区间数模糊 c 均值聚类中相对位置相异度的研究

袁 飞[1] 詹宜巨[2] 王永华[3]
(1. 中山大学信息科学与技术学院，广州 510006；2. 中山大学工学院，广州 510006；
3. 广东工业大学自动化学院，广州 510006)

摘 要：区间数模糊 c 均值聚类方法中，区间数距离公式存在无法描述区间数之间相对位置的问题，针对该问题，本文分析了该问题产生原因，提出了相对位置相异度公式，并将该相异度公式应用于区间数模糊 c 均值聚类中。理论分析说明相对位置相异度公式能定量描述区间数之间相异程度，还能描述区间数之间相对位置。仿真实验结果表明，相对于基于现有区间数距离公式的区间数模糊 c 均值聚类，基于相对位置相异度的区间数模糊 c 均值聚类方法具有更好的聚类效果。同时，给出了相对位置相异度公式中参数选择标准。

关键词：不确定数据；区间数；区间数距离；模糊 c 均值聚类
中图分类号：TP18 **文献标识码**：A **文章编号**：1003-0530(2012)10-1370-09

Research on Relative Position Dissimilarity in Interval-data Fuzzy C-Means Clustering

YUAN Fei[1] ZHAN Yi-ju[2] WANG Yong-hua[3]
(1. School of Information Science and Technology, SUN-YAT SEN University, Guangzhou 510006, China;
2. School of Engineering, SUN-YAT SEN University, Guangzhou 510006, China;
3. School of Automation, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: This paper discussed the problem that various distances in the interval-data fuzzy c-means clustering method (labeled IFCM) can't represent the relative position of interval data, and proposed the relative position dissimilarity. The relative position dissimilarity is constructed based on the fact that the differential value between distance of midpoint of interval data and sum of the half length of interval data could reflect the relative position of interval data. And the relative position dissimilarity satisfies the conditions: 1) it decreases as the decrease of the differential value; 2) it decreases as the increase of the sum of interval data length. In theory, the relative position dissimilarity depicts the difference of the interval data in quantity and the relative position of interval data. Meanwhile, the relative position dissimilarity was applied in the IFCM clustering method, which called as IFCM-RPD clustering method. Experimental results show that the IFCM-RPD clustering method has better clustering effect. As well, selection criteria of the parameters in the relative position dissimilarity are given.

Key words: uncertain data; interval data; interval data distance; fuzzy c-means clustering

1 引言

不确定数据是指因观测手段、观测设备和外界环境等因素造成观测数据与事物真实属性存在有限差别，且差别是模糊的、或随机的、或不精确的一类观测数据。普遍存在于金融[1][2]、气象[3][4][5]和通信[6]-[9]

收稿日期：2012-03-27；修回日期：2012-07-19
基金项目：国家自然科学基金项目(61071038)；国家自然科学基金项目(61102034)；广东省教育部产学研结合重点项目(2011A090200128)

等领域。为了处理不确定性数据,目前已发展了多种数学方法,如模糊数学、灰色系统理论、可拓学、属性数学、粗糙集理论、概率理论和区间数理论等。以上数学方法拓展了经典数学中定量化描述事物的方法,能更好地描述不确定事物,但客观事物的复杂性、环境的不确定性和人类思维的局限性,无法明确给出事物属性值,即使大量实验也不能准确得到属性值,只能给出一个区间范围,即以区间数形式表示。

区间数理论广泛应用于多属性决策、数据挖掘和符号数据分析(SDA)等领域,目前许多学者从不同角度对 SDA 领域中的区间数聚类分析进行了研究[10]-[14]。进行区间数据聚类分析中,需计算区间数之间距离,距离公式既要能定量描述区间数之间距离,又要符合区间数之间位置关系。文献[15][16]对区间数距离公式进行了改进,但仍然存在距离值无法描述两区间数间位置关系,存在距离大小关系与区间数相对位置不一致的情况。针对该问题,本文提出了相对位置相异度公式,该公式能合理描述区间数之间相异程度和相对位置,并通过实验仿真,说明该相异度公式对区间数模糊 c 均值(Interval-data Fuzzy C-Means — IFCM)聚类效果有明显改善作用。

2　区间数距离传统定义的缺陷

假设某对象的属性值为某区间范围内的某个值,则据区间数理论,可用该区间来描述该属性。区间数的定义如定义 1 所述。

定义 1. 区间数[17]

设 $\mathbf{R}$ 表示实数集,对任意的 $a^-, a^+ \in \mathbf{R}$,且 $a^- < a^+$,记

$$[\bar{A}]=[a^-,a^+] \tag{1}$$

称$[\bar{A}]=[a^-,a^+]$为一个标准的二元区间数,其中 a^+为上极限,称为二元区间的大元;a^-为下极限,称为二元区间的小元。

定义 2. 区间数向量

设有 p 个区间数分别为:$[\bar{A}_1]=[a_1^-,a_1^+]$,$[\bar{A}_2]=[a_2^-,a_2^+]$,$\cdots$,$[\bar{A}_p]=[a_p^-,a_p^+]$,则由这 p 个区间数构成的向量 $\boldsymbol{A}=([a_1^-,a_1^+],[a_2^-,a_2^+],\cdots,[a_p^-,a_p^+])$称为 p 维区间数向量。

根据区间数间相对位置关系,可将相对位置关系划分为相交、相离和相接三种,如图 1 所示。

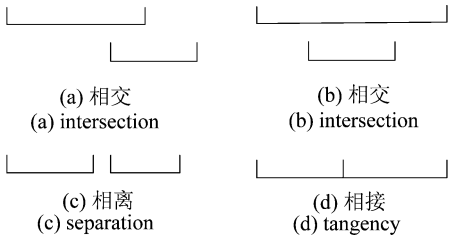


图 1　区间数相对位置关系

Fig. 1　Relative Position of the interval data

目前区间数距离公式多种多样,主要由点数据对象距离公式推广得到。若点数据对象 $\boldsymbol{a}$、$\boldsymbol{b}$ 分别为 $\boldsymbol{a}=[a_1,a_2,\cdots,a_p]$,$\boldsymbol{b}=[b_1,b_2,\cdots,b_p]$,区间数据对象分别为 $\boldsymbol{A}=([a_1^-,a_1^+],[a_2^-,a_2^+],\cdots,[a_p^-,a_p^+])$,$\boldsymbol{B}=([b_1^-,b_1^+],[b_2^-,b_2^+],\cdots,[b_p^-,b_p^+])$,则点数据对象 $\boldsymbol{a}$、$\boldsymbol{b}$ 间距离 $D(\boldsymbol{a},\boldsymbol{b})$与区间数据对象 $\boldsymbol{A}$、$\boldsymbol{B}$ 间距离 $D(A,B)$对比如表 1 所示。

区间数向量的 Cityblock 距离、Euclidean 距离和 Hausdorff 距离是 Minkowski 距离的特殊情况,下面将分析区间数间 Minkowski 距离中存在的区间数距离与区间数相对位置不一致的问题。

假设三个区间数分别为$[\bar{A}]=[a^-,a^+]$、$[\bar{B}]=[b^-,b^+]$、$[\bar{C}]=[c^-,c^+]$,且区间数$[\bar{A}]$与$[\bar{B}]$相交,$[\bar{A}]$与$[\bar{C}]$相离,据 Minkowski 距离公式可得:

$$D([\bar{A}],[\bar{B}])=(|a^--b^-|^n+|a^+-b^+|^n)^{1/n}$$

$$D([\bar{A}],[\bar{C}])=(|a^--c^-|^n+|a^+-c^+|^n)^{1/n}$$

其中 $n\geqslant 1$。

要证明 Minkowski 距离公式存在区间数距离与区间数相对位置矛盾的问题,则只需证明存在 $D([\bar{A}],[\bar{B}])\geqslant D([\bar{A}],[\bar{C}])$的情况即可,亦只需证明下式存在即可。

$$|a^--b^-|^n+|a^+-b^+|^n\geqslant|a^--c^-|^n+|a^+-c^+|^n$$

若区间数$[\bar{A}]$的区间长度足够小,且 a^-,a^+均趋近于 0,则上式可近似为:

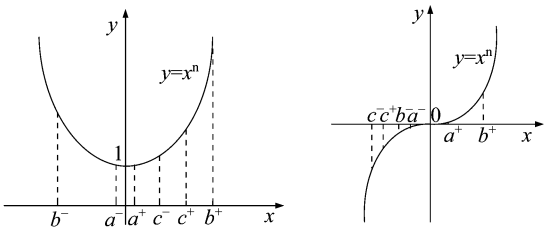
$$|b^-|^n+|b^+|^n\geqslant|c^-|^n+|c^+|^n \tag{2}$$

故若上式成立,则区间数间 Minkowski 距离值与区间数间相对位置矛盾。式(2)的证明可由图 2 所示进行说明。

表 1 点数据对象与区间数据对象距离对比

Tab. 1 Comparison of distances between point data objects and interval data objects

距离类型	$D(\boldsymbol{a},\boldsymbol{b})$	$D(\boldsymbol{A},\boldsymbol{B})$
Minkowiski 距离	$D(\boldsymbol{a},\boldsymbol{b}) = (\sum_{i=1}^{p} \lvert a_i - b_i \rvert^n)^{1/n}$ 其中 n 为自然数	$D(\boldsymbol{A},\boldsymbol{B}) = (\sum_{i=1}^{p} (\lvert a_i^- - b_i^- \rvert^n + \lvert a_i^+ - b_i^+ \rvert^n))^{1/n}$
Cityblock 距离	$D(\boldsymbol{a},\boldsymbol{b}) = \sum_{i=1}^{p} \lvert a_i - b_i \rvert$	$D(\boldsymbol{A},\boldsymbol{B}) = \sum_{i=1}^{p} (\lvert a_i^- - b_i^- \rvert + \lvert a_i^+ - b_i^+ \rvert)$
Euclidean 距离	$D(\boldsymbol{a},\boldsymbol{b}) = \sum_{i=1}^{p} (a_i - b_i)^2$	$D(\boldsymbol{A},\boldsymbol{B}) = (\sum_{i=1}^{p} ((a_i^- - b_i^-)^2 + (a_i^+ - b_i^+)^2))^{1/2}$
Hausdorff 距离	$D(\boldsymbol{a},\boldsymbol{b}) = \max_{1\leqslant i\leqslant p} (\lvert a_i - b_i \rvert)$	$D(\boldsymbol{A},\boldsymbol{B}) = \sum_{i=1}^{p} \max(\lvert a_i^- - b_i^- \rvert, \lvert a_i^+ - b_i^+ \rvert)$
Adaptive quadratic 距离	$D^2(\boldsymbol{a},\boldsymbol{b}) = (\boldsymbol{a}-\boldsymbol{b})^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{a}-\boldsymbol{b})$ $\boldsymbol{M}$ 为 $\boldsymbol{a}$、$\boldsymbol{b}$ 向量间的协方差矩阵	$D^2(\boldsymbol{A},\boldsymbol{B}) = (\boldsymbol{a}^- - \boldsymbol{b}^-)^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{a}^- - \boldsymbol{b}^-) + (\boldsymbol{a}^+ - \boldsymbol{b}^+)^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{a}^+ - \boldsymbol{b}^+)$ $\boldsymbol{a}^- = (a_1^-, a_2^-, \cdots, a_p^-)$, $\boldsymbol{a}^+ = (a_1^+, a_2^+, \cdots, a_p^+)$, $\boldsymbol{b}^- = (b_1^-, b_2^-, \cdots, b_p^-)$, $\boldsymbol{b}^+ = (b_1^+, b_2^+, \cdots, b_p^+)$, $\boldsymbol{M}$ 为 $\boldsymbol{A}$、$\boldsymbol{B}$ 向量间协方差矩阵



(a) n为偶数 (b) n为奇数

(a) n is an even number (b) n is an odd number

图 2 区间数 Minkowski 距离与区间相对位置不一致图示

Fig. 2 Inconsistency of Minkowski distance of interval data and their relative position

图 2(a)中,区间数$[\bar{A}]$与$[\bar{B}]$相交,$[\bar{A}]$与$[\bar{C}]$相离,且 a^-, a^+均趋近于 0,但 Minkowski 距离 $D([\bar{A}],[\bar{B}]) \geqslant D([\bar{A}],[\bar{C}])$,距离值与相对位置关系矛盾;图 2(b)中,同样存在 Minkowski 距离值与相对位置关系矛盾的情况。

若区间数向量 $\boldsymbol{A} = ([a_1^-, a_1^+], [a_2^-, a_2^+], \cdots, [a_p^-, a_p^+])$, $\boldsymbol{B} = ([b_1^-, b_1^+], [b_2^-, b_2^+], \cdots, [b_p^-, b_p^+])$,则据区间数二次型距离公式可知,$\boldsymbol{A}$ 与 $\boldsymbol{B}$ 间二次型距离为:

$$D^2(\boldsymbol{A},\boldsymbol{B}) = (\boldsymbol{a}^- - \boldsymbol{b}^-)^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{a}^- - \boldsymbol{b}^-) + (\boldsymbol{a}^+ - \boldsymbol{b}^+)^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{a}^+ - \boldsymbol{b}^+)$$

其中向量 $\boldsymbol{a}^- = (a_1^-, a_2^-, \cdots, a_p^-)$、$\boldsymbol{a}^+ = (a_1^+, a_2^+, \cdots, a_p^+)$、$\boldsymbol{b}^- = (b_1^-, b_2^-, \cdots, b_p^-)$、$\boldsymbol{b}^+ = (b_1^+, b_2^+, \cdots, b_p^+)$,矩阵 $\boldsymbol{M}$ 为区间数向量 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 间的属性协方差矩阵,为正定对称矩阵,则矩阵 $\boldsymbol{M}$ 可分解为:$\boldsymbol{M} = \boldsymbol{R}^{\mathrm{T}}\boldsymbol{R}$,$\boldsymbol{R}$ 为下三角矩阵。令向量 $\boldsymbol{X}^- = \boldsymbol{a}^- - \boldsymbol{b}^-$, $\boldsymbol{X}^+ = \boldsymbol{a}^+ - \boldsymbol{b}^+$,则 $D^2(\boldsymbol{A},\boldsymbol{B}) = (R\boldsymbol{X}^-)^{\mathrm{T}}(R\boldsymbol{X}^-) + (R\boldsymbol{X}^+)^{\mathrm{T}}(R\boldsymbol{X}^+)$,该公式形式与区间数欧氏距离的矩阵表示方式一样,则据区间数 Minkowiski 距离与区间数相对位置不一致的分析可知:$D^2(\boldsymbol{A},\boldsymbol{B}) = (R\boldsymbol{X}^-)^{\mathrm{T}}(R\boldsymbol{X}^-) + (R\boldsymbol{X}^+)^{\mathrm{T}}(R\boldsymbol{X}^+)$同样具有距离与区间数相对位置不一致的问题。

表 1 中区间数距离公式是点数据距离公式的推广,按照将区间数的上极限和下极限分别作为一维数据的原理,将原有 p 维区间数向量变为 $2p$ 维点数据向量,再据点数据向量的距离公式求得区间数距离。该扩展方法割裂了区间数向量中各区间数的结构,使得区间数距离与区间数相对位置存在了不一致问题。为了解决该问题,下一节将提出相对位置相异度公式,该公式能够反映区间数间相对位置,并能定量描述区间数间相异程度。同时,将该相异度公式应用到 IFCM 算法中,提出基于相对位置相异度的 IFCM 聚类方法(IFCM clustering method based on relative position dissimilarity. IFCM-RPD),以期得到更好的聚类效果。

3 相对位置相异度及其在 IFCM 聚类方法中的应用

3.1 区间数相对位置相异度

定义 3. 设区间数$[\bar{A}] = [a^-, a^+]$, $[\bar{B}] = [b^-, b^+]$,区间数$[\bar{A}]$的长度为 l_1,中心点坐标为 u_1,区间数$[\bar{B}]$的长度为 l_2,中心点坐标为 u_2,则区间数$[\bar{A}]$、$[\bar{B}]$相异度 $D([\bar{A}],[\bar{B}])$定义为:

$$D([\bar{A}],[\bar{B}]) = h^{d([\bar{A}],[\bar{B}])} \tag{3}$$

其中 $d([\bar{A}],[\bar{B}])=\dfrac{|u_2-u_1|-0.5\times(l_1+l_2)}{l_1+l_2+w}$,$h$、$w$ 为参数,且 $h>1$,$w\geqslant 0$,$d([\bar{A}],[\bar{B}])\in[-\dfrac{1}{2}\dfrac{l_1+l_2}{l_1+l_2+w},+\infty)$。

据文献[18]可知,数据集 $\boldsymbol{E}$ 中各元素间的相异度度量函数 D 必须为实数映射,即 $D:\boldsymbol{E}\times\boldsymbol{E}\rightarrow\mathbf{R}$,$\mathbf{R}$ 为实数集,且映射 D 需要满足以下条件:

(1) $D(a,b)=D(b,a)\qquad \forall a,b\in\boldsymbol{E}$

(2) $D(a,b)\geqslant D(a,a)\qquad \forall b\in\boldsymbol{E}$

(3) $D(a,b)\leqslant+\infty\qquad \forall a,b\in\boldsymbol{E}$

因为公式(3)中 $d([\bar{A}],[\bar{B}])=d([\bar{B}],[\bar{A}])$,故易得公式(3)满足条件(1);且公式(3)显而易见满足条件(3)。

若区间数 $[\bar{A}]=[a^-,a^+]$,$[\bar{B}]=[b^-,b^+]$,区间数 $[\bar{A}]$ 的长度为 l_1,中心点坐标为 u_1,区间数 $[\bar{B}]$ 的长度为 l_2,中心点坐标为 u_2,且令 $\delta=|u_2-u_1|-\dfrac{1}{2}(l_1+l_2)$,则 $\delta=[-\dfrac{1}{2}(l_1+l_2),+\infty)$,据公式(3)可得:

$$D([\bar{A}],[\bar{A}])=h^{d([\bar{A}],[\bar{A}])}$$

$$D([\bar{A}],[\bar{B}])=h^{d([\bar{A}],[\bar{B}])}$$

其中 $d([\bar{A}],[\bar{A}])=\dfrac{-l_1}{2\times l_1+w}$,$d([\bar{A}],[\bar{B}])=\dfrac{\delta}{l_1+l_2+w}$。

要证明公式(3)满足条件(2),则只需证明 $\dfrac{\delta}{l_1+l_2+w}\geqslant\dfrac{-l_1}{2\times l_1+w}$ 恒成立即可。

若 $l_1\geqslant l_2$,则

$$\min\left(\frac{\delta}{l_1+l_2+w}\right)=\frac{-0.5\times(l_1+l_2)}{l_1+l_2+w}$$

易得到 $\min\left(\dfrac{\delta}{l_1+l_2+w}\right)\geqslant\dfrac{-l_1}{2\times l_1+w}$,故 $\dfrac{\delta}{l_1+l_2+w}\geqslant\dfrac{-l_1}{2\times l_1+w}$ 恒成立;

若 $l_1<l_2$,则

当 $\delta\in[-\dfrac{l_1(l_1+l_2+w)}{2\times l_1+w},+\infty)$ 时,易得 $\dfrac{\delta}{l_1+l_2+w}\geqslant\dfrac{-l_1}{2\times l_1+w}$ 恒成立,

当 $\delta\in[-\dfrac{l_1+l_2}{2},-\dfrac{l_1(l_1+l_2+w)}{2\times l_1+w})$ 时,$\dfrac{\delta}{l_1+l_2+w}<\dfrac{-l_1}{2\times l_1+w}$。由于 $w\geqslant 0$,当 $w=0$ 时,可知 δ 所属区间为空,此时 $\dfrac{\delta}{l_1+l_2+w}\geqslant\dfrac{-l_1}{2\times l_1+w}$ 恒成立,且当 w 趋近于 0 时,该区间长度亦趋近于 0。在实际应用中,在保证 δ 分母不为零的前提下,w 可取 0 或取趋近于 0 的值。

由以上分析可知:当 $w=0$ 时,公式(3)满足条件(2);当 $w\neq 0$,且 δ 在区间 $[-\dfrac{l_1(l_1+l_2+w)}{2\times l_1+w},+\infty)$ 内时,公式(3)满足满足条件(2),δ 在区间 $[-\dfrac{l_1+l_2}{2},-\dfrac{l_1(l_1+l_2+w)}{2\times l_1+w})$ 内时,公式(3)不满足满足条件(2),但若 w 趋近于 0,区间 $[-\dfrac{l_1+l_2}{2},-\dfrac{l_1(l_1+l_2+w)}{2\times l_1+w})$ 的范围亦趋近于 0,此时可以近似认为公式(3)满足条件(2)。故由以上分析可知:定义 3 定义的区间数相异度公式满足相异度度量函数条件,可以定量描述区间数间的相异程度。

由定义 3 定义的区间数相异度公式易知,区间数 $[\bar{A}]$ 和 $[\bar{B}]$ 间相异度 $D([\bar{A}],[\bar{B}])$ 具有以下性质:

(1) $D([\bar{A}],[\bar{B}])$ 随 $d([\bar{A}],[\bar{B}])$ 的增加而增加。

(2) 若区间数 $[\bar{A}]$ 和 $[\bar{B}]$ 相交,则 $0<D([\bar{A}],[\bar{B}])<1$。

(3) 若区间数 $[\bar{A}]$ 和 $[\bar{B}]$ 相接,则 $D([\bar{A}],[\bar{B}])=1$。

(4) 若区间数 $[\bar{A}]$ 和 $[\bar{B}]$ 相离,则 $D([\bar{A}],[\bar{B}])>1$。

由以上性质可知,相异度 $D([\bar{A}],[\bar{B}])$ 能准确描述区间数间相对位置。

定义 4. 设两区间数向量分别为 $\boldsymbol{x}=(x_1,x_2,\cdots,x_p)$,$\boldsymbol{y}=(y_1,y_2,\cdots,y_p)$,其中 $x_i=[a_i^-,a_i^+]$,$y_i=[b_i^-,b_i^+]$,$i=1,2,\cdots,p$,区间数 x_i 与 y_i 的相异度为 $D(x_i,y_i)$,由式(3)求得,$i=1,2,\cdots,p$。则区间数向量 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 的相异度为:

$$D(\boldsymbol{x},\boldsymbol{y})=\prod_{i=1}^{p}D(x_i,y_i)\qquad(4)$$

据区间数相异度性质易知,区间数向量 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 间相异度 $D(\boldsymbol{x},\boldsymbol{y})$ 具有以下性质:

(1) 若 $D(\boldsymbol{x},\boldsymbol{y})<\min(D_1,D_2,\cdots,D_p)$,则区间数向量 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 相交;

(2) 若存在 $D(x_i,y_i)>1,i=1,2,\cdots,p$,则区间数向量 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 相离;

(3) 若存在 $D(x_i,y_i)=1,i=1,2,\cdots,p$,则区间数向量 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 相接。

3.2 基于相对位置相异度的 IFCM 聚类方法

假设符号数据分析系统对区间型数据进行模糊 c 均值聚类分析时,数据集 $\boldsymbol{X}=\{\boldsymbol{x}_k \mid k=1,2,\cdots,n\}$,其中 $\boldsymbol{x}_k$ 为 p 维区间数向量,$\boldsymbol{x}_k=(x_{k1},x_{k2},\cdots,x_{kp})$,$x_{kj}$ 为区间形式,$x_{kj}=[a_{kj}^-,a_{kj}^+],k=1,2,\cdots,n,j=1,2,\cdots,p$。簇中心 $\boldsymbol{g}_i$ $(i=1,2,\cdots,c)$ 描述为 $\boldsymbol{g}_i=([\alpha_{i1}^-,\alpha_{i1}^+],[\alpha_{i2}^-,\alpha_{i2}^+],\cdots,[\alpha_{ip}^-,\alpha_{ip}^+])$。则区间数的模糊 c 均值聚类算法即求解式(5)的最优规划问题。

$$\begin{aligned} &min \quad \boldsymbol{W}(\boldsymbol{G},\boldsymbol{U},\boldsymbol{X})=\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m D(\boldsymbol{x}_k,\boldsymbol{g}_i) \\ &s.t. \quad \sum_{i=1}^{c}u_{ik}=1 \quad k=1,2,\cdots,n \end{aligned} \tag{5}$$

其中 u_{ik} 表示第 k 个区间数向量数据属于第 i 个簇的隶属度,m 为模糊度系数,$\boldsymbol{U}$ 为 u_{ik} 形成的隶属度矩阵,$D(\boldsymbol{x}_k,\boldsymbol{g}_i)$ 为第 k 个区间数向量与第 i 个簇中心区间数向量间相异度,即两个 p 维区间数向量的相异度。

聚类是将簇内数据相异度小的数据归为一类,且保证簇间数据相异度大。IFCM-RPD 聚类算法中,区间数的相对位置相异度是以参数 h 为底的指数函数形式,则簇间相离数据间相异度会随着参数 h 的增大而增大,簇内相交数据的相异度会随着参数 h 的增大而减小,故参数 h 的选择应在计算能力允许下尽可能大。相对位置相异度也是参数 w 的减函数,即相异度随 w 的减少而增大,故参数 w 应在保证公式(3)中 $d([\bar{A}],[\bar{B}])$ 分母不为零的前提下尽可能小,并可以取 $w=0$。

根据实际情况选择好参数 h 和 w 后,IFCM-RPD 聚类算法按照以下步骤进行。

步骤 1:从数据集 $\boldsymbol{X}$ 中随机选择 c 个数据作为初始簇中心 $\boldsymbol{G}^0$,设定迭代计数值 $s=0$ 及迭代终止门限值 ε,并设定模糊度系数 $m=2$,参数 h 和 w;

步骤 2:基于 $\boldsymbol{G}^s$,根据式(6)计算 u_{ik},得到隶属度矩阵 $\boldsymbol{U}^s$

$$u_{ik}=\begin{cases}\left[\sum_{h=1}^{c}\dfrac{D(\boldsymbol{x}_k,\boldsymbol{g}_i)}{D(\boldsymbol{x}_k,\boldsymbol{g}_h)}\right]^{-1} & \boldsymbol{x}_k\neq\boldsymbol{g}_i \\ u_{ik}=1 \quad and \quad u_{jk}=0 \quad for \quad j\neq i & \boldsymbol{x}_k=\boldsymbol{g}_i\end{cases} \tag{6}$$

其中 $D(\boldsymbol{x}_k,\boldsymbol{g}_i)=\prod_{j=1}^{p}D(x_{kj},g_{ij})$,$D(x_{kj},g_{ij})$ 为据式(3)求得的区间数 x_{kj} 与 g_{ij} 的距离,$i=1,2,\cdots,c,k=1,2,\cdots,n$。

步骤 3:迭代计数值 s 加 1,根据式(7)计算第 s 步的 c 个簇中心 $\boldsymbol{G}^s$。

$$\alpha_{ij}^-=\frac{\sum_{k=1}^{n}(u_{ik})^2 a_{kj}^-}{\sum_{k=1}^{n}(u_{ik})^2} \qquad \alpha_{ij}^+=\frac{\sum_{k=1}^{n}(u_{ik})^2 a_{kj}^+}{\sum_{k=1}^{n}(u_{ik})^2} \tag{7}$$

步骤 4:根据步骤 2 中的式(6)计算得到新的隶属度矩阵 $\boldsymbol{U}^s$,

步骤 5:若 $\|\boldsymbol{U}^s-\boldsymbol{U}^{(s-1)}\|<\varepsilon$,则终止聚类迭代,否则重复步骤 3、4、5。

下一节将对 IFCM-RPD 聚类方法进行聚类效果仿真,以说明 IFCM-RPD 聚类方法相对于现有主要 IFCM 聚类方法具有更好的聚类效果。

4 仿真实验

本节仿真实验分为两部分,第一部分采用合成数据,对比相对位置相异度公式中不同参数 h 下的 IFCM-RPD 聚类效果,验证参数 h 对 IFCM-RPD 聚类效果的影响情况,为实际参数选择提供参考依据。另外,对比 IFCM-RPD 聚类与基于现有距离公式的 IFCM 聚类效果。第二部分采用实测城市气温数据,通过基于现有主要区间数距离公式的 IFCM 聚类效果和 IFCM-RPD 聚类效果的对比,说明 IFCM-RPD 聚类算法的有效性。

4.1 合成数据的仿真实验分析

4.1.1 相对位置相异度公式中参数 h 的选择

本实验中,构造的数据集由 3 类二维区间向量组成,每类包括 100 个区间数向量,共 300 个区间数向量。构造区间数向量时,首先按照二维正态分布生成二维数据点 (x,y),再为点 (x,y) 随机分配区间长度 r_1、r_2,r_1、r_2 从区间 $[0,8]$ 随机选取,则数据点 (x,y) 对应的区间数向量为 $([x-r_1/2,x+r_1/2],[y-r_2/2,y+$

$r_2/2$]）。据此构造 300 个二维区间数向量,数据集各区间数向量的中心服从的正态分布参数如表 2 所示。

表 2　数据集参数

Tab. 2　Parameters of data sets

	数量	均值	方差	相关系数
Cluster_1	100	$\mu_1=28,\mu_2=23$	$\sigma_1=144,\sigma_2=16$	$\rho_{12}=0$
Cluster_2	100	$\mu_1=62,\mu_2=30$	$\sigma_1=81,\sigma_2=49$	$\rho_{12}=0$
Cluster_3	100	$\mu_1=50,\mu_2=15$	$\sigma_1=49,\sigma_2=81$	$\rho_{12}=0$

本实验中,采用文献[12]中的总异质性指数(overall heterogeneity index)R 来衡量聚类效果,指数 R 越大,聚类效果越好。首先对构造好的数据集进行基于欧氏距离的 IFCM 聚类,得到相应的总异质性指数 R1;同时对相同的数据集,固定参数 $w=1$,参数 h 的取值如表 3 所示,进行 IFCM-RPD 聚类并得到相应的总异质性指数 R2。重复聚类实验 100 次,重复实验时,重新随机选取区间数向量长度。相对位置相异度公式中参数 h 对聚类效果指标 R 的影响效果的统计结果如表 3 所示。

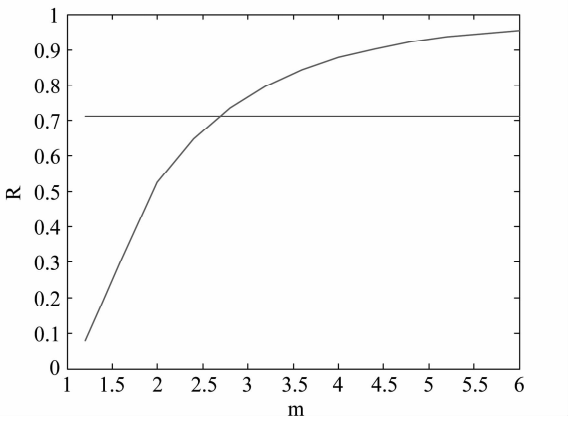


图 3　指标 R 随参数 h 变化趋势图

Fig. 3　Trend graph of index R with parameter h

由表 3 数据结果可知:对于合成的 100 个数据集,基于欧式距离的 IFCM 聚类的总异质性指数 R1 均值为 0.7110,方差为 0,即基于欧式距离的 IFCM 聚类的总异质性指数 R1 不随各区间数长度的变化而变化,保持为 0.7110;IFCM-RPD 聚类的总异质性指数 R2 随参数 h 的增大而增大,并且在参数 h 大于区间[2.4,2.8]中某个数时,R2>R1。因为对于相离的两区间数,其相对位置相异度会随着参数 h 的增大而增大,对于相交的两区间数,其相对位置相异度会随着参数 h 的增大而减小,故参数 h 越大,相对位置相异度公式对区间数集的区分能力越强,IFCM-RPD 的聚类效果也随着参数 h 的增大而增加。总异质性指数 R 随参数 h 的变化趋势图如图 3 所示。

4.1.2　距离公式对聚类效果影响的对比

该部分实验中,数据集构造方式与 4.1.1 部分的实验数据集构造方式一样。据此方式生成 4 个数据集,每个数据集包括 300 个区间数向量,并分为 3 类,每类数据的参数如表 4 所示。

对各合成数据集进行基于 Cityblock Distance 的 IFCM 聚类、基于 Euclidean Distance 的 IFCM 聚类、基于 Hausdorff Distance 的 IFCM 聚类、基于 Adaptive Quadratic Distance 的 IFCM 聚类和 IFCM-RPD 聚类运算,通过比较各聚类算法的总异质性指数 R,来比较各聚类算法的聚类效果。IFCM-RPD 聚类算法中相对位置相异度公式中参数 $h=5$、$w=1$。每次聚类运算时,数据集中各区间数向量的长度和初始簇中心都重新随机选取,并且重复 50 次聚类运算,求得这 50 次聚类运算的总异质性指数 R 的均值与方差,表 5 列出了不同数据集下,基于不同距离公式的 IFCM 聚类总异质性指数 R 及迭代次数的统计结果。

表 3　参数 h 与指标 R 关系

Tab. 3　Relation of parameter h and index R

参数 h	R1		R2		参数 h	R1		R2	
	均值	方差	均值	方差		均值	方差	均值	方差
1.2	0.7110	0.0000	0.0807	0.0000	4	0.7110	0.0000	0.8781	0.0012
1.6	0.7110	0.0000	0.3079	0.0015	4.4	0.7110	0.0000	0.9030	0.0010
2	0.7110	0.0000	0.5254	0.0007	4.8	0.7110	0.0000	0.9217	0.0008
2.4	0.7110	0.0000	0.6462	0.0011	5.2	0.7110	0.0000	0.9360	0.0007
2.8	0.7110	0.0000	0.7341	0.0014	5.6	0.7110	0.0000	0.9470	0.0005
3.2	0.7110	0.0000	0.7979	0.0015	6	0.7110	0.0000	0.9555	0.0004
3.6	0.7110	0.0000	0.8442	0.0014					

表 4 合成数据集参数

Tab. 4 Patterns of synthetic data sets

数据集 1	$\mu_1=28,\mu_2=22,\sigma_1^2=100,\sigma_2^2=9,\rho_{12}=0$ $\mu_1=60,\mu_2=30,\sigma_1^2=9,\sigma_2^2=144,\rho_{12}=0$ $\mu_1=45,\mu_2=38,\sigma_1^2=9,\sigma_2^2=9,\rho_{12}=0$	数据集 3	$\mu_1=28,\mu_2=22,\sigma_1^2=100,\sigma_2^2=9,\rho_{12}=0.8$ $\mu_1=60,\mu_2=30,\sigma_1^2=9,\sigma_2^2=144,\rho_{12}=0.7$ $\mu_1=45,\mu_2=38,\sigma_1^2=9,\sigma_2^2=9,\rho_{12}=0.6$
数据集 2	$\mu_1=28,\mu_2=23,\sigma_1^2=100,\sigma_2^2=9,\rho_{12}=0$ $\mu_1=62,\mu_2=30,\sigma_1^2=81,\sigma_2^2=16,\rho_{12}=0$ $\mu_1=50,\mu_2=15,\sigma_1^2=100,\sigma_2^2=16,\rho_{12}=0$	数据集 4	$\mu_1=28,\mu_2=23,\sigma_1^2=100,\sigma_2^2=9,\rho_{12}=0.7$ $\mu_1=62,\mu_2=30,\sigma_1^2=81,\sigma_2^2=16,\rho_{12}=0.8$ $\mu_1=50,\mu_2=15,\sigma_1^2=100,\sigma_2^2=16,\rho_{12}=0.7$

表 5 基于不同距离公式 IFCM 聚类效果统计表

Tab. 5 Clustering effect of IFCM clustering based on various distance formulas

距离公式	统计量	数据集 1	迭代次数	数据集 2	迭代次数	数据集 3	迭代次数	数据集 4	迭代次数
Cityblock 距离	均值	0. 5084	133	0.4538	60	0.4920	89	0.4748	48
	方差	0.0000	308.0	0.0000	11.6	0.0001	2724.6	0.0000	4.2
Euclidean 距离	均值	0. 6940	43	0.7259	68	0.6973	36	0.7480	41
	方差	0.0000	0.3	0.0000	4.2	0.0000	0.3	0.0000	0.3
Hausdorff 距离	均值	0.4160	135	0.3741	124	0.4087	110	0.3932	154
	方差	0.0000	513.4	0.0000	195.5	0.0001	142.3	0.0000	208.5
Adaptive quadratic 距离	均值	0.7521	200	0.7887	200	0.7722	200	0.7843	200
	方差	0.0009	0	0.0000	0	0.0018	0	0.0001	0
RPD 相异度	均值	0.9255	33	0.9088	55	0.9133	29	0.8964	37
	方差	0.0007	15.2	0.0007	19.5	0.0005	6.0	0.0005	7.3

由表 5 仿真结果可知:对数据集 1 进行 IFCM 聚类时,IFCM-RPD 聚类算法总异质性指数 R 的均值为 0.9255,方差为 0.0007,迭代次数的均值为 33,方差为 15.2。IFCM-RPD 聚类效果指数 R 的均值相对于表 5 中其他 IFCM 聚类算法的总异质性指数 R 具有明显改善,且迭代次数明显较其他 IFCM 聚类方法少。对数据集 2、3、4 进行 IFCM 聚类仿真时,IFCM-RPD 聚类算法的总异质性指数 R 比其他 IFCM 聚类算法的总异质性指数 R 同样具有明显改善,迭代次数明显较少。

由于区间数相对位置相异度会随着区间数间相离程度而呈指数增加,则某区间数离较近的某区间数的相异度与离较远区间数的相异度差值会较大,相异度能更清晰地描述该区间数离哪个簇中心较近,提高聚类判断的有效性,从而提高聚类效果。同时,聚类迭代过程中,区间数 k 离簇 i 越近,式(6)求得的 u_{ik} 会更趋近于 1,迭代过程的收敛速度更快,故 IFCM-RPD 聚类效果相对表 5 中其他 IFCM 聚类效果有较大提高,迭代收敛速度更快。

4.2 城市气温数据的仿真实验分析

文献[19]给出了 36 个城市 12 个月的温度区间,并且对这 36 个城市按照地理特征进行了划分。该部分仿真实验针对文献[19]的数据集,进行基于不同区间数距离公式的 IFCM 聚类,通过文献[20]中 Rand 指标进行聚类效果比较。Rand 指标描述了两个分类之间的相似性,Rand 指标越大,说明这两个分类的相似程度越大。表 6 列出了基于各区间数距离公式的 IFCM 算法的 Rand 指标。

表 6 基于各距离公式的 IFCM 聚类 Rand 指标

Tab. 6 Rand index of IFCM based on various distance formulas

距离公式类型	Rand 指标 M
Cityblock 距离	0.6647
Euclidean 距离	0.7425
Hausdorff 距离	0.7615
Adaptive quadratic 距离	0.6351
RPD 相异度	0.7679

由表 6 结果可知: IFCM-RPD 聚类算法的 Rand 指标 M 值为 0.7679,比基于其他距离公式的 IFCM 聚类算法的 Rand 指标 M 高。说明基于相对位置距离的 IFCM 算法的聚类结果与文献[19]给出的城市分类的

相似度最高,聚类效果最好。

5　总结

本文在分析了现有区间数距离公式存在无法描述区间数之间相对位置的缺点后,提出了相对位置相异度公式。并通过理论分析说明了该相异度公式不仅满足相异度公理,能定量描述区间数之间距离,还能描述区间数之间的相对位置。通过将相对位置相异度公式应用于 IFCM 聚类中,提出了 IFCM-RPD 聚类方法。

仿真实验结果表明:IFCM-RPD 聚类方法中,由于相对位置相异度公式为指数形式,参数 h 越大,相离区间数之间的相异度差异更大,相交区间数之间相异度差异更小,具有更好的相异度描述效果,故参数 h 在计算能力允许范围内应尽可能大;相对位置相异度公式相对于现有区间数距离公式具有更好的区间数相异度度量效果,可以提高 IFCM 聚类效果,且聚类收敛速度更快。

参考文献

[1] Yehia M, Chedid R, Ilic M, Zobian A, Tabors R, Lacalle-Melero J. A Global Planning Methodology For Uncertain Environments: Application to the Lebanese Power System [J]. IEEE Transactions on Power Systems, 1995, 10(1): 332-338.

[2] Bonissone PP, Dutta S and Wood NC. Merging Strategic and Tactical Planning in Dynamic and Uncertain Environments[J]. IEEE TRANSACTION ON SYSTEMS, MAN, AND CYBERNETICS, 1994, 24(6): 841-862.

[3] Shahi A, Atan Rodziah binti and Sulaiman MN. Decision Making for Uncertain Data in Dynamic Environment using Hybrid Method[C]. 2009 IEEE International Conference on Control and Automantion Christchurch, New Zealand, 2009. 9-11.

[4] Windhorst R, Field M, and Karahan S. Covective Weather avoidance with uncertain weather forecasts [C]. Digital Avionics Systems Conference, 2009. DASC'09. Orlando, 2009, 3. D. 4-1-3. D. 4-10.

[5] Hipel KW and Yakov Ben-Haim. Decision Making in an Uncertain World: Information-Gap Modeling in Water Resources Management [J]. IEEE TRANSACTION ON SYSTEMS, MAN, AND CYBERNETICS-PART C: APPLICATION AND REVIEWS, 1999, 29(4): 506-517.

[6] Batista DM and Fonseca Nelson L. S da. Scheduling Grid Task in Face of Uncertain Communication Demands[J]. IEEE TRANSACTION ON NETWORK AND SERVICE MANAGEMENT, 2011, .8(2): 92-103.

[7] Wang S, Wang GR, Gao X, Tan Z. Frequent Items Computation over Uncertain Wireless Sensor Network[C]. 2009 Ninth International Conference on Hybrid Intelligent System. Shenyang, 2009, 223-228.

[8] Charalambous CD, Farhadi A, Denic S and Rezaei F. Robust Control over Uncertain Communication Channels [C]. Proceedings of the 13th Mediterranean Conference on Control and Automation Limassol, Cyprus, 2005, 27-29.

[9] 宋成,王飞雪,庄钊文. 辅助型 GPS 接收机中载波频偏及其不确定度估计算法研究[J]. 信号处理. 2009, 25(11): 1694-1700.
Song Cheng, Wang Fei-xue, Zhuang Zhao-wen. An Estimate Algorithm for Carrier Frequency Shift and Its Uncertainty in the Assisted GPS Receiver[J], Signal Processing. 2009, 25(11): 1694-1700. (in Chinese)

[10] Peng Wei and Li Tao. Interval Data Clustering with Application[C]. Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence. 2006: 355-362.

[11] Francisco de A. T. de Carvalho, Fuzzy c-means clustering methods for symbolic interval data[J]. Pattern Recognition Letter, 2007, 28(6): 423-437.

[12] Chuang Chen-Chia, Jeng Jin-Tsong and Li Chih-Wen. Fuzzy C-Means Clustering Algorithm with Unknown Number of Clustering for Symbolic Interval Data[C]. SICE Annual Conference. Japan. 2008: 358-363.

[13] Francisco de A. T. de Carvalho and Yves Lechevallier. Dynamic Clustering of Interval-Valued Data Based on Adaptive Quadratic Distance[J]. IEEE TRANSACTION ON SYSTEMS, MAN AND CYBERNETICS-PART A: SYSTEMS AND HUMANS, 2009. 39(6): 1295-1306.

[14] Pimentel BA, Anderson F. B. F. da Costa, Renata M. C. R. de Souza. Kernel-Based Fuzzy Clustering of Interval Data. 2011 IEEE International Conference on Fuzzy Systems. Taipei, Taiwan. 2011: 497-501.

[15] Tran L, Duckstein L. Multi Objective Fuzzy Regression

with Central Tendency and Possibilistic Properties[J]. Fuzzy Sets and Systems. 2002,130(1):21-31.

[16] 李霞,张绍林,张淼,刘华. 基于新距离测度的区间数排序[J]. 西华大学学报,2008,27(1):87-90.
Li Xia, Zhang Shao-lin, Rank of Interval Numbers Based on a New Distance Measure[J]. JOURNAL OF XIHUA UNIVERSITY. 2008. 27(1):87-90. (in Chinese)

[17] A. Sengupta, T. K. Pal. On comparing interval numbers. European Journal of Operational Research. 2000,127(1): 28-43.

[18] L. Nieddu, A. Rizzi. Proximity measures in symbolic data analysis. STATISTICA. 2003. 63(2):195-211.

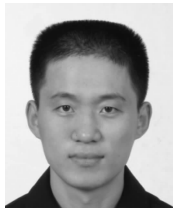
[19] Guru DS, Kiranagi BB and Nagabhushan P. Multivalues type proximity measure and conceptes of mutual similarity value for clustering symbolic patterns[J]. Pattern Recognit. Letter,2004,25(10):1203-1213.

[20] Hubert L and Arabie P. Comparing Partitions[J]. Journal of Classification 1985,2(1):193-218.

作者简介

袁 飞(1984-),男,湖南常德人,现为中山大学信息科学与技术学院博士研究生,主要研究方向:不确定数据管理、智能信息处理。
E-mail:eric_f_y@ foxmail. com

詹宜巨(1955-),男,博士,教授、博士生导师,主要研究方向:自动识别技术、物联网以及智能信息处理。
E-mail:zhanyiju@ mail. sysu. edu. cn



王永华(1979-),男,博士,讲师,主要研究方向:物联网、RFID 智能识别以及智能信号处理。
E-mail:wangyonghua@ gdut. edu. cn