基于联系数的位置不确定性数据 UCNK-Means 聚类算法

王 骏 黄德才
(浙江工业大学计算机科学与技术学院 杭州 310023)

摘 要 摘要位置不确定性数据的聚类是一个新的不确定性数据聚类问题。其聚类方法主要包括获取对象的概率密度函数,通过积分计算对象间的期望距离来进行聚类分析和以区间数表示对象,通过区间数的系列运算来进行聚类分析这两大类。前者存在概率密度函数获取困难、计算复杂、实用性不强的缺陷;后者在区间数转化为实数过程中,忽略了区间数变化范围对聚类效果的影响,其聚类质量不佳。鉴于此,提出一种基于联系数的不确定对象聚类新算法 UCNK-Means。该算法用联系数巧妙地表示不确定性对象,并专门定义了对象间的联系距离,运用联系数态势值比较联系距离大小,克服了现有算法的不足。仿真实验表明,UCNK-Means 具有聚类精度高、计算复杂度低、实用性强的特点。

关键词 不确定性数据,联系数,聚类,数据挖掘
中图法分类号 TP181 **文献标识码** A

UCNK-Means Clustering Method for Position Uncertain Data Based on Connection Number

WANG Jun HUANG De-cai
(College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract Clustering for position uncertain data is a new problem of uncertain data clustering. Mainly there are two solutions to this new problem. The first is clustering acquiring the probability density function or probability distribution function of uncertain object and getting the expected distance with integral operation. The second is clustering by series of operation of interval data. However, the former has the disadvantages of getting probability density function hard, complex operation and poor practicability, and the latter ignores the effect of the range of interval data to the result of clustering. Therefore, a new uncertain data clustering method UCNK-Means was put forward. This method uses connection number as the model of uncertain object and defines connection distance between two objects and uses the situation-value to compare the connection distance, which overcome the disadvantages existed in the two solutions above. The experiment illustrates that UCNK-Means has high precision of clustering, low complexity and strong praticability.

Keywords Uncertain data, Connection number, Clustering, Data mining

1 引言

在现实生活中,数据采集工作由于在不同的应用领域常常因为存在误差或人为干扰带来差错或受到数据采集设备的限制,获取到的数据具有不确定性[1]。

从不确定性数据的表现形式,大致可将其分为实例存在不确定性和属性不确定性两类。

实例存在不确定性(Existential Uncertainty)[2]也称元组存在不确定性或数据对象存在不确定性,即一个数据元组存在与否具有一定的不确定性,通常用一个概率值来表示。这种不确定性又分为实例之间存在相互依赖关系和实例之间相互独立两种情况。实例存在不确定性通常简称为存在不确定性。

一个元组的存在不确定性一般可采用点概率模型(point probability case)。在该模型中,元组的属性值是确定的,而元组的存在具有不确定性,并用一个[0,1]之间的概率值来表示。

属性级不确定性(Attribute Level Uncertainty)[3]也称属性值的不确定性,它是指一个属性的取值存在不确定性。在元组数目和数据模型已经确定的前提下,通常用概率密度函数或其它统计量来描述属性的不确定信息。这种不确定性又可以根据属性的类型分为数值不确定性和非数值不确定性。

本文研究的数据是数值属性级不确定性数据中的一种,也叫做位置不确定性对象,它区别于传统的数值属性级不确定性数据。例如,在一个野生动物园中,每个动物身上都装有发射该动物当前位置信号的设备,这些被发射的信号能够被收集到动物园控制中心的服务器中,并以一个数据点的形式显示在一张三维图中。服务器会每隔一段时间收集一次动物的位置信号,然后追加显示在图中。对于某一个动物 A 而言,在某一个时刻 B 所发射的位置信号会以圆点形式显示在

本文受水利部公益性行业科研专项(201401044)资助。
王 骏(1990-),男,硕士生,主要研究方向为数据挖掘;黄德才(1958-),男,博士,教授,博士生导师,主要研究方向为数据挖掘、人工智能等,E-mail:hdc@zjut.edu.cn(通信作者)。

图中，代表动物 A 在 B 时刻出现在了图中圆点处的位置。因此，在一个星期内，动物 A 的信号能够在图中形成许多圆点，这些圆点形成的区域代表了这个星期内动物 A 所出现过的位置，即动物 A 这个星期内的活动区域。同样，其他动物也能在图中形成相应的位置区域，则一个动物代表了一个位置不确定性对象，对这些区域，即位置不确定性对象进行聚类分析，能够揭示这些动物间可能的族群关系或者亲密关系。

对于这一类位置不确定性对象的聚类分析，目前普遍的做法是：文献[4-10]用概率密度函数 pdf（probability density function）来表示对象，然后通过积分来求得对象间的期望距离 EED（Expected Euclidean Distance），从而进行聚类分析。这一类处理方法虽然理论上较为精确，但是实际操作难度较大，积分计算复杂度高，现实中想要得到准确的概率密度函数难度非常大。除此之外，文献[11]采用区间数表示对象，通过区间数的四则运算，得出一个区间数表示对象间的距离，最后把区间数转换为一个确定性数值来进行聚类分析，即 UIDK-Means 算法。该算法大大减小了计算期望距离 EED 所带来的复杂度；但是该方法最后将一个区间数转化为一个确定性数值来表示对象间的距离，没有额外考虑区间范围变化。例如，区间数$[a_1,b_1]$，$[a_2,b_2]$分别表示对象 O_1 与 O_2 的距离和对象 O_1 和 O_3 间的距离，UIDK-Means 算法将两个区间数通过 $\lambda*a_1+(1-\lambda)*b_1$，$\lambda*a_2+(1-\lambda)*b_2$ 公式计算得两个确定性数 dis_1，dis_2，通过比较 dis_1，dis_2 的大小来决定对象 O_1 更接近 O_2 还是 O_3。这时可能会出现 $d_1=d_2$，但是两个区间数范围不一致的情况；如取 $\lambda=0.5$，$[a_1,b_1]$，$[a_2,b_2]$分别为$[1,5]$，$[2,4]$，则计算得 $dis_1=dis_2=3$，按照 UIDK-Means 方法，O_1 与 O_2，O_1 与 O_3 两者的距离一样，但实际上，O_1 与 O_2 的距离有可能达到 1，比 O_1 与 O_3 的亲密度更进一步，因此，UIDK-Means 算法对于以上这一种情况的聚类精度有待改进。

鉴于以上问题，本文提出一种新的不确定性数据聚类算法 UCNK-Means（Uncertain Connection NumberK-Means），该方法用联系数[12]表示对象，并自定义联系数运算法则，新的对象间距离-联系距离，以及新的对象与簇间距离衡量标准，然后根据联系数的态势值理论来衡量联系距离的大小，从而对不确定性对象进行聚类。经过实验仿真发现该新方法不但大大减小了计算复杂度，而且在聚类精度方面也有提高，聚类效果优良，现实可操作性强。

2 相关定义

本节对本文提出的算法中所用到的一些概念进行定义。

2.1 不确定性对象

对于不确定性对象，如下定义。

定义 1 对象$\overline{O_p}=(O_{p1},O_{p1},\cdots,O_{pm})$是一个 n 维的不确定性对象（$p=1,2,\cdots,m$），其中 $O_{pq}=\{o_{pq1},o_{pq2},\cdots,o_{pqf(p,q)}\}(q\in\{1,2,\cdots,n\})$表示维度 q 上有着 $f(p,q)$个不确定的离散取值，$f(p,q)$是关于 p,q 的函数，值随着 p,q 的变化而变化。对任意的 $h\in\{1,2,\cdots,f(p,q)\}$，$o_{pqh}$ 表示对象$\overline{O_p}$在第 q 维上的第 h 个取值。

如：在一个野生动物园中，一只会飞的鸟就是一个不确定性对象，记作$\overline{O_1}=(O_{11},O_{12},O_{13})$，其中 $O_{11}=\{o_{111},o_{112},\cdots,o_{11f(1,1)}\}$表示这只鸟在三维空间中位置投射到 x 轴上对应的 $f(1,1)$个数值，$\{o_{111},o_{112},\cdots,o_{11f(1,1)}\}$中每一个值代表了 x 轴上的一个数值；同理，O_{12} 与 O_{13} 分别为这只鸟在三维空间中位置投射到 y 轴与 z 轴上的数值。

2.2 联系数

联系数理论[12]是在集对分析 SPA（Set Pair Analysis）[13-16]理论的基础上延伸出来的一种新的处理不确定性问题的新理论方法和数学工具。它是集对中衡量不确定的一个重要指标。所谓集对实际是指具有联系的两个集合组成的对，这两个集合对具有同一性（同）、差异性（异）和对立性（反）3 个属性，并通过一个三元联系数来描述这种同异反的特性。引进联系数的最初目的是为了应用的方便，其理论意义则在于拓广了数的概念。一方面它把可确定数与所在范围的数与值联系起来，另一方面它把宏观层次上的确定量和微观层次上的不确定量联系起来。虽然联系数 $u=a+bi+cj$ 的表达形式是统一的，但 a,b,c 和 i,j 的语义可根据所研究问题的实际背景具体定义和解释。

目前，联系数已经取得了广泛的应用，如文献[17]用联系数来解决属性值为区间数的多属性决策问题；文献[18]用联系数来解决准则权重不完全确定的群决策问题；文献[19]用联系数解决了信任表达中的模糊性和不确定性难题，为主观信任评价和决策研究提供了一种有价值的新思路；文献[20]用联系数来很好地融入考虑等级标准边界的模糊性，从而给出水资源系统评价的新方法。

以下定义联系数的基本概念，以及联系数的一些运算。

定义 2 设用 u 表示二元联系数[18]，则有：

$$u=a+bi \tag{1}$$

其中，a 称为确定性联系分量，或者同一度，b 称为不确定性联系分量，也称为差异度，i 是不确定性联系分量系数，$i\in[-1,1]$，a 为实数，b 为非负实数，统称为联系数 u 的联系分量。

定义 3 设用 u 表示三元联系数[18]，则有：

$$u=a+bi+cj \tag{2}$$

其中，a,b 与定义 1 中二元联系数一样，分别为确定性联系分量和不确定性联系分量；j 表示“反”，表示 c 与 a 的取值“相反”，对立，c 称为对立度联系分量；a 为实数，b,c 为非负实数。

所谓对立，当 c 与 a“正负对立”，$j=-1$；当 c 与 a“倒数对立”，$j=1/a$；有“虚实对立”，$j=\sqrt{-1}$。本文联系数的对立类型属于正负对立。

对于二元联系数的四则运算，有如下定义。

定义 4 设有两个联系数，$u_1=a_1+b_1i$，$u_2=a_2+b_2i$，则二元联系数加法[12]如下：

$$\begin{aligned}u_1+u_2&=(a_1+b_1i)+(a_2+b_2i)\\&=(a_1+a_2)+(b_1+b_2)i\end{aligned} \tag{3}$$

因此，联系数之和仍然是一个联系数。

如：$u_1=1+2i$，$u_2=2+3i$，则 $u_1+u_2=(1+2i)+(2+3i)=(1+2)+(2+3)i=3+5i$。

定义 5 设有两个联系数，$u_1=a_1+b_1i$，$u_2=a_2+b_2i$，则二元联系数减法如下：

$$\begin{aligned}u_1-u_2&=(a_1+b_1i)-(a_2+b_2i)\\&\triangleq(a_1-a_2)+(b_1+b_2)i\end{aligned} \tag{4}$$

在文献[12]中，将 u_1-u_2 定义为$(a_1-a_2)+(b_1-b_2)i$；本文为了保留数据不确定性部分，对于联系数的减法，定义其不确定性分量相加，而不是相减。

因此，联系数之差仍然是一个联系数。

如：$u_1=1+2i, u_2=2+3i, u_1-u_2=(1+2i)-(2+3i)=(1-2)+(2+3)i \triangleq -1+5i$。

定义 6 设有两个联系数，$u_1=a_1+b_1i, u_2=a_2+b_2i$，则二元联系数乘法为：

$$\begin{aligned} u_1\times u_2&=(a_1+b_1i)\times(a_2+b_2i)\\ &=a_1a_2+(a_1b_2+a_2b_1)i+b_1b_2i^2\\ &\triangleq a_1a_2+b_1b_2i \end{aligned} \tag{5}$$

注意，在文献[12]中将$(a_1+b_1i)\times(a_2+b_2i)$定义为$(a_1a_2+b_1b_2)+(a_1b_2+a_2b_1)i$，但本文首先将式子$(a_1+b_1i)\times(a_2+b_2i)$展开为$a_1a_2+(a_1b_2+a_2b_1)i+b_1b_2i^2$，然后，保留$i$的最高次幂项和$i$的0次幂项，舍去其他部分，即$a_1a_2+b_1b_2i^2$，最后将$i$的最高次幂还原为$i$的1次幂，即$a_1a_2+b_1b_2i$。

因此，联系数之积仍然是一个联系数。

如 $u_1=1+2i, u_2=2+3i, u_1\times u_2=(1+2i)\times(2+3i)=2+(3+4)i+(2\times3)i^2 \triangleq 2+6i$。

定义 7 设一个联系数，$u=a+bi$，则二元联系数根号运算为：

$$\sqrt{u}=\sqrt{(a+bi)}\triangleq\sqrt{a}+\sqrt{b}i \tag{6}$$

即只对二元联系数的确定性分量和不确定性分量直接进行根号运算，计算值直接作为结果的确定性分量和不确定性分量部分。

因此，联系数根号运算后仍然是一个联系数。

如：$u=4+9i$，则$\sqrt{u}=\sqrt{4+9i}\triangleq 2+3i$。

定义 8 给定一个三元联系数，$u=a+bi+cj$，则三元联系数的态势值记为：

$$Shi(u)=\frac{a}{c} \tag{7}$$

其为该联系数的态势。通常来说，$Shi(u)$越大，代表该联系数u所表示的不确定值平均值越大。

态势值理论是根据联系数的态势函数表思想转化而来：联系数的态势函数是联系数中联系分量大小关系的一种数学表达式，利用联系数的态势函数，有时可以从整体上比较联系数所描述问题的发展“势头”，如表1是三元联系数的态势函数表。

表1 三元联系数态势函数表

联系数	联系分量大小关系	联系数态势
a+bi+cj	a>c	同势
	a=c	均势
	a<c	反势

态势表是通过比较一个联系数中的确定性联系分量与对立性联系分量的大小来判断态势；以一个表示两个不确定性对象间距离的联系数举例，它的“势头”就是距离的大小。“同势”表示“势头”发展向确定性联系分量处发展，也即距离很大；“均势”表示“势头”发展的确定联系分量与对立性联系分量势均力敌，也即距离一般；“反势”表示“势头”向对立分量发展，也即距离偏小。

本文直接用一个联系数中的确定性联系分量a与对立性联系分量c的商值来表示该联系数的态势值。

因此，如果给定一个联系数，表示两个对象的距离，则可以用该联系数的态势值$Shi(u)$来反映其平均大小，这既解决了不确定性数据的一个大小比较，又解决了由多个样本点求平均值的计算复杂问题。

如：$u=1+2i+3j$，则该u的态势值就为$1/3=0.333$。

2.3 位置不确定性数据的联系数表示

以下定义给出不确定性数据的联系数表示法。

定义 9 对于定义1中的不确定对象$\overline{O_p}=(O_{p1},O_{p2},\cdots,O_{pm})$，其中$O_{pq}=\{o_{pq1},o_{pq2},\cdots,o_{pqf(p,q)}\}$，可以用一个联系数向量表示该对象，即$\overline{O_p}=(u_{p1},u_{p2},\cdots,u_{pm})$，其中$u_{pq}=a_{pq}+b_{pq}i$表示维度$q$的联系数；在$O_{pq}=\{o_{pq1},o_{pq2},\cdots,o_{pqf(p,q)}\}$中找出最大值$\max_{pq}$和最小值$\min_{pq}$，使得$\min_{pq}\leqslant p_{pqh}\leqslant\max_{pq}$，其中$h\in\{1,2,\cdots,f(p,q)\}$，则$a_{pq}=(\max_{pq}+\min_{pq})/2$，$b_{pq}$表示第$q$维度上的值变化半径，即$b_{pq}=\max_{pq}-a_{pq}$。

如：给出对象$\overline{O_p}=(\{1,2,3\},\{2,4,5\})$，则对应的$O_{p1}=\{1,2,3\}$，$O_{p2}=\{2,4,5\}$，则可以转换为联系数来表示对象，$\overline{O_p}=(u_{p1},u_{p2})$，$u_{p1}=a_{p1}+b_{pl}i$，$u_{p2}=a_{p2}+b_{p2}i$，$\max_{p1}=3$，$\min_{p1}=1$，则$a_{p1}=(1+3)/2=2$，$b_{p1}=3-2=1$，$u_{p1}=2+i$；同理可得，$u_{p2}=3.5+1.5i$；因此$\overline{O_p}=(2+i,3.5+1.5i)$。

定义 10 给定两个对象，$\overline{O_1}=\{u_{11},u_{12},\cdots,u_{1n}\}$，$\overline{O_2}=\{u_{21},u_{22},\cdots,u_{2n}\}$，则两个对象间的联系距离为：

$$rd(\overline{O_1},\overline{O_2})=\sqrt{(u_{11}-u_{21})^2+(u_{12}-u_{22})^2+\cdots+(u_{1n}-u_{2n})^2} \tag{8}$$

根据式(4)可将式(8)式化解为$\sqrt{w_1^2+w_2^2+\cdots+w_n^2}$；再根据式(5)，可继续将式子化解为$\sqrt{r_1+r_2+\cdots+r_n}$；再根据式(3)，将式子化解为$\sqrt{z}$；再根据式(6)，可最终将式子化解成为一个二元联系数。

因此，两个不确定性对象间的联系距离是一个二元联系数。

定义 11 给出一个对象集合$set(O)=\{\overline{O_1},\overline{O_2},\cdots,\overline{O_m}\}$，根据式(8)计算两两对象间的联系距离$rd$，则可得联系数集合$set(u)=\{u_{12},u_{13},\cdots,u_{1m},u_{23},\cdots,u_{2m},\cdots,u_{(m-1)m}\}$，根据定义9中，对于每一个$u_{(p1)(p2)}=a_{(p1)(p2)}+b_{(p1)(p2)}i(p1,p2\in\{1,2,\cdots,m\})$记$\max_{(p1)(p2)}=a_{(p1)(p2)}+b_{(p1)(p2)}$，对于整个$set(u)$集合中找出最大的$\max_{(p1)(p2)}$记作$\max_{set}$，该$\max_{set}$为该对象集合$set(O)$的最大距离上界；则记：

$$abmax_{set}=\max_{set}+\max_{set}*factor, factor>0 \tag{9}$$

称$abmax_{set}$为当前对象集的绝对距离上界(这里$factor$只要不为0即可，效果无多大差别)。

如：在一个联系数集合$set(u)=\{1+2i,2+4i,3+2i,5+6i\}$中，$\max_{set}=11$(对应联系数$5+6i$)，则$abmax_{set}=11+11*factor$，若取$factor=0.3$，则$abmax_{set}=11+11*0.3=14.3$。

定义 12 可以把二元联系数转换为三元联系数，对于$u_1=a_1+b_1i$转换为$u_2=a_2+b_2i+c_2j$，其中$a_2=a_1, b_2=b_1, c_2=abmax_{set}-(a_1+b_1)$。

如：给出一个二元联系数$1+2i$，假设$abmax_{set}=5$，则对应三元联系数为$1+2i+(5-(1+2))j=1+2i+2j$。

2.4 对象与簇

定义 13 给定一个簇$C_y=\{O_1,O_3,O_4\}$，一个对象O_2，根据式(7)可以求得$Shi_{12}, Shi_{23}, Shi_{24}$，则记：

$$Shi_{C_yO_2}=(Shi_{12}+Shi_{23}+Shi_{24})/3 \tag{10}$$

表示对象O_2与簇C_y间的态势值。

一个对象与自身的态势值记为0，即如给定一个簇$C_y=\{O_1,O_2,O_3,O_4\}$，一个对象O_2，则记：

$$Shi_{C_yO_2}=(Shi_{12}+Shi_{23}+Shi_{24}+0)/4 \tag{11}$$

表示对象 O_2 与簇 C_y 间的态势值。

定义 14 给定一个簇 $C_y=\{O_1,O_3,O_4\}$，一个对象 O_2，根据式(8)可以求得对象 O_2 与 C_y 簇内各个对象间的联系距离，$u_{12}=a_{12}+b_{12}i$，$u_{23}=a_{23}+b_{23}i$，$u_{24}=a_{24}+b_{24}i$，则记：

$$b_{C_yO_2}=(b_{12}+b_{23}+b_{24})/3 \tag{12}$$

表示对象 O_2 与簇 C_y 间的距离变化范围。

3 UCNK-Means 聚类方法

3.1 算法描述

由定义 1 的对象表示格式，给出不确定性数据集 $set(O)=\{\overline{O_1},\overline{O_2},\cdots,\overline{O_m}\}$，给出 m 个不确定性对象，每个对象都为 n 维的数据，运用 UCNK-Means 方法对其进行聚类，设分为 k 个类。

步骤 1 根据定义 9，把该对象集表示成联系数形式，即

$\overline{O_1}=\{u_{11},u_{12},\cdots,u_{1n}\}$

$\overline{O_2}=\{u_{21},u_{22},\cdots,u_{2n}\}$

…

$\overline{O_m}=\{u_{m1},u_{m2},\cdots,u_{mm}\}$

步骤 2 运用定义 10 中式(8)，计算数据集中对象之间两两的联系距离，生成对象间的联系距离矩阵：

$$\begin{pmatrix} - & rd_{12} & \cdots & rd_{1m} \\ rd_{21} & - & & rd_{2m} \\ \vdots & & \ddots & \vdots \\ rd_{m1} & rd_{m2} & \cdots & - \end{pmatrix} \tag{13}$$

如 rd_{12} 表示对象 $\overline{O_1}$ 与对象 $\overline{O_2}$ 间的联系距离，其值为一个二元联系数。

步骤 3 根据式(13)矩阵中对象间联系距离(二元联系数)，按照定义 11 的式(9)，找出绝对距离上界，并根据定义 12 的式(13)矩阵更新对象间联系距离(三元联系数)。

步骤 4 根据式(13)矩阵以及式(7)，求得对象间的态势值矩阵：

$$\begin{pmatrix} - & Shi_{12} & \cdots & Shi_{1m} \\ Shi_{21} & - & & Shi_{2m} \\ \vdots & & \ddots & \vdots \\ Shi_{m1} & Shi_{m2} & \cdots & - \end{pmatrix} \tag{14}$$

步骤 5 选取 k 个初始簇中心，然后根据式(14)矩阵，将其余各个对象与初始簇中心对象间的态势值小的分配给该初始簇中心所在的簇。

步骤 6 如果对象 O_{p1} 与 O_{p2}，O_{p1} 与 O_{p3} 的态势值相等，则直接比较两组对象间三元联系数的 b_{p1p2}，b_{p1p3} 的值大小，可以通过查询式(14)矩阵得到；变化范围大的两个对象归为一类。

因为 b 表示不确定性部分，所以 b 越大，表示两个对象间距离的变化范围越大，即两者靠近的可能性越大，如图 1 所示。

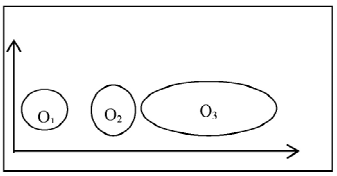

图 1 不确定变化范围与亲密性

在图 1 中，对象 O_1 与对象 O_2 的态势和对象 O_2 与对象 O_3 的态势相等，但是对象 O_2 与对象 O_3 的三元联系数 b 比较大，同时也可以看到，O_2 与 O_3 两个对象接近的概率比较大，因此 O_2 与 O_3 应该归为一类。

此步骤也是本算法 UCNK-Means 区别于 UIDK-Means 算法的关键一步，UIDK-Means[11] 引进参数 λ 后，相当于在区间数中指定了某一位置处的值来代替整个区间数，如 $\lambda=0.5$，就是取了区间数的中间位置的数；于是就会发生区间大小不同，但是区间数的中点值相等的情况，UIDK-Means 无法区分该情况。

步骤 7 第一次迭代后，已经形成了 k 个簇，重新计算每个对象与各个簇之间的“距离”，这个距离根据定义 13 来计算，然后记录在对象与簇之间的态势值矩阵中。

$$\begin{pmatrix} Shi_{O1C1} & Shi_{O1C2} & \cdots & Shi_{O1Ck} \\ Shi_{O2C1} & Shi_{O2C2} & \cdots & Shi_{O2Ck} \\ \vdots & & \ddots & \vdots \\ Shi_{OmC1} & Shi_{OmC2} & \cdots & Shi_{OmCk} \end{pmatrix} \tag{15}$$

然后，根据定义 14，计算对象与各个簇之间的距离变化范围，记录在对象与簇间距离变化矩阵中：

$$\begin{pmatrix} b_{O1C1} & b_{O1C2} & \cdots & b_{O1Ck} \\ b_{O2C1} & b_{O2C2} & \cdots & b_{O2Ck} \\ \vdots & & \ddots & \vdots \\ b_{OmC1} & b_{OmC2} & \cdots & b_{OmCk} \end{pmatrix} \tag{16}$$

则根据式(15)矩阵重新分配每个对象到与簇的态势值较小的簇，如果遇到对象与两个簇的态势值相等，则根据式(16)矩阵，将与簇的距离变化范围大的分配给该簇；如果连距离变化范围也相同，则任意选择一个簇。

步骤 8 继续迭代，直到聚类结果不再变化，或者迭代次数超过指定次数，结束，输出最终结果。

算法 1 UCNK-Means 算法

输入：不确定性数据集 set(O)，簇数目 k，初始簇中心 $\{O_1,O_2,\cdots,O_k\}$

输出：k 个簇 $set(C)=\{C_1,C_2,\cdots,C_k\}$

Step1 for set(O)中的每一个对象 O_p

用联系数表示对象 $O_p=\{u_{p1},u_{p2},\cdots,u_{pn}\}$；

endfor

Step2 for set(O)中的每一个对象 O_p

计算与 set(O)中其余对象间的态势值，记录于式(14)的矩阵中；

endfor

Step3 for set(O)中的每一个对象 O_p(除了被选为初始簇中心的对象)

根据式(14)矩阵，分配每个对象到距离初始簇中心态势值小的簇中，与之归为一类；更新 set(C)；

endfor

Step4 Repeat

for set(O)中的每一个对象 O_p

for $\{C_1,C_2,\cdots,C_k\}$ 中每一个 C_y

根据式(10)、式(11)，计算 O_p 与 C_y 的态势值，记录在式(15)矩阵中；根据式(12)，计算 O_p 与 C_y 的距离变化范围，记录在式(16)矩阵中；

endfor

endfor

Step5 for set(O)中的每一个对象 O_p

根据式(15)矩阵，重新分配对象 O_p 到簇，根据 O_p 到簇的态势值最小原则；

If O_p 到簇 C_{y1}，C_{y2} 态势值相等；

根据式(16)矩阵，将 O_p 到簇 C_{y1}，C_{y2} 的距离变化范围 b_{y1}，b_{y2} 大的，归为一类；

endif

更新 set(C)；

endfor

Step6 Until(超过迭代次数或聚类结果不再变化)

Step7 返回 set(C)；

3.2 计算复杂度分析

由于目前一些主流算法都是基于 k-Means 算法的，包括本文的 UCNK-Means 算法，因此它们的时间复杂度都是一样的，为 $O(tKn)$；其中 n 是数据的总数，k 是聚类的数目，t 是算法循环的次数。

因此，本文主要比较算法过程中的计算量的复杂度。

在目前存在的一些算法当中，大多采用给定概率密度，通过积分运算来求值的方式，如文献[4-9]中指出较为典型的 UK-Means 用概率函数 pdf 表示一个不确定性对象，若要计算两个对象 $\overline{O_1}$，$\overline{O_2}$ 间的距离，则是通过下式：

$$E(d(\overline{O_1},\overline{O_2}))=\int_{UD(\overline{O_1})}\int_{UD(\overline{O_2})} D(x_1,x_2)f_1(x_1)f_2(x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \tag{17}$$

其中，$f_1(x_1)$，$f_2(x_2)$ 表示连续变量 x_1，x_2 所对应的概率密度。

可以发现，如果对象的维度是高维且为连续区域，则涉及到多维向量积分，计算复杂度不言而喻；区域是离散的样本点，则根据下式：

$$E(d(\overline{O_1},\overline{O_2}))=\sum_{t=1}^{|\overline{O_1}|}\sum_{z=1}^{|\overline{O_2}|} d(x_t,x_z)F(x_t)F(x_z) \tag{18}$$

来计算两个对象间的距离，其中 $F(x_t)F(x_z)$ 表示离散变量 x_tx_z 所对应的概率密度。

这时可以发现，如果 O_1 有 T 个样本点，O_2 有 Z 个样本点，则计算一次 $E(d(\overline{O_1},\overline{O_2}))$ 需要进行 $T*Z*2$ 次乘法，$T*Z+T-2$ 次加法。

另外，文献[11]指出基于区间数的方法 UIDK-Means 来处理。该方法在给定一个对象各个维度上的区间表示后，计算两个对象 O_1，O_2 时，需要先计算出两个对象间各个维度上的距离最大最小值，一个维度需要两次加减；然后再通过平方和求得两个对象的距离区间数。因此，对于两个 2 维的对象来说，计算一次两个对象的区间数，总共需要 6 次加法、4 次乘法、2 次根号运算。

此外，UIDK-Means 算法在得到对象每个维度上的区间数 $[x-r,x+r]$ 时，采用该维度上的均值作为 x，以该维度上的标准差作为 r；假设取该区间上的样本点数为 n，则计算标准差需要 n 次减法、$n-1$ 次加法、n 次乘法、1 次根号运算。

而对于本文提出的 UCNK-Means 算法，在计算复杂度方面有显著的提升。当给定了两个用联系数表示的对象后，计算一次两个对象间的距离，假设两个对象是 2 维的，则需要 4 次加减、4 次乘法、4 次根号运算；与 UIDK-Means 相比，少两次加法，多了两次根号运算；但是，在对象表示上面，用联系数表示对象时，某一维度上，$a+bi$，同样用该维度的均值作为 a，但是不用计算标准差，直接用该维度上区间最大值减去 a 作为 b，因此只需要做一次加减；并且在后续算法迭代时，UIDK-Means 需要不断计算标准差与均值，而 UCNK-Means 只需要简单地通过查询第一阶段计算好的态势计算均值即可。具体对比如表 2 所列。

表 2 计算一次对象间距离的复杂度(2 维)

	UK-Means	UIDK-Means	UCNK-Means
乘法	T*Z*2	4	4
加减	T*Z+T-2	6	4
根号	—	2	4

注：T，Z 为两个对象的样本点数。

表 3 对象准备阶段的计算复杂度(2 维)

	UIDK-Means	UCNK-Means
乘法	n+2	2
加减	2n	2
根号	1	0

注：n 为一个对象内的样本点数。

由此可以看出，UIDK-Means 与 UCNK-Means 两个算法的计算复杂度远远低于 UK-Means，无论 UK-Means 是否按照连续区域来计算；其次，UCNK-Means 在对象表示以及算法迭代中的计算复杂度远远低于 UIDK-Means。因此，在计算复杂度上，UCNK-Means 算法有着独到的优势。

4 UCNK-Means 实验及性能

本节实验主要对 UCNK-Means 进行 Iris，Wine，Glass 3 个数据集上的聚类实验，然后再给出自定义的不确定性数据集进行实验；实验与 K-Means 和 UIDK-Means 进行比较，并用 F-measure[8,9,21] 来进行评估。

F-measure(范围是[0,1])是一种用来度量聚类结果之间相似度的常用指标[12]，较高的 F-measure 值表明两个聚类结果具有很高的相似度。

4.1 UCI 数据集上的聚类精度

UCI 机器学习数据集有 Iris，Wine 和 Glass 数据集，其中 Iris(Iris 是一种鸢尾属植物)数据集总共有 150 个数据点，每个数据点含有 Iris 花的萼片长度、萼片宽度、花瓣长度和花瓣宽度等 4 种属性；Wine 数据集描述 3 种不同意大利葡萄酒的化学分析结果；Glass 数据集通过 10 种化学成分的值来描述每一种玻璃。表 4 列出了这 3 种数据集的主要特性。

表 4 UCI 数据集

数据集	数据项个数	属性个数	类别数
Iris	150	4	3
Wine	178	13	3
Glass	214	10	6

利用 F-Measure 来评估 K-Means，UIDK-Means，UCNK-Means 的精度，并进行对比；K-Means 算法直接对 UCI 数据集进行聚类，对于 UIDK-Means，UCNK-Means 算法采用先对 UCI 数据进行不确定性化，方法如下。

对于一个对象 O_p，原先是一个确定性点，假设为 2 维数据，则可表示为 (x_p,y_p)，对每一个维度进行不确定化，转化为一个区间数 $[a_p-b_p,a_p+b_p]$，其中 a_p 为该区间的中点，b_p 为该区间半径；如对于 x_p 而言，在区间 $[x_p-0.2*x_p,x_p+0.2*x_p]$ 间产生一个随机数作为 a_{px}，然后在区间数 $[0.2*x_p,0.3*x_p]$ 间产生一个随机数作为 b_{px}，则 x_p 的不确定性范围可表示为 $[a_{px}-b_{px},a_{px}+b_{px}]$，同理 y_p 可转化为 $[a_{py}-b_{py},a_{py}+b_{py}]$；因此对于 O_p 对象不确定性化结束。

按照以上要求，对 Iris，Wine 和 Glass 数据集进行聚类，进行 20 次聚类(对于 UIDK-Means，UCNK-Means 每次聚类

都用不同的不确定性化后的数据)，取平均值，结果如表 5 所列。

表 5　UCI-Iris(certain-uncertain)数据集

	K-Means	UIDK-Means	UCNK-Means
数据集	certain	uncertain	uncertain
聚类精度 (F-Measure)	0.8917	0.7675	0.8279

表 6　UCI-Wine(certain-uncertain)数据集

	K-Means	UIDK-Means	UCNK-Means
数据集	certain	uncertain	uncertain
聚类精度 (F-Measure)	0.80223	0.5502	0.6950

表 7　UCI-Glass(certain-uncertain)数据集

	K-Means	UIDK-Means	UCNK-Means
数据集	certain	uncertain	uncertain
聚类精度 (F-Measure)	0.6289	0.6099	0.6123

注：对于 UIDK-Means，取其算法中参数 $k_{(\text{UIDK-Means})}=1,\lambda=0.5$(由文献[11]指出，在 $k_{(\text{UIDK-Means})}=1,\lambda=0.5$ 时，UIDK-Means 算法聚类精度较高)；UCNK-Means 中 $factor$ 取 0.3。

由上述实验结果可以发现，由于对数据不确定性化后，对象间的位置重叠部分较大，因此相较于原来的类标号，UIDK-Means，UCNK-Means 的聚类精度相对较之原数据(确定性数据)的 K-Means 有一定差距，但是本文的 UCNK-Means 比 UIDK-Means 聚类精度要高，且保持了相对较好的抗噪性；对于数据不确定化变化，相距原来确定数据的 K-Means 精度差距较小，在解决位置不确定性数据聚类问题上有着独到的优势。

4.2　自定义数据上聚类效果

本节实验采用自定义的数据，随机产生不确定性数据，然后对其进行 UCNK-Means 的聚类，结果用图形可视化方式和聚类密集性指标来进行衡量。

聚类密集性是一种有关聚类内方差的测量，方差越小，说明数据集的同一性越高。

给定一个数据集 X，簇内方差被定义为：

$$var(X)=\sqrt{\frac{1}{N}\sum_{i=1}^{N}d^2(x_i,\bar{x})} \tag{19}$$

其中，$d(x_i,\bar{x})$是矢量 x_i 与$\bar{x}$之间的距离；N 是数据集 X 中数据总个数；$\bar{x}$为 X 数据集中均值，可定义为：

$$\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i \tag{20}$$

聚类输出结果 $C_1,C_2,\cdots,C_n$，表示最终结果有 n 个簇；则聚类密集性可定义为：

$$cmp=\frac{1}{n}\sum_{i=1}^{N}(var(C_i)/var(X)) \tag{21}$$

其中，n 为最终聚类个数，$var(C_i)$为簇 C_i 的方差。每个簇内的成员尽可能的接近，所以聚类密集型越小越好。在极端情况下，每个单独的数据对象被分为单独的簇时，聚类密集型最小，为 0。

本节实验对位置不确定性对象的指标评估采用将位置不确定对象的每一个样本点看做一个单独的数据点参与到式(19)－式(21)的计算。

数据按照以下方式产生不确定数据：

在二维平面中，生成 20 个对象，对于每一个对象 O_p，首先在[0,800]范围上生成 2 个值，分别记作 x_p,y_p，然后在[50,200]范围内生成 2 个值，作为区域半径，记作 r_{x_p},r_{y_p}；这样 O_p 就可以表示为 $O_p=([x_p-r_{x_p},x_p+r_{x_p}],[y_p-r_{y_p},y_p+r_{y_p}])$，为了能够在二维图中表示，在对象 O_f 区域内随机生成 15 个样本点。

案例 1　对象分布如图 2 所示。

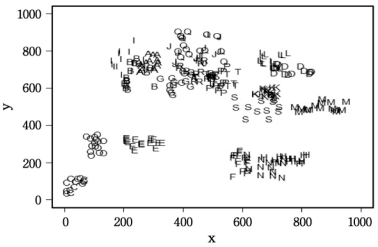

图 2　自定义数据分布图(一)

图 2 中，一种字母代表一个对象，一个字母代表该字母所表示的对象的一个样本点，因此可以看出图中总共有 A－T 20 个对象，每个对象有 15 个样本点。

对该数据集进行 UCNK-Means 聚类，选择聚类参数 $k=4$，即生成 4 个簇。

当初始簇中心选择$\{T,S,H,E\}$，则聚类结果生成的 4 个簇为$\{O,E,C\}$，$\{P,R,I,J,T,A,B,Q,G\}$，$\{N,H,F\}$，$\{D,S,M,L,K\}$。

当初始簇中心选择$\{A,T,S,M\}$时，结果还是$\{O,E,C\}$，$\{P,R,I,J,T,A,B,Q,G\}$，$\{N,H,F\}$，$\{D,S,M,L,K\}$。

在二维平面中的的划分如图 3 所示。

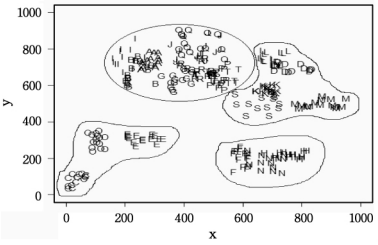

图 3　自定义数据分布聚类结果图(一)

根据式(19)－式(21)可得聚类密集性：$Cmp=0.3165$，在(0,1)间比较靠近 0。

案例 2　对象分布如图 4 所示。

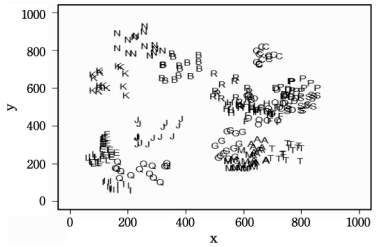

图 4　自定义数据分布图(二)

对该数据集进行 UCNK-Means 聚类，选择聚类参数 $k=4$，即生成 4 个簇。

当初始簇中心选择$\{T,S,H,E\}$，则聚类结果生成的 4 个簇为$\{S,O,D,R,P,H,F,C\}$，$\{M,A,T,G\}$，$\{E,I,J,Q,L\}$，$\{B,N,K\}$。

当初始簇中心选择$\{A,T,S,M\}$时，结果还是$\{S,O,D,R,P,H,F,C\}$，$\{M,A,T,G\}$，$\{E,I,J,Q,L\}$，$\{B,N,K\}$。

在二维平面中的划分如图5所示。

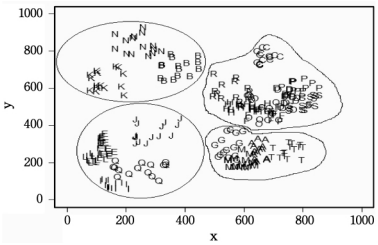

图5　自定义数据分布聚类结果图(二)

根据式(19)－式(21)可得聚类密集性：$Cmp=0.3597$，在(0,1)间比较靠近0。

因此，从以上两个案例可以看出，无论是从直观图形角度，还是从聚类密集性指标上来讲，UCNK-Means的聚类效果都是比较好的；无论初始中心是否选得合理，最后聚类效果都是非常符合直观感受的，与确定性数据的K-Means算法效果一致，表明了UCNK-Means算法能够很好地解决位置不确定性对象聚类的难题。

结束语　针对位置型不确性数据的聚类问题，目前的主流算法有以UK-Means为代表和以概率密度表示对象、以UIDK-Means为代表和以区间数表示对象这两类算法；但是UK-Means一类算法计算复杂度高，包含向量多次积分，并且概率密度函数现实问题中难以轻易得到；而UIDK-Means一类算法相对于UK-Means算法复杂度有所降低，取消了积分运算，但是它含有区间内数据的标准差计算，也需要一定复杂度；并且UIDK-Means只考虑区间整体大小，没有考虑到区间变化范围对聚类结果的影响。针对以上问题，本文提出了以联系数表示对象的UCNK-Means算法，该算法在UIDK-Means基础上又大幅度降低了计算复杂度，不但考虑了不确定性数据间的整体距离，而且也考虑了自身不确定性变化范围的大小，因此聚类精度更高，区别性更强。

在经过不确定性化的UCI数据集上的实验发现，UCNK-Means的F-Measure精度基本接近原先UCI数据的K-Means算法的精度，并且在同样的不确定性数据下，UCNK-Means精度要比UIDK-Means精度要高；在自定义的数据集上进行聚类实验，可以直观地发现，UCNK-Means能够准确地识别出应该归为一类的数据，聚类密集型指标较小，聚类效果较好，并且受到初始簇中心的影响较小，具有优良的聚类效果。

本文是基于K-Means算法的思想，因此无法识别离群点以及任意形状的数据类；但是本文证明了基于联系数的不确定性对象表示，可以很好地解决聚类问题，并且能够在减少计算复杂度以及提高聚类精度上有显著效果。进一步的研究，可以继续运用联系数表示不确定性数据来进行基于密度的聚类、数据流的聚类等。

参 考 文 献

[1] 周傲英，金澈清，王国仁，等. 不确定性数据管理技术研究综述[J]. 计算机学报，2009，32(1)：1-16

[2] 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(3)：541-546

[3] 王梁，周光焱，王黎维，等. 不确定关系数据属性级溯源表示与概率计算[J]. 软件学报，2014，25(4)：863-879

[4] Xu Lei，Hu Qing-hua，Hung E，et al. A heuristic approach to effective and efficient clustering on uncertain objects[J]. Knowledge-Based Systems，2014，66(9)：112-125

[5] Wang K N，Kao Ben，Chun K C D，et al. Efficient Clustering of Uncertain Data[C]// Proceedings of the Sixth International Conference on Data Mining(ICDM'06). IEEE，2006：436-445

[6] Hung E，Xiao Lu-rong，Hung R Y S. An Efficient Representation Model of Distance Distribution Between Uncertain Objects[C]// Computational Intelligence，2012，28(3)：373-397

[7] Xiao Lu-rong，Hom H，Hung E，et al. An Efficient Distance Calculation Method for Uncertain Objects[C]// Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Minging. 2007：10-17

[8] Yun C H，Yang J. Reducing UK-Means to K-Means[C]// Proceedings of the 6th IEEE International Conference on Data Mining(ICDM 2007). 2007：483-488

[9] Gullo F，Ponti G. A Tagarelli. Clustering Uncertain Data Via K-Medoids[C]// International Conference on Scalable Uncertainty Management. 2008：229-242

[10] Chau M，Cheng R，Kao B，et al. Uncertain data mining：An example in clustering location data[C]// The 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining. lnai，2006，3918：199-204

[11] 彭宇，罗清华，彭喜元. UIDK-means：多维不确定性测量数据聚类算法[J]. 仪器仪表学报，2011，32(6)：1201-1207

[12] 黄德才，张丽君，赵克勤. 基于a＋bi型联系数的不确定网格静态调度算法[J]. 计算机科学，2007，34(8)：126-129

[13] Su H S，Mi G S. Set pair analysis applied for identifyingpower transformer faults[C]// IntConf on MachineLearning and Cybernetics. Dalian，2006：1708-1713

[14] Dong L，Li G G，He Z X. Pattern recognition based onall set theory and SPA in complexsystem innovativecomputing[C]// 1st Int Conf. on Information and Control. Beijing，2006：204-208

[15] Huang D C，Zhao K Q. Uncertainty network planningmethodology based on the connection number a＋ b i＋ cj[C]// 5th World Congress on Intelligent Contrl and Automation. Hangzhou，2004：2863-2866

[16] 杨俊杰，周建中，方仍存，等. 基于集对分析的不确定多属性决策方法[J]. 控制与决策，2008，23(12)：1423-1426

[17] 刘健，刘思峰，吴顺祥. 基于优势关系的多属性决策对象排序研究[J]. 控制与决策，2012，27(4)：632-635，640

[18] 汪新凡，王坚强，杨恶恶. 基于二元联系数集结算子的多准则群决策方法[J]. 控制与决策，2013，28(11)：1630-1636

[19] 陶利民，黄德才. 开放网络环境下基于多元联系数的主观信任评估与决策研究[J]. 小型微型计算机系统，2012，33(6)：1202-1206

[20] 王文圣，金菊良，丁晶，等. 水资源系统评价新方法——集对评价法[J]. 中国科学E辑：技术科学，2009，39(9)：1529-1534

[21] Habich D，Volk P B，Diittmann R，et al. Error-a-ware density-based clustering of imprecise measurement values[C]// Proc. of the 23nd IEEE Int'l Conf. on Data Mining(ICDM 2007). 2007：417-476