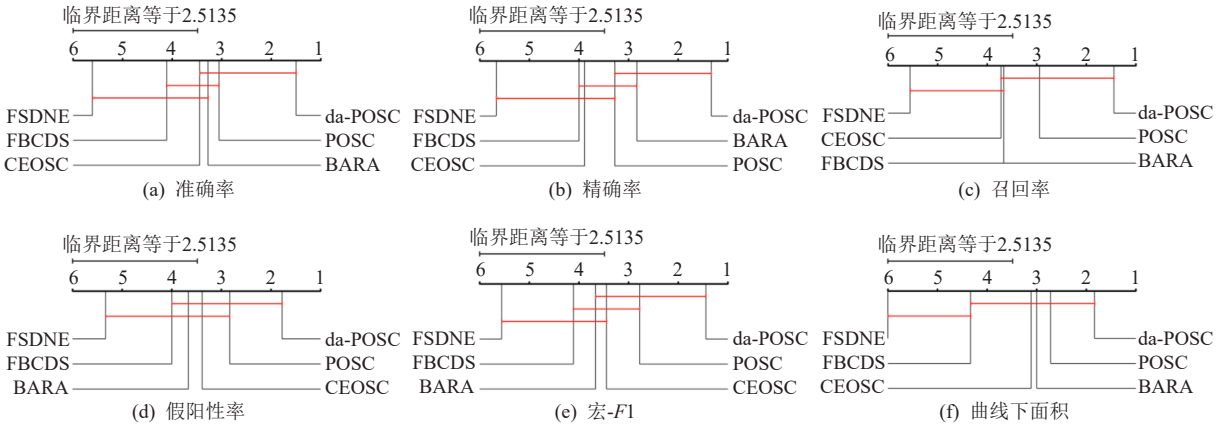



Fig. 3 Critical difference figures of six algorithms under KNN classifier on six classification indexes

图 3 6 种算法在 KNN 分类器下 6 种分类指标的临界差异图

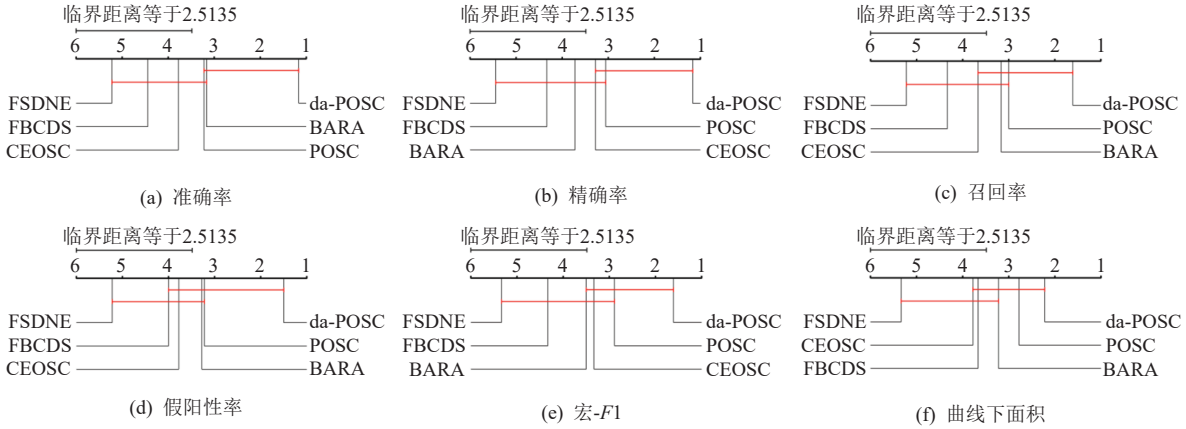



Fig. 4 Critical difference figures of six algorithms under NB classifier on six classification indexes

图 4 6 种算法在 NB 分类器下 6 种分类指标的临界差异图

效果的同时，显著降低计算时间。未来在本文研究的基础上，可进一步研究动态混合数据的多尺度知识更新表示与动态多尺度知识获取问题。

**作者贡献声明**：王蕾晰撰写论文，提出并实现算法思路和实验方法；吴伟志、王熙照提出指导意见并修改论文。

## 参 考 文 献

[1] 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): 111−127

[2] Liang Jiye, Qian Yuhua, Li Deyu, et al. Theory and method of granular computing for big data mining[J]. 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[J]. 中国科学: 信息科学, 2015, 45(11): 1355−1369)

[3] Xu Ji, Wang Guoyin, Yu Hong. Review of big data processing based on granular computing[J]. Chinese Journal of Computers, 2015, 38(8): 1497−1517 (in Chinese)
(徐计, 王国胤, 于洪. 基于粒计算的大数据处理[J]. 计算机学报, 2015, 38(8): 1497−1517)

[4] Miao Duoqian, Zhang Qinghua, Qian Yuhua, et al. From human intelligence to machine implementation model: Theories and applications based on granular computing[J]. CAAI Transactions on Intelligent Systems, 2016, 11(6): 743−757 (in Chinese)
(苗夺谦, 张清华, 钱宇华, 等. 从人类智能到机器实现模型——粒计算 理论与方法[J]. 智能系统学报, 2016, 11(6): 743−757)

[5] Zhang Qinghua, Wang Yutai, Zhao Fan, et al. Multi-granularity computing framework for complex problem solving[J]. SCIENTIA SINICA Informationis, 2025, 55(5): 1122−1139 (in Chinese)
(张清华, 王宇泰, 赵凡, 等. 复杂问题求解的多粒度计算框架[J]. 中国科学: 信息科学, 2025, 55(5): 1122−1139)

[6] Qin Jindong, Martínez L, Pedrycz W, et al. An overview of granular computing in decision-making: Extensions, applications, and challenges[J]. Information Fusion, 2023, 98: 101833

[7] Guo Doudou, Xu Weihua. R-FCCL: An approach of fuzzy-based concept-cognitive learning with robustness for high-dimensional data[J]. Journal of Computer Research and Development, 2025, 62(2): 383−396 (in Chinese)
(郭豆豆, 徐伟华. R-FCCL: 一种面向高维数据的稳健模糊概念认

知学习方法[J]. 计算机研究与发展, 2025, 62(2): 383−396)

[8] Behzadidoost R, Mahan F, Izadkhah H. Granular computing-based deep learning for text classification[J]. Information Sciences, 2024, 652: 119746

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

[10] Campagner A, Ciucci D, Denœux T. Belief functions and rough sets: Survey and new insights[J]. International Journal of Approximate Reasoning, 2022, 143: 192−215

[11] Huang Zhehuang, Li Jinjin. Covering based multi-granulation rough fuzzy sets with applications to feature selection[J]. Expert Systems with Applications, 2024, 238: 121908

[12] Jiang Jiefang, Zhang Xianyong, Yuan Zhong. Feature selection for classification with Spearman's rank correlation coefficient-based self-information in divergence-based fuzzy rough sets[J]. Expert Systems with Applications, 2024, 249: 123633

[13] Qu Kanglin, Gao Pan, Dai Qun, et al. Attribute reduction using self-information uncertainty measures in optimistic neighborhood extreme-granulation rough set[J]. Information Sciences, 2025, 686: 121340

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

[15] Wang Jianzhou, Wang Kang, Li Zhiwu, et al. Short-term power load forecasting system based on rough set, information granule and multi-objective optimization[J]. Applied Soft Computing, 2023, 146: 110692

[16] Wu Weizhi, Leung Y. Theory and applications of granular labelled partitions in multi-scale decision tables[J]. Information Sciences, 2011, 181(18): 3878−3897

[17] Chen Yingsheng, Li Jinhai, Li Jinjin, et al. A further study on optimal scale selection in dynamic multi-scale decision information systems based on sequential three-way decisions[J]. International Journal of Machine Learning and Cybernetics, 2022, 13(5): 1505−1515

[18] Huang Xianfeng, Zhan Jianming, Xu Zeshui, et al. A prospect-regret theory-based three-way decision model with intuitionistic fuzzy numbers under incomplete multi-scale decision information systems[J]. Expert Systems with Applications, 2023, 214: 119144

[19] Li Feng, Hu Baoqing. A new approach of optimal scale selection to multi-scale decision tables[J]. Information Sciences, 2017, 381: 193−208

[20] Li Runkang, Yang Jilin, Zhang Xianyong. Optimal scale selection based on three-way decisions with decision-theoretic rough sets in multi-scale set-valued decision tables[J]. International Journal of Machine Learning and Cybernetics, 2023, 14(11): 3719−3736

[21] Yao Yibin, Zhan Jianming, Xu Zeshui, et al. Heterogeneous group consensus method with three-way decisions based on fuzzy social networks and manipulation behaviors under multiscale information systems[J]. IEEE Transactions on Fuzzy Systems, 2024, 32(9): 5122−5135

[22] Zhang Xiaoyan, Feng Zihan. Feature selection based on contradictory state sequence for multi-scale interval valued decision table[J]. Information Sciences, 2024, 677: 120926

[23] Zhang Qinghua, Yang Ying, Cheng Yunlong, et al. Information fusion for multi-scale data: Survey and challenges[J]. Information Fusion, 2023, 100: 101954

[24] Cheng Yunlong, Zhang Qinghua, Wang Guoyin, et al. Optimal scale selection and attribute reduction in multi-scale decision tables based on three-way decision[J]. Information Sciences, 2020, 541: 36−59

[25] Li Feng, Hu Baoqing, Wang Jun. Stepwise optimal scale selection for multi-scale decision tables via attribute significance[J]. Knowledge-Based Systems, 2017, 129: 4−16

[26] Mou Qiong, Cheng Yunlong. Optimal scale combinations selection for incomplete generalized multi-scale decision systems[C]//Proc of the 2021 IEEE Int Conf on Progress in Informatics and Computing. Piscataway, NJ: IEEE, 2021: 270−275

[27] Xie Zhenhuang, Wu Weizhi, Wang Leixi, et al. Entropy based optimal scale selection and attribute reduction in multi-scale interval-set decision tables[J]. International Journal of Machine Learning and Cybernetics, 2024, 15: 3005−3026

[28] Zhang Qinghua, Cheng Yunlong, Zhao Fan, et al. Optimal scale combination selection integrating three-way decision with Hasse diagram[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(8): 3675−3689

[29] Zhan Jianming, Deng Jiang, Xu Zeshui. A three-way decision methodology with regret theory via triangular fuzzy numbers in incomplete multiscale decision information systems[J]. IEEE Transactions on Fuzzy Systems, 2023, 31(8): 2773−2787

[30] Chen Baiyang, Li Yongxiang, Peng Dezhong, et al. Fusing multi-scale fuzzy information to detect outliers[J]. Information Fusion, 2024, 103: 102133

[31] Hu Qinghua, Xie Zongxia, Yu Daren. Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation[J]. Pattern Recognition, 2007, 40(12): 3509−3521

[32] Hu Qinghua, Yu Daren, Xie Zongxia. Information-preserving hybrid data reduction based on fuzzy-rough techniques[J]. Pattern Recognition Letters, 2006, 27(5): 414−423

[33] Kumar A, Prasad P S. Scalable feature subset selection with fuzzy rough sets and fuzzy min-max neural network in hybrid decision system[J]. IEEE Transactions on Fuzzy Systems, 2024, 33(2): 669−679

[34] Sun Lin, Wang Lanying, Ding Weiping, et al. Neighborhood multi-granulation rough sets-based attribute reduction using Lebesgue and entropy measures in incomplete neighborhood decision systems[J]. Knowledge-Based Systems, 2020, 192: 105373

[35] Zhang Qinli, Song Yan, Peng Yichun, et al. Attribute reduction algorithms with an anti-noise mechanism for hybrid data based on fuzzy evidence theory[J]. Engineering Applications of Artificial Intelligence, 2024, 129: 107659

[36] Wang Leixi, Wu Weizhi, Xie Zhenhuang, et al. Optimal scale combination selection in generalized multi-scale hybrid decision systems[J]. Information Sciences, 2025, 689: 121429

[37] Yang Ying, Zhang Qinghua, Zhao Fan, et al. Optimal scale combination selection based on genetic algorithm in generalized multi-scale decision systems for classification[J]. Information Sciences,

2025, 693: 121685

[38] Hu Jun, Chen Yan, Zhang Qinghua, et al. Optimal scale selection for generalized multi-scale set-valued decision systems[J]. Journal of Computer Research and Development, 2022, 59(9): 2027−2038 (in Chinese)
(胡军, 陈艳, 张清华, 等. 广义多尺度集值决策系统最优尺度选择[J]. 计算机研究与发展, 2022, 59(9): 2027−2038)

[39] Wang Pei, He Jiali, Li Zhaowei. Attribute reduction for hybrid data based on fuzzy rough iterative computation model[J]. Information Sciences, 2023, 632: 555−575

[40] Yuan Zhong, Chen Hongmei, Li Tianrui, et al. Unsupervised attribute reduction for mixed data based on fuzzy rough sets[J]. Information Sciences, 2021, 572: 67−87

[41] Liao Shujiao, Lin Yidong, Li Jinjin, et al. Attribute-scale selection for hybrid data with test cost constraint: The approach and uncertainty measures[J]. International Journal of Intelligent Systems, 2022, 37(6): 3297−3333

[42] Zhang Qinli, Qu Liangdong, Li Zhaowen. Attribute reduction based on D-S evidence theory in a hybrid information system[J]. International Journal of Approximate Reasoning, 2022, 148: 202−234

[43] Shannon C E. A mathematical theory of communication[J]. The Bell System Technical Journal, 1948, 27(3): 379−423

[44] Wan Jihong, Chen Hongmei, Yuan Zhong, et al. A novel hybrid feature selection method considering feature interaction in neighborhood rough set[J]. Knowledge-Based Systems, 2021, 227: 107167

[45] Xu Weihua, Yuan Ziting, Liu Zheng. Feature selection for unbalanced distribution hybrid data based on *k*-nearest neighborhood rough set[J]. IEEE Transactions on Artificial Intelligence, 2024, 5(1): 229−243

[46] Dai Jianhua, Xu Qing, Wang Wentao, et al. Conditional entropy for incomplete decision systems and its application in data mining[J]. International Journal of General Systems, 2012, 41(7): 713−728

[47] Dai Jianhua, Wang Wentao, Xu Qing. An uncertainty measure for incomplete decision tables and its applications[J]. IEEE Transactions on Cybernetics, 2012, 43(4): 1277−1289

[48] Sun Lin, Zhang Xiaoyu, Qian Yuhua, et al. Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification[J]. Information Sciences, 2019, 502: 18−41

[49] Friedman M. A comparison of alternative tests of significance for the problem of *m* rankings[J]. The Annals of Mathematical Statistics, 1940, 11(1): 86−92

[50] Demšar J. Statistical comparisons of classifiers over multiple data sets[J]. The Journal of Machine Learning Research, 2006, 7: 1−30



**Wang Leixi**, born in 1999. Master candidate. Her research interests include rough set and granular computing.

**王蕾晰**, 1999 年生。硕士研究生。主要研究方向为粗糙集、粒计算。



**Wu Weizhi**, born in 1964. PhD, professor, PhD supervisor. Senior member of CCF. His research interests include rough set, granular computing, data mining, and artificial intelligence.

**吴伟志**, 1964 年生。博士, 教授, 博士生导师。CCF 高级会员。主要研究方向为粗糙集、粒计算、数据挖掘、人工智能。



**Wang Xizhao**, born in 1963. PhD, professor, PhD supervisor. His main research interest includes uncertainty modeling and machine learning for big data.

**王熙照**, 1963 年生。博士, 教授, 博士生导师。主要研究方向为大数据的不确定性建模和机器学习。