文章编号: 1001-0920(2018)08-1481-08 DOI: 10.13195/j.kzyjc.2017.0485

基于多元异构不确定性案例学习的广义区间灰数熵权聚类模型

张 秦†, 方志耕, 蔡佳佳, 刘思峰

(南京航空航天大学 经济与管理学院, 南京 211106)

摘 要: 现实生活中多数聚类对象具有多元异构不确定性特征,表现为对象聚类指标体系异构化以及对象信息具有多元不确定性特点,而现有的不确定性多属性聚类决策方法对此类对象的聚类研究具有局限性. 为此,针对聚类问题,首先,根据聚类对象多元不确定性信息的特点,提出广义区间灰数的概念,证明多元不确定性信息可统一用广义区间灰数进行表征;然后,结合极大熵思想,构建基于多元异构不确定性案例学习的广义区间灰数熵权配置模型,通过对对象相关的历史案例进行充分学习,测算各层指标的广义区间灰数熵权,以此确定各指标的聚类权重,再结合广义区间灰数的白化权函数对对象的新案例进行聚类分析;最后,通过案例研究验证所提出聚类模型的合理性和可行性.

关键词: 聚类; 多元异构不确定性; 广义区间灰数; 案例学习; 熵权配置

中图分类号: C394 **文献标志码:** A

Generalized interval grey entropy-weight clustering model based on multiple heterogeneous uncertainty cases study

ZHANG Qin†, FANG Zhi-geng, CAI Jia-jia, LIU Si-feng

(College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: Most of the clustering objects have the characteristic of multiple heterogeneous uncertainty in real life, which is presented via the heterogeneity of the clustering indicator system of objects and multi-uncertainty of the objects information. However, the present uncertainty multiple attribute clustering decision methods have limitations to research the clustering of these kinds of objects. Therefore, firstly the concept of generalized interval grey number is proposed according to the characteristics of multi-uncertainty information. It is proved that multi-uncertainty information can be all represented with generalized interval grey number in some specific conditions. Then, the rule of maximum entropy is introduced, the generalized interval grey entropy-weight allocation model based on multiple heterogeneous uncertainty cases study is built. The entropy-weight of multiple level is calculated through studying the cases about the objects, and then the weight of every indexes is got. Furthermore, the white function of generalized interval grey is used to analyze the clustering for the new cases of objects. Finally, the case study verifies the rationality and feasibility of the proposed clustering model.

Keywords: clustering; multiple heterogeneous uncertainty; generalized interval grey number; cases study; entropy-weight allocation

0 引 言

由于测量技术的固有限制,客观事物的复杂性以及人类认知的局限性等,现实生活中的大多数决策都为不确定性多属性决策. 其中针对不确定性多属性聚类决策的研究一直是学者们关注的热点和难点,通过对大量文献的梳理发现,目前的不确定性多属性聚类决策方法主要包括两类:基于计算机算法的聚类方法以及基于模型的聚类方法[1-2].

基于计算机算法的聚类方法是根据不确定性数据挖掘的准则,利用模式识别和机器学习等技术把有限的无标签的待聚类对象划分为多个“相似的”簇集,将相似度高的对象聚为一类,典型的聚类算法有

收稿日期: 2017-04-22; **修回日期:** 2017-06-21.

基金项目: 基本科研业务费科研基地创新基金项目(NP20150036, NP20150037).

作者简介: 张秦(1992–), 男, 博士生, 从事管理科学与工程的研究; 方志耕(1962–), 男, 教授, 博士, 从事管理科学与工程等研究.

†通讯作者. E-mail: 2468279002@qq.com

K-近邻算法[3]、模糊C-均值算法[4-5]、高维不确定数据高效聚类(HDUDEC)算法[6]、神经网络算法[7],以及不确定-围绕中心点划分(U-PAM)和不确定测量-围绕中点划分(UM-PAM)算法[8]等. 基于模型的聚类方法是从分析对象中提取多个特征指标,并以指标属性为准则构建分析对象的评价指标体系,通过建模的思想对聚类对象进行全面客观地评价. 其中,运用最广泛的不确定性多属性聚类模型主要是灰色聚类分析模型. 自从灰色系统理论[9]出世以来,灰色聚类模型作为重要的聚类分析方法,已被广泛应用于经济、军事、企业评估,以及交通等领域[10-14]. 但是,随着人们对事物的深度认知,越来越多的学者对传统的灰色聚类方法进行了改进,例如利用熵权的思想建立灰色熵权聚类模型[15-17],更加有效地解决了实际生活中存在的问题. 然而,不确定性多属性聚类决策问题是很复杂的,主要是因为多数聚类对象具有多元异构不确定性特征,具体表现在对象聚类决策的复杂性决定了对象聚类指标体系呈现异构的特点,即聚类对象指标之间存在着上下层级的关系或是并列的关系,且对象信息具有多元不确定性特点,例如随机性、灰色性、模糊性等.

就现有研究而言,传统的不确定性多属性聚类决策的局限性主要集中于3个方面: 1) 基于计算机算法的聚类方法是建立在大数据、大样本挖掘的基础上,其主要的缺点是如果样本量少于百万级,则聚类效果往往很差,而现实生活中的对象样本量多数不满足这个条件; 2) 基于模型的聚类方法虽然没有样本数量上的限定要求,但大都集中于单层特征指标的聚类问题,不适合研究多层次指标体系的对象聚类问题; 3) 基于模型的聚类方法只考虑了单一的不确定性方法对属性指标的评价,而多元异构不确定性对象的属性指标往往会同时具有灰色性、模糊性、随机性等. 因此,传统的不确定性多属性聚类方法对研究多元异构不确定性对象聚类问题具有一定的局限性.

为解决多元异构不确定性对象的聚类问题,本文首先针对多元不确定性信息下各指标属性具有随机性、灰色性以及模糊性等特点,提出广义区间灰数的概念以及严谨的证明过程,证明多种不确定数均可用其表征,并设计相应的广义区间灰数的白化权函数; 然后,引入极大熵准则,构造基于多元异构不确定性案例学习的广义区间灰数熵权配置模型,通过对对象相关的历史案例进行充分学习,计算出各层指标的广义区间灰数熵权,以此得到各指标的权重,再结合广义区间灰数的白化权函数对新案例进行聚类分析,得出聚类结果. 为解决多元异构不确定性多属性聚类问题提供一种新方法.

1 广义区间灰数表征及其白化权函数设计

在现实生活中,某些事物之所以能够归属到同一类中,是因为其自然本质属性具有较高的相似性,因此人们往往通过评价事物属性之间的相似程度,将评价对象划分到不同的类别中去. 然而,评价对象往往具有多元异构不确定性特点,使得聚类指标存在一定的层次结构以及指标属性的评价值可能同时包括灰数、模糊数、区间值模糊数、概率数等多种不确定性数,因此对象的聚类研究面临着较大的挑战. 为此,在对对象进行聚类研究时,可根据影响对象的各种重要因素,将其分成不同的特征指标以及属性指标,而一个待聚类对象含有若干个一级特征指标$y_i(i = 1,2,\cdots,n)$,每个一级指标又含有多个二级指标$y_{ij}(j = 1,2,\cdots,h)$,以此类推,直到最后一层指标就是属性指标$y_{i\cdots k}(k = 1,2,\cdots,m)$,其中需要人为评价的是属性指标,如图1所示. 由于每个层次的各个指标对对象间相似性的影响程度不同,应对指标赋以相应的权重w以合理地区分其重要程度.

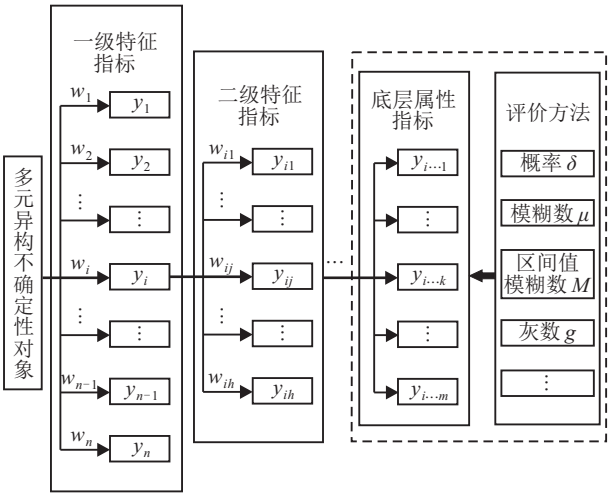

图1 多元异构不确定性对象指标体系

多元异构不确定性对象的底层指标评价规则众多,而传统的评价方法所研究的对象单一,例如模糊理论、灰色系统或者概率论等,其未对所有的不确定性对象进行综合研究,因此针对多元异构不确定性对象的评价,导致传统的评价方法失真. 本文经过深入研究发现,在某种条件下,灰集均可与概率数集、模糊集、区间值模糊集等相互转换,于是提出广义区间灰数的概念,证明在某种条件下,多种不确定数均可用其表征,并结合传统的白化权函数,设计广义区间灰数的白化权函数.

1.1 广义区间灰数表征

定义1 令$g^{\pm}\in R$为一个未知实数,且属于某个开闭区间的并集[18],即

$$g^{\pm}\in\bigcup_{i=1}^{n}[a_i,b_i], \tag{1}$$

则$g^{\pm}$被称为灰数,$g^{-}=\inf\limits_{a_i\in g^{\pm}}a_i$为$g^{\pm}$的下限,$g^{+}=\sup\limits_{b_i\in g^{\pm}}b_i$为$g^{\pm}$的上限. 其中:$i=1,2,\cdots,n;0<n<\infty$,且$n$为整数;$a_i,b_i\in R,b_{i-1}\leqslant a_i\leqslant b_i\leqslant a_{i+1}$. 对于任一区间,有$[a_i,b_i]\subseteq\bigcup\limits_{i=1}^{n}[a_i,b_i]$.

定义2 令值域$D\subset R,g^{\pm}\in D,d_{\min},d_{\max}\in R$是$D$的最小值与最大值,则$g^{\pm}$的灰度[19]表示为

$$g^0=\frac{|g^{+}-g^{-}|}{|d_{\max}-d_{\min}|}. \tag{2}$$

当且仅当$g^{+}=g^{-},d_{\max}-d_{\min}\neq 0$时,$g^0=0$,此时$g^{\pm}$为概率数;当$g^{+}=d_{\max},g^{-}=d_{\min},d_{\max}-d_{\min}\neq 0$时,$g^0=1,g^{\pm}$为黑数.

定义3 令U为有限论域,对于集合$A\subseteq U$,如果任一x关于A的特征函数值能表示为灰数$g_A^{\pm}(x)\in\bigcup\limits_{i=1}^{n}[a_i,b_i]\in D[0,1]^{\pm}$,且有

$$\delta_A:U\longrightarrow D[0,1]^{\pm}, \tag{3}$$

则A为灰集[19],其中$D[0,1]^{\pm}$指在区间$[0,1]$的所有灰数的集合.

定义4 概率数是单一确定的数,令U为有限论域,对于集合$A\subseteq U$,如果任一x关于A的特征函数值能表示为单一的概率数$v\in[0,1]$,且有

$$\delta_A:U\longrightarrow[0,1], \tag{4}$$

则A是一个概率数集[20].

定理1 对于灰集$A\subseteq U,g_A^{\pm}(x)\in D[0,1]^{\pm}$,若$A$的灰度$g_A^0=0$,则灰集$A$为概率数集.

证明　灰集A的灰度为$g_A^0=0$,且有

$$g_A^{-}(x)=g_A^{+}(x). \tag{5}$$

由定义2可知,对于任一x,均有$g_A^{\pm}(x)$为概率数,且$g_A^{\pm}(x)\in[0,1]\subseteq D[0,1]^{\pm}$,所以灰集$A$为概率数集. □

定义5 令U表示论域,在此论域中的模糊集A可定义为序对集合[21]

$$A=\{\langle x,\mu_A(x)\rangle:x\in U\}. \tag{6}$$

其中:$\mu_A(x):U\longrightarrow[0,1]$为模糊集$A$的隶属度函数,$\mu_A(x)$为$x$属于$A$的隶属程度. 隶属函数值可以是$[0,1]$之间的任意值,表示模糊的概念以及集合的边界.

定理2 对于灰集$A\subseteq U,g_A^{\pm}(x)\in D[0,1]^{\pm}$,$g_A^0=0$,则灰集$A$为模糊集.

证明　首先证明模糊集是概率数集. 令δ为概率数集A的特征函数,则有

$$\delta_A:U\longrightarrow[0,1], \tag{7}$$

得到概率数集,当δ为隶属度函数时,令$\mu=\delta$,则有

$$\mu_A:U\longrightarrow[0,1]. \tag{8}$$

得到模糊集A,当将隶属度函数μ作为特征函数时,亦可得到概率数集. 又由定理1可知,当$g_A^{\pm}(x)\in D[0,1]^{\pm}$,且$g_A^0=0$时,灰集$A$为概率数,所以在此条件下,灰集$A$为模糊集. □

定义6 令$D[0,1]$为区间$[0,1]$的所有闭合子区间的集合,U为论域,x为论域U中的一个元素,则论域U下的区间值模糊集[22]可用集合A表示为

$$A=\{\langle x,M_A(x)\rangle:x\in U\}, \tag{9}$$

其中$M_A(x):U\longrightarrow D[0,1]$. 因此,单个元素的隶属度用区间形式表示,而不是单个值.

定理3 对于任一元素$x\in U$,有$g_A^{\pm}(x)$均为连续灰数,并且区间表示有明确边界,确切指未知的数,则集合A为一个区间值模糊集[18].

定义7 对于某一属性指标集合A,各个特征指标x属性值$\otimes_A(x)$的类型包括概率数、模糊数、区间值模糊数(连续区间)以及灰数等,且属性值值域均为$D[0,1]$,则这些特征指标属性值可用广义区间灰数统一表示为

$$G(\otimes_A(x))\in\{\delta_A(x)\bigcup\mu_A(x)\bigcup M_A(x)\bigcup g_A^{\pm}(x)\}\in\bigcup_{i=1}^{n}[a_i,b_i]\in D[0,1]^{\pm}. \tag{10}$$

其中:$i=1,2,\cdots,n$为第i个特征指标;A为特征指标属性集合;$\delta_A(x)$为概率数;$\mu_A(x)$为模糊数;$M_A(x)$为区间值模糊;$g_A^{\pm}(x)$为灰数,a_i,b_i分别为第i个特征指标属性值的下限与上限;$D[0,1]^{\pm}$为在区间$[0,1]$的所有广义区间灰数的集合.

1.2 广义区间灰数的白化权函数设计

定义8 设有n个具有多元异构不确定性特点的待聚类对象,m个具有多元不确定性特征的聚类指标,s个现有的广义灰类,则通过第i个待聚类对象关于j聚类指标的广义区间灰数观测值,将第i个待聚类对象纳入第k个广义灰类中,这个过程称为广义区间灰数的广义灰色聚类.

定义9 设$A\subseteq R,\forall a,b\in R,G(\otimes)\in[a,b]\in D[0,1]^{\pm}$为广义区间灰数,则$[A]$表示集合$A$中所有广义区间灰数的集合.

定义10[23] 设$f:A\to R$是A上的连续函数,$F:[A]\to R,G(\otimes)\in[a,b]\in D[0,1]^{\pm}$,则$f$关于$[A]$的积分均值函数表达形式为$F(\otimes)=\dfrac{\int_a^b f(t)\mathrm{d}t}{b-a}$.

定义11 已知$f_j^k(\cdot)$为指标属性值是实数的白化权函数,若属性值为广义区间灰数,即$G(\otimes)\in[a,$

$b] \in D[0,1]^{\pm}$,则定义广义区间灰数的白化权函数为$f_j^k(\cdot)$所对应的积分均值函数

$$F(\otimes)=\begin{cases}\dfrac{\int_a^b f_j^k(x)\mathrm{d}x}{b-a},\ a\neq b;\\ f_j^k(x),\ a=b.\end{cases}\tag{11}$$

其中:j为聚类指标,k为灰类.

命题1 当$a=b$时,广义区间灰数的上限测度白化权函数为

$$f_j^k(x)=\begin{cases}0,\ x<x_j^k(1);\\ \dfrac{x-x_j^k(1)}{x_j^k(2)-x_j^k(1)},\ x\in[x_j^k(1),x_j^k(2)];\\ 1,\ x\geqslant x_j^k(2).\end{cases}\tag{12}$$

当$a\neq b$时,其上限测度白化权函数为

$$F_j^k(\otimes)=$$

$$\begin{cases}0,\ b\leqslant x_j^k(1);\\ \dfrac{(b-x_j^k(1))^2}{2(b-a)(x_j^k(2)-x_j^k(1))},\ a\leqslant x_j^k(1)<b\,x_j^k(2);\\ \dfrac{a+b-2x_j^k(1)}{2(x_j^k(2)-x_j^k(1))},x_j^k(1)<a<b\leqslant x_j^k(2);\\ \dfrac{(x_j^k(2)-a)[x_j^k(2)+a-2x_j^k(1)+2(b-x_j^k(2))]}{x_j^k(2)-x_j^k(1)},\\ x_j^k(1)\leqslant a<x_j^k(2)<b;\\ 1,\ x_j^k(2)\leqslant a.\end{cases}\tag{13}$$

同理,可按照定义11设计出典型的、适中测度以及下限测度的广义区间灰数白化权函数,这里不再一一详解.

2 基于多元异构不确定性案例学习的广义区间灰数熵权配置模型

本文首先对多元异构不确定性对象案例进行有效学习和分析,挖掘其本质规律,发现多元异构不确定性对象的聚类指标权重与案例聚类情况有着十分紧密的关系. 所以本文引入极大熵思想[24],在已知部分信息的基础上,认为权重熵值达到最大且满足约束条件时所得到的权重值出现的可能性最大,以此构建基于多元异构不确定性案例学习的广义区间灰数熵权配置模型,再结合灰数“核”[25],得到各指标的权重,具体步骤如下:

Step 1 目标函数对于一个具有由不同评价方面、要素和因素形成的多层次指标的聚类对象,其指标熵权$\overline{w_i}$是指标i在其所在的评价层指标集合中所占的广义区间比重,是一个随机变量,并具有一定的不确定性. 按照极大熵思想,构造广义区间灰数熵权配置模型的目标函数

$$\max P=[p_1(\overline{w^{\otimes}})\cdots p_i(\overline{w^{\otimes}})\cdots p_n(\overline{w^{\otimes}})].\tag{14}$$

其中:$p_i(\overline{w^{\otimes}})=-\sum\limits_{k=1}^{n_i}\overline{w_k^{\otimes}}\lg\overline{w_k^{\otimes}}$,$\overline{w^{\otimes}}$为各个特征指标的广义区间灰数熵权,$n_i$为在第$i$指标层的特征指标个数,$p_i(\overline{w^{\otimes}})$为第$i$指标层所对应的函数.

Step 2 约束条件.

定义12 对于任一广义区间灰数$\otimes_i=[a_i,b_i]$,$a_i\leqslant b_i,i=1,2,\cdots,n$,均可表征为$\otimes_i=a_i+c_i\cdot\gamma_i$,此形式称为广义标准灰数. 其中:$a_i$称为$\otimes_i$的广义白部;$c_i\cdot\gamma_i$称为$\otimes_i$的广义灰部,在广义灰部$c_i\cdot\gamma_i$中,$c_i=(b_i-a_i)$称为广义灰系数,$\gamma_i$称为广义单位灰数.

定义13 令多元异构不确定对象的底层指标集A的广义区间灰数向量为$\overrightarrow{\otimes_{\boldsymbol{A}}}=[\otimes_A(1),\otimes_A(2),\cdots,\otimes_A(n)]$. 其中:$\otimes_A(i)=a_i+c_i\gamma_i,\gamma_i\in[0,1],i=1,2,\cdots,n$. 集合$B$的广义区间灰数向量为$\overrightarrow{\otimes_{\boldsymbol{B}}}=[\otimes_B(1),\otimes_B(2),\cdots,\otimes_B(n)]$. 其中:$\otimes_B(i)=b_i+d_i\gamma_i,\gamma_i\in[0,1],i=1,2,\cdots,n$. 则广义区间灰数向量$A$与$B$之间的相似度为

$$\otimes_{\mathrm{sim}}^{(A,B)}=\cos(\overrightarrow{\otimes_{\boldsymbol{A}}},\overrightarrow{\otimes_{\boldsymbol{B}}})=\frac{\sum\limits_{i=1}^{n}[\otimes_A(i)\times\otimes_B(i)]}{\sqrt{\sum\limits_{i=1}^{n}\otimes_A(i)^2}\times\sqrt{\sum\limits_{i=1}^{n}\otimes_B(i)^2}}.\tag{15}$$

定义14 若对象有n层特征指标,其底层指标的数量为k_l且属性值均为广义区间灰数,则$\otimes_{\mathrm{sim}}^{l}$为底层指标集合广义区间灰数向量的相似度,对应的广义区间灰数熵权为$\overline{w_{h\cdots l}^{\otimes}}$,并设$i$层指标体系第$j$个指标的熵权为$\overline{w_{h\cdots j}^{\otimes}}(j=1,2,\cdots,k_i)$,则两个实际案例$S_1,S_2$的相似度用下式计算求出:

$$\otimes_{\mathrm{sim}}^{(S_1,S_2)}=\sum_{h=1}^{k_n}\overline{w_h^{\otimes}}\cdots\sum_{j=1}^{k_i}\overline{w_{h\cdots j}^{\otimes}}\cdots\sum_{l=1}^{k_l}\otimes_{\mathrm{sim}}^{l}\overline{w_{h\cdots l}^{\otimes}}.\tag{16}$$

1) 当广义灰数集合$\gamma=[\gamma_1,\cdots,\gamma_i,\cdots,\gamma_n]$($\gamma_i\in[0,1],i=1,2,\cdots,n$)赋予某一具体数值时,有$\gamma^*=[\gamma_1^*,\cdots,\gamma_i^*,\cdots,\gamma_n^*]$,且指标体系各层的熵权和为1,即

$$\sum_{h=1}^{k_n}\overline{w_h^{\otimes}}=1,\sum_{j=1}^{k_i}\overline{w_{h\cdots j}^{\otimes}}=1,\cdots,\sum_{l=1}^{k_l}\overline{w_{h\cdots l}^{\otimes}}=1.\tag{17}$$

2) 若有几起案例属于同一类,则它们分别与其他任意案例的相似度必定会小于同一类案例间的相似度,即$\otimes_{\mathrm{sim}}^{(S_i,S_j)}>\otimes_{\mathrm{sim}}^{(S_i,S_k)}$,对象$S_i$与$S_j$属于类别$L_\alpha$,

对象S_k属类别L_β,则以$\otimes_{\rm sim}^{(S_i,S_j)} > \otimes_{\rm sim}^{(S_i,S_k)}$作为目标函数的约束条件之一.

Step 3　模型构建.

在上述目标函数与约束条件的基础上,建立广义区间灰数熵权配置模型:

$$\max P = [p_1(\overline{w^{\otimes}}),\cdots,p_i(\overline{w^{\otimes}}),\cdots,p_n(\overline{w^{\otimes}})].$$
$$\begin{cases}\displaystyle\sum_{h=1}^{k_n}\overline{w_h^{\otimes}}=1,\sum_{j=1}^{k_i}\overline{w_{h\cdots j}^{\otimes}}=1,\cdots,\sum_{l=1}^{k_l}\overline{w_{h\cdots l}^{\otimes}}=1;\\ \otimes_{\rm sim}^{(S_i,S_j)} > \otimes_{\rm sim}^{(S_i,S_k)};\\ S_i,S_j\in L_\alpha,\ S_k\in L_\beta;\\ \gamma^*=[\gamma_1^*,\cdots,\gamma_i^*,\cdots,\gamma_n^*],\ \gamma_i^*\in[0,1].\end{cases} \tag{18}$$

利用灰数“核”的思想,使各层次指标的权重为$w^{\otimes}=1/2(\overline{w^{\otimes L}}+\overline{w^{\otimes U}})$,其中$\overline{w^{\otimes L}}$为广义区间灰数熵权$\overline{w^{\otimes}}$的下限,$\overline{w^{\otimes U}}$为广义区间灰数熵权$\overline{w^{\otimes}}$的上限.

3　基于广义区间灰数熵权配置模型的聚类分析

根据以上研究内容,本文可进一步得到基于广义区间灰数熵权的聚类分析算法,主要分为3个步骤.

Step 1　根据对象i历史案例关于底层属性指标l的评价值$\otimes_{il}(i=1,2,\cdots,n;\ l=1,2,\cdots,m)$,利用广义区间灰数熵权配置模型确定各个指标的熵权$\overline{w^{\otimes}}$,可分为4个小步骤.

1) 根据定义12,利用广义标准灰数$\otimes=a+c\gamma$对评价值$\otimes_{il}$进行标准化处理;

2) 根据定义13和定义14计算出对象的两两相似度$\otimes_{\rm sim}^{(S_{i1},S_{i2})}$;

3) 利用广义区间灰数熵权配置模型求解出各层指标的广义区间灰数熵权$\overline{w^{\otimes}}\in[\overline{w^{\otimes L}},\overline{w^{\otimes U}}]$;

4) 取$w^{\otimes}=1/2(\overline{w^{\otimes L}}+\overline{w^{\otimes U}})$作为各层指标的聚类权重.

Step 2　根据命题1中的广义区间灰数的白化权函数$F(\otimes)$,计算求得新案例底层属性指标的白化权函数值$F_l^k(\otimes)(i=1,2,\cdots,n;\ l=1,2,\cdots,m)$,再根据Step 1得到的聚类权重$w^{\otimes}$,计算广义灰色变权聚类系数

$$\sigma_i^k=\sum_{h=1}^{k_n}w_h^{\otimes}\cdots\sum_{j=1}^{k_i}w_{h\cdots j}^{\otimes}\cdots\sum_{l=1}^{k_l}F_l^k(\otimes_{il})\cdot w_{h\cdots l}^{\otimes}. \tag{19}$$

Step 3　若$\max\limits_{1\leqslant k\leqslant s}\sigma_i^k=\sigma_i^{k*}$,则判定对象$i$属于灰类$k^*$.

4　案例研究

本文以安徽某市公安局对入室盗窃类、抢劫类以及扒窃类等刑事犯罪串并案为例,刑事犯罪串并案是典型的多元异构不确定性对象,其具有多层次指标体系以及底层属性指标具有多元不确定性特点. 本案例研究选择3类6起案件集作为研究对象,首先建立串并案指标体系,然后利用广义标准灰数确定底层指标的属性值,接着运用广义区间灰数熵权配置模型对各个指标进行权重分配,最后建立广义灰色变权聚类系数矩阵对新案件作聚类分析.

4.1　串并案指标体系建立与底层指标属性值确定

选取入室盗窃类、抢劫类以及扒窃类案件6起$S=\{S_1,S_2,S_3,S_4,S_5,S_6\}$,其中$S_1,S_2$同属于入室盗窃类,$S_3,S_4$同属于抢劫类,$S_5,S_6$同属于扒窃类,串并案指标体系如图2所示.

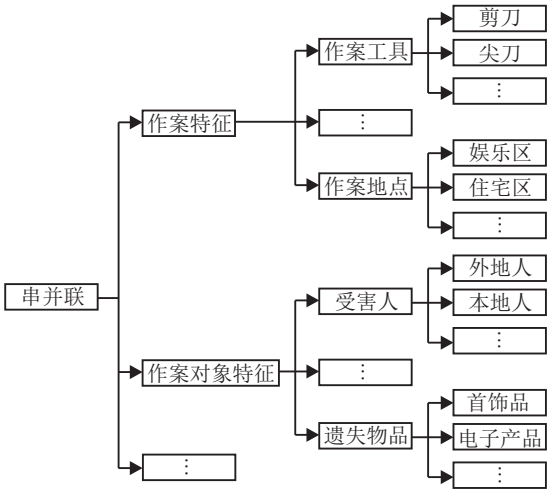

图 2　串并案指标体系

由于案件属性指标具有差异性、评价专家的认知性,相同案件底层指标的属性值集合会同时含有多种不确定性数. 比如针对“作案工具”指标而言:若现有的信息比较充分,不确定性低,则专家会选择统计概率的方法计算其可能性,得出底层指标(剪刀、尖刀、其他)的概率为(0.3, 0.7, 0);若现存的信息相对较少,则专家会选择灰数模型对底层指标(剪刀、尖刀、其他)进行分析,计算得出其灰数值为([0.3, 0.6], [0.4, 0.7], 0). 而针对“受害人”等特征指标,如果信息具有较高的模糊性,则专家会选择模糊理论对此类指标进行评价. 本文对6 起案件的底层指标进行打分,评价结果如表1、表2所示.

在表1和表2中:$*$表示白化数(概率数),$**$表示模糊数,$***$表示区间值模糊数(连续区间),$****$表示灰数.

表 1 作案特征属性指标评价结果

案例编号	作案工具 (剪刀、尖刀、其他)	作案地点 (娱乐区、住宅区、其他)
1	$(0.3,0.7,0)^{*}$	$([0.3,0.5],[0.6,0.7],0)^{***}$
2	$([0.3,0.6],[0.4,0.7],0)^{****}$	$(0,1,0)^{*}$
3	$(1,0,0)^{*}$	$(0.2,0.8,0)^{*}$
4	$(0.6,0.4,0)^{*}$	$(0.8,0.2,0)^{*}$
5	$(0,1,0)^{*}$	$(0.1,0.8,0.1)^{*}$
6	$(0,1,0)^{*}$	$(0.9,0.1,0.2)^{**}$

表 2 作案对象特征属性指标评价结果

案例编号	受害人 (外地人、本地人、其他)	损失物品 (首饰品、电子产品、其他)
1	$(0.7,0.3,0)^{*}$	$(0.9,0.1,0)^{*}$
2	$(0.2,0.7,0)^{**}$	$(0.3,0.7,0)^{*}$
3	$([0.1,0.2],[0.6,0.7],0)^{***}$	$(0.1,0.8,0.2)^{**}$
4	$(0.2,0.9,0)^{**}$	$(0.4,0.6,0)^{*}$
5	$(0.9,0.3,0)^{**}$	$([0.2,0.3],[0.7,0.8],0)^{****}$
6	$(0.7,0.3,0)^{*}$	$(0.3,0.7,0)^{*}$

4.2 串并案广义区间灰数熵权配置模型构建与求解

Step 1 底层指标属性值标准化根据定义12,利用广义标准灰数对多种不确定性数进行标准化,如表3、表4所示.

表 3 作案特征属性指标广义标准灰数表征

案例编号	作案工具 (剪刀、尖刀、其他)	作案地点 (娱乐区、住宅区、其他)
1	$(0.3,0.7,0)$	$(0.3+0.2\gamma_{11},0.6+0.1\gamma_{12},0)$
2	$(0.3+0.3\gamma_{21},0.4+0.3\gamma_{22},0)$	$(0,1,0)$
3	$(1,0,0)$	$(0.2,0.8,0)$
4	$(0.6,0.4,0)$	$(0.8,0.2,0)$
5	$(0,1,0)$	$(0.1,0.8,0.1)$
6	$(0,1,0)$	$(0.9,0.1,0.2)$

表 4 作案对象特征属性指标广义标准灰数表征

案例编号	受害人 (外地人、本地人、其他)	损失物品 (首饰品、电子产品、其他)
1	$(0.7,0.3,0)$	$(0.9,0.1,0)$
2	$(0.2,0.7,0)$	$(0.3,0.7,0)$
3	$(0.1+0.1\gamma_{31},0.6+0.1\gamma_{32},0)$	$(0.1,0.8,0.2)$
4	$(0.2,0.9,0)$	$(0.4,0.6,0)$
5	$(0.9,0.3,0)$	$(0.2+0.1\gamma_{41},0.7+0.1\gamma_{42},0)$
6	$(0.7,0.3,0)$	$(0.3,0.7,0)$

由于专家评价案件时,个人评价标准与规则相同,而不同专家的评价标准与规则各异,本文使相同案件的广义单位灰数相同,并区分不同案件的广义单位灰数,即$\gamma_1=\gamma_{11}=\gamma_{12},\gamma_2=\gamma_{21}=\gamma_{22},\gamma_3=\gamma_{31}=\gamma_{32},\gamma_4=\gamma_{41}=\gamma_{42}$.

Step 2 两两案件的相似度分析. 根据定义13和定义14计算6起案件的相似度

$$\otimes_{\rm sim}^{(1,2)}=\frac{(0.37+0.3\gamma_2)\cdot\overline{W_1^{\otimes}}\cdot\overline{W_5^{\otimes}}}{0.762\sqrt{(0.3+0.3\gamma_2)^2+(0.4+0.3\gamma_2)^2}}+\frac{(0.6+0.1\gamma_1)\cdot\overline{W_2^{\otimes}}\cdot\overline{W_5^{\otimes}}}{\sqrt{(0.3+0.2\gamma_1)^2+(0.6+0.1\gamma_1)^2}}+0.631\overline{W_3^{\otimes}W_6^{\otimes}}\cdot0.493\overline{W_4^{\otimes}}\cdot\overline{W_6^{\otimes}}.$$

同理计算出其他案件的相似度. 其中:$\overline{W_1^{\otimes}}$为作案工具的熵权,$\overline{W_2^{\otimes}}$为作案地点的熵权,$\overline{W_3^{\otimes}}$为受害人的熵权,$\overline{W_4^{\otimes}}$为损失物品的熵权,$\overline{W_5^{\otimes}}$为作案特征的熵权,$\overline{W_6^{\otimes}}$为作案对象特征的熵权.

Step 3 广义区间灰数熵权配置模型构建.

6起案件S_1,S_2,S_3,S_4,S_5,S_6的聚类情况为

$$\otimes_{\rm sim}^{(1,2)}>\otimes_{\rm sim}^{(1,3)},\otimes_{\rm sim}^{(1,4)},\otimes_{\rm sim}^{(1,5)},\otimes_{\rm sim}^{(1,6)},\otimes_{\rm sim}^{(2,3)},\otimes_{\rm sim}^{(2,4)},\otimes_{\rm sim}^{(2,5)},\otimes_{\rm sim}^{(2,6)};$$

$$\otimes_{\rm sim}^{(3,4)}>\otimes_{\rm sim}^{(1,3)},\otimes_{\rm sim}^{(2,3)},\otimes_{\rm sim}^{(3,5)},\otimes_{\rm sim}^{(3,6)},\otimes_{\rm sim}^{(1,4)},\otimes_{\rm sim}^{(2,4)},\otimes_{\rm sim}^{(4,5)},\otimes_{\rm sim}^{(4,6)};$$

$$\otimes_{\rm sim}^{(5,6)}>\otimes_{\rm sim}^{(1,5)},\otimes_{\rm sim}^{(2,5)},\otimes_{\rm sim}^{(3,5)},\otimes_{\rm sim}^{(4,5)},\otimes_{\rm sim}^{(1,6)},\otimes_{\rm sim}^{(2,6)},\otimes_{\rm sim}^{(3,6)},\otimes_{\rm sim}^{(4,6)}.$$

从而建立相对应的熵权配置模型如下:

$$\max P=\Big\{-\sum_{i=1}^{6}\overline{W_j^{\otimes}}\lg\overline{W_j^{\otimes}}\Big\};$$

$$\begin{cases}\overline{W_1^{\otimes}}+\overline{W_2^{\otimes}}=1,\\ \overline{W_3^{\otimes}}+\overline{W_4^{\otimes}}=1,\\ \overline{W_5^{\otimes}}+\overline{W_6^{\otimes}}=1,\\ \quad\vdots\\ (\overline{W_1^{\otimes}}+0.252\overline{W_2^{\otimes}})\overline{W_5^{\otimes}}+\Big(0.997\overline{W_3^{\otimes}}+\dfrac{0.55+0.1\gamma_4}{0.762\sqrt{(0.2+0.1\gamma_4)^2+(0.7+0.1\gamma_4)^2}}\cdot\overline{W_4^{\otimes}}\Big)\overline{W_6^{\otimes}}>\\ 0.339\overline{W_2^{\otimes}W_5^{\otimes}}+\Big(\dfrac{0.25+0.1\gamma_3}{0.583\sqrt{(0.1+0.1\gamma_3)^2+(0.6+0.1\gamma_3)^2}}\overline{W_3^{\otimes}}+0.993\overline{W_4^{\otimes}}\Big)\overline{W_6^{\otimes}},\\ (\overline{W_1^{\otimes}}+0.252\overline{W_2^{\otimes}})\overline{W_5^{\otimes}}+\Big(0.997\overline{W_3^{\otimes}}+\dfrac{0.55+0.1\gamma_4}{0.762\sqrt{(0.2+0.1\gamma_4)^2+(0.7+0.1\gamma_4)^2}}\cdot\overline{W_4^{\otimes}}\Big)\overline{W_6^{\otimes}}>\\ (0.555\overline{W_1^{\otimes}}+0.968\overline{W_2^{\otimes}})\overline{W_5^{\otimes}}+\\ (0.583\overline{W_3^{\otimes}}+0.983\overline{W_4^{\otimes}})\overline{W_6^{\otimes}}.\end{cases}$$

Step 4　广义区间灰数熵权配置模型求解. 利用仿真软件Matlab编程求解得到

$$\begin{cases}\overline{W_1^{\otimes}}=[0.703\,1,\ 0.757\,4],\\ \overline{W_2^{\otimes}}=[0.242\,6,\ 0.296\,9],\\ \overline{W_3^{\otimes}}=[0.245\,9,\ 0.395\,7],\\ \overline{W_4^{\otimes}}=[0.604\,3,\ 0.754\,1],\\ \overline{W_5^{\otimes}}=[0.725\,7,\ 0.801\,1],\\ \overline{W_6^{\otimes}}=[0.198\,9,\ 0.274\,3].\end{cases}$$

由于$W^{\otimes}=1/2(\overline{W^{\otimes L}}+\overline{W^{\otimes U}})$,则各个指标权重

$$W_j^{\otimes}=(0.730\,3,\ 0.269\,7,\ 0.320\,8,\ 0.679\,2,\ 0.763\,4,\ 0.236\,6).$$

4.3　基于广义区间灰数熵权的新案件聚类分析

通过上述方法已求得历史案件各层次指标的权重值,本文利用得到的指标权重对新案件作聚类分析. 首先对3起新案件的属性评价指标进行评分,评分结果如表5、表6所示.

表 5　新案件作案特征属性指标评分结果

案例编号	作案工具 (剪刀、尖刀、其他)	作案地点 (娱乐区、住宅区、其他)
1	$(0.2,0.8,0)^{*}$	$([0.2,0.3],[0.8,0.9],0)^{***}$
2	$([0.1,0.2],[0.8,0.9],0)^{****}$	$(0.4,0.6,0)^{*}$
3	$(1,0,0)^{*}$	$(0.7,0.3,0)^{*}$

表 6　新案件作案对象特征属性指标评分结果

案例编号	受害人 (外地人、本地人、其他)	损失物品 (首饰品、电子产品、其他)
1	$(0.1,0.9,0)^{*}$	$(0.5,0.5,0)^{*}$
2	$(0.8,0.1,0)^{**}$	$(0.8,0.2,0)^{*}$
3	$([0.1,0.2],[0.6,0.7],0)^{***}$	$(0.7,0.3,0)^{*}$

按照“入室盗窃类”、“抢劫类”和“扒窃类”3个灰类,采用白化权函数对3起新案件进行聚类,可得

$$f_1^1[-,-,0.1,0.3],\ f_2^1[0.3,0.5,-,0.75],$$
$$f_3^1[-,-,0.1,0.2],\ f_4^1[0.75,0.8,-,-],$$
$$f_5^1[0.6,0.8,-,-],\ f_6^1[-,-,0.1,0.25],$$
$$f_7^1[-,-,0.25,0.4],\ f_8^1[0.65,0.8,-,-],$$
$$f_1^2[0.3,0.4,-,0.65],\ f_2^2[0.75,0.9,-,-],$$
$$f_3^2[0.2,0.4,-,0.7],\ f_4^2[0.4,0.6,-,0.8],$$
$$f_5^2[0.3,0.4,-,0.6],\ f_6^2[0.55,0.6,-,-],$$
$$f_7^2[0.4,0.55,-,0.6],\ f_8^2[0.45,0.5,-,0.65],$$
$$f_1^3[0.65,0.8,-,-],\ f_2^3[-,-,0.05,0.1],$$
$$f_3^3[0.7,0.8,-,-],\ f_4^3[-,-,0.25,0.4],$$
$$f_5^3[-,-,0.1,0.15],\ f_6^3[0.2,0.3,-,0.6],$$
$$f_7^3[0.6,0.75,-,-],\ f_8^3[-,-,0.3,0.5].$$

由定义11和命题1可得相应的属性指标白化权函数值$F_j^k(\otimes_{ij})$,再通过广义灰色变权聚类系数$\sigma_i^k=\sum_{h=1}^{k_n}w_h^{\otimes}\cdots\sum_{j=1}^{k_i}w_{h\cdots j}^{\otimes}\cdots\sum_{l=1}^{k_l}F_l^k(\otimes_{il})\cdot w_{h\cdots j\cdots l}^{\otimes}$计算得到

$$(\sigma_i^k)=\begin{bmatrix}0.687\,8 & 0.236\,2 & 0.076\,8\\ 0.569\,7 & 0.780\,3 & 0.736\,8\\ 0 & 0.118\,2 & 0.428\,0\end{bmatrix}.$$

$\max\limits_{1\leqslant k\leqslant 3}\sigma_1^k=\sigma_1^1=0.687\,8$, $\max\limits_{1\leqslant k\leqslant 3}\sigma_2^k=\sigma_2^2=0.780\,3$, $\max\limits_{1\leqslant k\leqslant 3}\sigma_3^k=\sigma_3^3=0.428\,0$,表明案件1属于“入室盗窃类”,案件2属于“抢劫类”,案件3属于“扒窃类”.

实际上,“入室盗窃类”案件通常是本地人作案,因为嫌疑人习惯熟悉地形,易于观察并选择合适的作案时间,并且选择地点是住宅区,作案工具为尖刀,盗窃的物品为中小型、昂贵的物件,如首饰品和电子产品等,与案件1的情况相似. “抢劫类”案件通常是外地人作案,为了不易被人发现,其作案手段简单,常用尖刀之类的工具威胁受害人,作案地点偏向于过往人少的住宅区,且盗窃物品多为昂贵的首饰品,与案件2的情况相似. “扒窃类”案件多为本地人作案,因为能更好地熟悉当地的扒窃环境和规律,扒窃地点通常选择热闹的娱乐区,借助拥挤的环境,用剪刀作案,扒窃首饰品等轻小型物件,与案件3的情况相似. 因此本文提出的聚类模型较为科学合理,符合实际情况.

5　结　论

本文针对多元异构不确定性对象的聚类问题,提出广义区间灰数的概念以及严谨的证明过程,证明了多种不确定数均可用其表征,并设计了相应的广义区间灰数的白化权函数,然后构造了基于多元异构不确定性案例学习的广义区间灰数熵权配置模型,得到各指标的聚类权重,并结合广义区间灰数白化权函数,对新的案例进行聚类分析,最后以刑事犯罪串并案为例对聚类模型进行验证分析,得到的结果符合现实情况,表明了本文提出的聚类模型具有一定的可行性和有效性,为解决多元异构不确定性对象的聚类研究提供了新方法.

参考文献(References)

[1] Liu B, Shen Y, Chen Y, et al. A two-layer weight determination method for complex multi-attribute large-group decision-making experts in a linguistic environment[J]. Information Fusion, 2015, 23(C): 156-165.

[2] Wu Y, Xu H, Xu C, et al. Uncertain multi-attributes

decision making method based on interval number with probability distribution weighted operators and stochastic dominance degree[J]. Knowledge-Based Systems, 2016, 113(1): 199-209.

[3] 陆亿红, 夏聪. 不确定数据的最优k近邻和局部密度聚类算法[J]. 控制与决策, 2016, 31(3): 541-546.
(Lu Y H, Xia C. Optimal k-nearest neighbors and local density-based clustering algorithm for uncertain data[J]. Control and Decision, 2016, 31(3): 541-546.)

[4] 陈健美, 陆虎, 宋余庆, 等. 一种隶属关系不确定的可能性模糊聚类方法[J]. 计算机研究与发展, 2008, 45(9): 1486-1492.
(Chen J M, Lu H, Song Y Q, et al. A possibility fuzzy clustering algorithm based on the uncertainty membership[J]. J of Computer Research and Development, 2008, 45(9): 1486-1492.)

[5] 郑爱武. 基于模糊k-均值聚类模型的移动终端业务故障诊断[J]. 统计与决策, 2014, 3(11): 83-85.
(Zhen A W. Fault diagnosis of mobile terminal based on fuzzy k-means clustering model[J]. Statistics and Decision, 2014, 3(11): 83-85.)

[6] Jian H, Shu-bin S, Yi-min M, et al. High dimensional uncertain data efficient clustering algorithm[J]. Computer Knowledge & Technology, 2014, 10(4): 673-676.

[7] 李慧, 张庆范, 段培永, 等. 一种基于聚类的超闭球模糊神经网络[J]. 控制与决策, 2011, 26(12): 1803-1807.
(Li H, Zhang Q F, Duan P Y, et al. Hyperball fuzzy neural network based on clustering[J]. Control and Decision, 2011, 26(12): 1803-1807.)

[8] 何云斌, 张志超, 万静, 等. 不确定数据聚类的U-PAM算法和UM-PAM算法的研究[J]. 计算机科学, 2016, 43(6): 263-269.
(He Y B, Zhang Z C, Wan J, et al. Research for uncertain data clustering algorithm: U-PAM and UM-PAM algorithm[J]. Computer Science, 2016, 43(6): 263-269.)

[9] Deng J L. Control problem of grey system[J]. Systems and Control Letters, 1982, 1(5): 288-294.

[10] Liu H Q, Fang Z G, Li W D, et al. Object-oriented multi-attribute differences matrix grey clustering method and its application[J]. Control and Decision, 2015, 30(2): 366-370.

[11] 王嵩华, 朱建军, 方志耕. 基于灰色聚类的大规模群体语言评价信息集结研究[J]. 控制与决策, 2012, 27(2): 271-275.
(Wang H H, Zhu J J, Fang Z G. Group aggregation method on large-scale linguistic evaluation information based on grey cluster[J]. Control and Decision, 2012, 27(2): 271-275.)

[12] Yuan C Q, Liu S F. Core of grey cluster and its application in evaluation of scientific and technological strength[J]. J of Grey System, 2012, 24(4): 327-336.

[13] Pei L L, Wang Z X. An optimized grey cluster model for evaluating quality of labor force[J]. J of Software, 2013, 8(10): 2489-2494.

[14] Li S Z, Zhang Z D, He R S. Application of grey clustering evaluations in coal railway transportation[J]. Kybernetes, 2012, 41(5/6): 714-725.

[15] 钱丽丽, 刘思峰, 谢乃明. 基于熵权和区间灰数信息的灰色聚类模型[J]. 系统工程与电子技术, 2016, 38(2): 352-356.
(Qian L L, Liu S F, Xie N M. Grey clustering model based on entropy-weight and grey numbers[J]. Systems Engineering and Electronics, 2016, 38(2): 352-356.)

[16] 王斐, 梁晓庚, 郭超, 等. 灰云熵权聚类的制导仿真系统可信度评估[J]. 系统仿真学报, 2015, 27(8): 1703-1707.
(Wang F, Liang X G, Guo C, et al. Guidances simulation credibility evaluation based on clustering with grey cloud and entropy weight[J]. J of System Simulation, 2015, 27(8): 1703-1707.)

[17] 高志扬, 雒赵飞, 景国勋, 等. 综采工作面环境状况灰色熵权聚类分析[J]. 煤炭技术, 2016, 35(6): 145-147.
(Gao Z Y, Luo Z F, Jing G X, et al. Grey entropy weight clustering analysis of environment state of fully mechanized mining faces[J]. Coal Technology, 2016, 35(6): 145-147.)

[18] Yang Y, John R. Grey sets and greyness[J]. Information Sciences, 2012, 185(1): 249-264.

[19] 刘思峰, 党耀国, 方志耕. 灰色系统理论及其应用[M]. 北京: 科学出版社, 2004: 14-25.
(Liu S F, Dang Y G, Fang Z G. The theory of grey system and its application[M]. Beijing: Science Press, 2004: 14-25.)

[20] Halberstam H, Elliott P D. Probabilistic number theory I and II[J]. Mathematical Gazette, 1982, 66(435): 92.

[21] Zadeh L A. Fuzzy sets[J]. Information and Control, 1965, 8(3): 338-353.

[22] Zhang H, Shu L. Generalized interval-valued fuzzy rough set and its application in decision making[J]. Int J of Fuzzy Systems, 2015, 17(2): 279-291.

[23] 周伟杰, 党耀国, 熊萍萍, 等. 区间灰数的灰色变权与定权聚类模型[J]. 系统工程理论与实践, 2013, 33(10): 2590-2595.
(Zhou W J, Dang Y G, Xiong P P, et al. Grey clustering model for interval grey number with variable and fixed weights[J]. Systems Engineering —— Theory & Practice, 2013, 33(10): 2590-2595.)

[24] 邱菀华. 管理决策与应用熵学[M]. 北京: 机械工业出版社, 2002: 79-96.
(Qiu W H. Management decision making and application of entropy[M]. Beijing: Mechanical Industry Press, 2002: 79-96.)

[25] 刘思峰, 方志耕, 谢乃明. 基于核和灰度的区间灰数运算法则[J]. 系统工程与电子技术, 2010, 32(2): 313-316.
(Liu S F, Fang Z G, Xie N M. Algorithm rules of interval grey numbers based on the kernel and the degree of greyness of grey numbers[J]. Systems Engineering and Electronic, 2010, 32(2): 313-316.)

(责任编辑: 闫 妍)