一般分布区间型符号数据的 K 均值聚类方法①

郭均鹏，陈 颖，李汶华
（天津大学管理与经济学部，天津 300072）

摘要： 对于区间型符号数据聚类分析的研究，现有方法大多假设个体在区间内服从均匀分布，这往往并不符合实际情况. 针对此问题，研究一般分布的区间型符号数据 K 均值聚类方法，给出了一般分布区间型符号数据的定义，并基于经验分布理论研究其描述统计. 基于 Hausdorff 距离，考虑区间数所包含个体的分布信息，提出了一种新的区间型符号数据距离度量. 给出了一般分布的区间型符号数据 K 均值聚类算法. 通过随机模拟试验对该方法进行了有效性评价，结论表明，在各种实验设计的条件下，考虑一般分布的 K 均值聚类算法有效性均优于均匀分布假设下的 K 均值聚类算法. 最后将文中方法应用于汽车的聚类分析，进一步体现了文中方法在解决实际问题中的优势.

关键词： 区间数；一般分布；符号数据分析；聚类分析
中图分类号： O212.4 **文献标识码：** A **文章编号：** 1007－9807(2013)03－0021－08

0 引 言

现代社会中数据信息的丰富促进了对高效的数据分析方法的需求. 传统的数据分析技术在处理数据结构过于冗杂的数据集合时，有很大的局限性，主要困难在于：由于样本容量和变量维数的影响，往往使得计算工作量很大，并且难以把握数据属性的内在关系，无法获得隐含在数据中的重要知识资源[1]. 符号数据分析(symbolic data analysis，简称 SDA)是研究如何从海量数据中发掘系统知识的理论和方法[2]，其运用数据打包的思想，不仅使得计算量减少，并且能从整体上把握样本的特性. 例如，对股票进行评价，若决策者希望从全局上研究各股票板块的表现，而不关心个股的表现，这样就可对股票按板块打包，此处，打包后的样本个体称为符号对象. 相应地，样本数据的性质就发生了变化，由原来的“点数据”变为“符号数据”. 符号数据可能是定量数据，也可能是定性数据，可有多种表现形式，其中区间型符号数据是最常用的一种符号数据类型[2,3]. 例如研究某股票板块符号对象，由该板块的所有股票在某天的收盘价的最小值和最大值，构成区间型符号数据 $X=[25, 36]$. 聚类分析是知识发现最重要的技术手段之一，常用的聚类分析方法包括系统聚类、K 均值聚类等. 然而，当传统的“点数据”通过 SDA 的数据打包技术变为“符号数据”之后，传统的聚类分析方法无法奏效. 此时，对传统的聚类分析方法进行拓展研究，使其能够处理符号数据，十分必要.

符号数据的聚类分析是符号数据分析领域中的研究热点，近年来提出了许多针对各种类型的符号数据的聚类分析方法. 文献[3]提出了转换算法来对分布式符号变量进行聚类划分，文献[5]针对有约束的多值型符号数据提出了数据分解规则和与之相适应的相似性度量及聚类方法. 由于区间型符号数据在 SDA 中的重要地位，国内外学者对区间型符号数据的聚类分析的研究成果

① 收稿日期：2011－06－15；修订日期：2012－10－19.
基金项目：国家自然科学基金资助项目(71271147；71003072).
作者简介：郭均鹏(1973—)，男，山东昌邑人，博士，教授. Email：guojp@tju.edu.cn

更为丰富. 文献 [6] 介绍了基于欧式距离的划分聚类方法,以及各种对聚类结果进行解释的工具; 文献 [7] 提出了基于马氏距离(Mahalanobi distances) 的模糊聚类分析方法; 文献 [8] 介绍了基于二次距离的模糊 K 均值聚类方法; 文献 [9] 给出了基于城市—街区距离(city-block distances) 和 Hausdorff 距离的划分聚类方法,以及各种对聚类结果进行解释的工具; 文献 [10] 提出了基于 Wasserstein 距离的区间距离量度,归纳了该距离对其它几种区间距离的概括关系,并在此基础上对区间型符号数据的动态聚类分析方法进行了整理和统一; 文献 [11] 介绍了基于核方法的区间数模糊聚类算法,该方法通过设计合适的核函数,采用区间数遗传算法来求取高度非凸聚类优化问题得到聚类问题的全局最优解,并使用仿真实验说明了该方法的有效性.

上述对于区间型符号数据的聚类分析的研究,均假设个体在区间内服从均匀分布,然而在许多应用领域,个体在区间内往往服从如正态分布等非均匀分布. 本文在现有的针对均匀分布区间数据进行聚类分析的研究成果基础上,研究一般分布的区间型符号数据 K 均值聚类分析问题. 首先给出一般分布区间型符号数据的定义,并基于经验分布理论研究其描述统计,然后提出一般分布的区间型符号数据 K 均值聚类算法,并通过随机模拟试验对文中方法进行有效性评价,最后给出应用举例.

1 一般分布区间型符号数据及其描述统计量

首先给出一般分布区间型符号数据定义.

定义 1 [12] 若随机变量 X 服从某任意分布,且其观测值在 $[a,b]$ 取值,则称 X 为一般分布的区间型符号变量,简称一般分布区间变量或区间变量; $[a,b]$ 称为一般分布区间型符号数,简称一般分布区间数或区间数.

设 $X_1,\cdots,X_p$ 为 p 个一般分布的区间变量, $k=1,\cdots,n$ 表示 n 个符号对象, $[a_{kj},b_{kj}]$ 表示变量 X_j 在第 k 个符号对象的区间观测值,则符号对象的数据矩阵可表示为

$$\boldsymbol{X}=\begin{bmatrix} [a_{11},b_{11}] & \cdots & [a_{1p},b_{1p}] \\ \vdots & & \vdots \\ [a_{n1},b_{n1}] & \cdots & [a_{np},b_{np}] \end{bmatrix} \tag{1}$$

基于经验分布理论,可得一般分布区间变量 X_j 的经验均值为 [12]

$$\overline{X}_j=\frac{1}{n}\sum_{k=1}^{n}\mu_{kj} \tag{2}$$

式中 μ_{kj} 为一般分布区间变量 X_{kj} 的均值,由于 μ_{kj} 不易获得,实际运算中可用其样本均值 $\overline{X}_{kj}$ 来估计,即可以通过“打包”之前的原始点数据样本的样本均值进行求解. 式(2) 说明区间变量 X_j 的均值是各符号对象对应变量的均值的平均值.

一般分布区间变量 X_j 的经验方差为 [12]

$$S_j^2=\frac{1}{n}\sum_{k=1}^{n}[\sigma_{kj}^2+(\overline{X}_j-\mu_{kj})^2] \tag{3}$$

式中 μ_{kj} 为一般分布区间数 X_{kj} 的均值; σ_{kj}^2 为一般分布区间数 X_{kj} 的方差. 它们可分别由形成区间的原始点数据的样本均值 $\overline{X}_{kj}$ 和样本方差 S_{kj}^2 进行估计.

2 一般分布区间型符号数据的 K 均值聚类方法

距离度量是聚类分析的基础,为此,先研究一般分布区间型符号数据的距离度量.

2.1 一般分布区间型符号数据的距离度量

基于对传统 Hausdorff 距离的改进,本节将提出一般分布区间数的新距离. Hausdorff 距离用来定义 R^P 空间上两个紧集之间的距离,对于区间数据 $A=[a,b]$ 和 $B=[c,d]$,将区间数视为一个紧集,可得区间数间的 Hausdorff 距离 [9,13]

$$H(A,B)=|c(A)-c(B)|+|r(A)-r(B)| \tag{4}$$

式中 $c(X)$ 表示区间数 X 的中点; $r(X)$ 表示区间数 X 的半径. 此处 $X=A$ 或 B.

分析式(4) 所定义的 Hausdorff 距离可知,对于均匀分布的区间数,式中的中点 $c(X)$ 描述了区间数的集中位置,半径 $r(X)$ 描述了区间数的离散程度. 然而,对于非均匀分布的区间数来说,如此描述不再合适. 一般来讲,对于一般分布区间数,当区间数内部点数据可获得时,考虑用均值表示

集中位置,用标准差描述离散程度. 于是,可对 Hausdorff 距离加以改进,使之适用于服从一般分布的区间数.

定义 2　设 A、B 为任两个一般分布区间数,当区间数内点数据可得或其服从分布已知时,A、B 间的 $\mu\sigma$ 距离定义为

$$D_S(A,B) = |\bar{A} - \bar{B}| + \sqrt{3}\,|S_A - S_B| \quad (5)$$

式中 $\bar{A}$、$\bar{B}$ 分别表示区间 A、B 内点数据的样本均值或其服从分布的均值; S_A、S_B 分别表示区间 A、B 内点数据的标准差或其服从分布的标准差.

式(5) 中将标准差之差的系数设定为 $\sqrt{3}$,是因为当区间 A 和 B 内的点数据服从均匀分布(或只知道区间端点,无内部信息) 时,注意到

$$\bar{X} = \frac{\underline{x} + \bar{x}}{2}, \qquad S_X = \frac{\bar{x} - \underline{x}}{2\sqrt{3}} = \frac{r(X)}{\sqrt{3}} \quad (6)$$

则式(5) 退化为区间数的 Hausdorff 距离.

易得如下定理.

定理 1　式(5) 为距离度量.

下面给出 1 个算例,分别计算 $\mu\sigma$ 距离和均匀分布假设下的 Hausdorff 距离,以比较这两种距离对一般分布区间数进行距离度量的有效性.

例 1　给定 3 个区间数 A、B_1、B_2,其中 $A = [0,16]$ 为内部点数据服从左偏态分布的区间数,其内部数据为 $\{0,2,8,10,10,11,11,11,12,12,12,14,14,14,14,16,16,16,16,16\}$,$B_1 = [2,18]$ 和 $B_2 = [-2,14]$ 为两个服从均匀分布的区间数. 讨论 B_1 和 B_2 与 A 的距离. 区间数 B_1 的位置要偏向于区间 A 点数据密度较大的一侧(右侧),而区间数 B_2 的位置要偏向于区间 A 点数据密度较小的一侧(左侧),故 B_1 和 A 的距离应小于 B_2 和 A 的距离.

然而,由式(4) 计算区间数的 Hausdorff 距离,得 $D_H(A, B_1) = D_H(A, B_2) = 2$,而由式(5) 计算区间数的 $\mu\sigma$ 距离,得 $D_S(A, B_1) = 2.150\,66$, $D_S(A, B_2) = 6.150\,66$, $D_S(A, B_1) < D_S(A, B_2)$,表明使用基于描述统计量的 $\mu\sigma$ 距离计算距离,计算结果与实际情况相符.

由此算例可知,当区间内部点数据服从一般分布时,均匀分布假设下的 Hausdorff 距离忽略了内部点数据的分布信息,使计算结果偏离了实际情况,而 $\mu\sigma$ 距离考虑了区间内部点数据的描述统计量,使计算结果更加准确有效.

考虑区间向量的距离,假设

$$\boldsymbol{X} = (x_1, x_2, \cdots, x_p)^T = ([a_1,b_1], [a_2,b_2], \cdots, [a_p,b_p])^T$$

$$\boldsymbol{Y} = (y_1, y_2, \cdots, y_p)^T = ([c_1,d_1], [c_2,d_2], \cdots, [c_p,d_p])^T$$

为两个 p 维区间向量,将实数空间中的欧氏距离进行拓广可以得到两个区间向量之间的距离为

$$d_S(\boldsymbol{X},\boldsymbol{Y}) = \sqrt{\sum_{i=1}^{p} d(x_i,y_i)^2} = \sqrt{\sum_{i=1}^{p} [\,|\bar{x}_i - \bar{y}_i| + \sqrt{3}\,|S_{x_i} - S_{y_i}|\,]^2} \quad (7)$$

2.2　基于 $\mu\sigma$ 距离的 K 均值聚类算法

K 均值聚类是基于样本间相似性度量的间接聚类方法,其基本思想为给定或随机产生个先验划分,得到初始的聚类中心(聚类原型),计算每个对象到各聚类中心的距离,根据最小距离准则将对象并入与其距离最近的类; 更新聚类中心,重新计算每个对象与新的聚类中心的距离,重复这两个过程直至聚类中心不再改变(即聚类划分不再改变)[13].

根据传统实数据的 K 均值聚类分析方法,可以得到一般分布区间型符号数据的 K 均值聚类分析基本算法如下.

算法 1　基于 $\mu\sigma$ 距离的 K 均值聚类算法.

考虑将 n 个对象 $\{X_1,\cdots,X_n\}$ 聚为 K 类 $\{P_1,\cdots,P_K\}$,对应的聚类中心为 $\{y_1,\cdots,y_k\}$,K 均值聚类分析旨在获得最优划分使得目标函数 J 达到最小

$$J = \sum_{k=1}^{K} \sum_{i \in P_k} \phi(x_i, y_k)$$

式中 $\phi(x_i,y_k)$ 是对象 x_i 和聚类中心 y_k 之间的距离量度.

步骤 1　初始化　从 n 个对象中任选 K 个对象作为初始聚类中心 $\{y_1,\cdots,y_K\}$,并令参数 $test = 0$.

步骤 2　聚类划分　由式(7) 计算对象 $x_i(i = 1,\cdots,n)$ 和聚类中心 $y_k(k = 1,\cdots,K)$ 之间

的 $\mu\sigma$ 距离，若对象与 P_j 的聚类中心 y_j 的距离最小，则将 x_i 并入类 P_j 中.

步骤3 调整聚类划分 设任意对象 $x_i(i=1,\cdots,n)$ 原来在类 P_h 中，其与类 P_{h^*} 的聚类中心的 $\mu\sigma$ 距离最小，如果 $h \neq h^*$，则 $test=1$，将 x_i 并入类 P_{h^*} 中，即令 $P_{h^*}=P_{h^*}\cup\{i\}$ 且 $P_h=P_h\backslash\{i\}$.

步骤4 更新聚类中心 利用第 k 类中对象 $x_i=[a_i,b_i](x_i\in P_k)$ 的经验描述统计量更新 P_k 聚类中心 $y_k(k=1,\cdots,K)$，即根据式(2)和式(3)求得 P_k 的聚类中心 y_k 的经验均值 $\bar{y}_k$ 和经验方差 S_{y_k}.

步骤5 检验循环终止条件 若 $test=0$，迭代终止；否则令 $test=0$，转步骤2.

需说明的是，关于该算法中更新后的聚类中心，是基于类中所有区间型符号数据的经验描述统计量来定义的，在此基础上，计算待聚类对象 x_i 与该聚类中心的 $\mu\sigma$ 距离，进而调整聚类划分. 这是该算法与传统 K 均值聚类算法的区别，也恰是该算法的主要创新点.

3 聚类有效性评价的随机模拟

3.1 评价指标的选取

用 CR 指数(corrected rand index)[14] 来衡量同一个数据集的两个不同划分之间的差距. 本文将其作为聚类有效性的评价指标，其定义如下.

设共有 n 个样品，$U=\{u_1,\cdots,u_i,\cdots,u_R\}$，$V=\{v_1,\cdots,v_j,\cdots,v_C\}$ 是这同一组样品的两个不同的划分，分别包含 R 类和 C 类，则指数

$$CR=\frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2}-\binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i_o}}{2}\sum_{j=1}^{C}\binom{n_{oj}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{R}\binom{n_{io}}{2}+\sum_{j=1}^{C}\binom{n_{oj}}{2}\right]-\binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i_o}}{2}\sum_{j=1}^{C}\binom{n_{oj}}{2}} \tag{8}$$

式中 $\binom{n}{2}=\frac{n(n-1)}{2}$；$n_{ij}$ 表示属于聚类 u_i 和 v_j 中相同的样品数；n_{i_o} 表示聚类 u_i 中的样品数，n_{oj} 表示聚类 v_j 中的样品数.

CR 指数取值在 $[-1,1]$ 区间，越接近于 1 表示 U 和 V 两种划分越相似，等于 1 时两种划分完全相同；反之，若接近 0 或为负数，说明两种划分差异较大. 该指数是外部评价指数，它的计算结果不依赖于聚类分析中所选择的距离度量，较为公正客观，但它只能应用于标准聚类划分已知的情况下. 一般在随机模拟试验中，令 U 为原始标准划分，V 为聚类分析所获得的划分，从而用 CR 指数反映聚类分析得到的划分与已知的原始标准划分之间的差距，进而评价聚类算法的有效性.

3.2 随机模拟试验设计

本试验首先构造已划分类别的区间数据集，并在数据集中模拟生成区间内散点数据集，而后在得到的数据集上实施聚类分析，得到划分结果后使用 CR 指数分别与已知划分进行比对. 根据蒙特卡洛模拟思想，当随机模拟次数足够多时，可以认为试验获得的 CR 指数的平均值是近似于真实情况的.

如何产生随机区间数是本试验的关键所在，具体模拟生成方法如下.

1) 生成区间中点的随机点数据集

首先随机生成包含 350 个二维实数数据的点数据集作为区间数据集的中点，这 350 个实数点是由 3 类相互独立数据集组成的，其中有两类各包含 150 个点，另一类包含 50 个点，此即初始类别划分情况. 数据集中每一个点由两个服从标准正态分布的变量(x 和 y 轴坐标)确定，变量的均值向量和协方差矩阵记为

$$\boldsymbol{\mu}=\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix},\ \boldsymbol{\Sigma}=\begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix}$$

3 个聚类中的点根据如下参数分别随机产生：

data 1：$\mu_1=22,\mu_2=20,\sigma_1^2=100,\sigma_2^2=9$；

data 2：$\mu_1=64,\mu_2=30,\sigma_1^2=9,\sigma_2^2=36$；

data 3：$\mu_1=40,\mu_2=42,\sigma_1^2=9,\sigma_2^2=9$.

2) 产生随机区间数据集

依据已生成的随机点数据集，将每个点作为中心，在某一给定范围内随机生成参数 r_1 和 r_2(例如设定随机数取值范围为 $[1,4]$)，将其作为中心点向 x 轴和 y 轴两个方向上扩展的半径，则对任意点 (x,y)，有($[x-r_1,x+r_1]$，$[y-r_2,y+r_2]$)，从而使得每个实数点扩展成一个形为矩形的区间数，得到随机区间数据集，如图 1 所示.

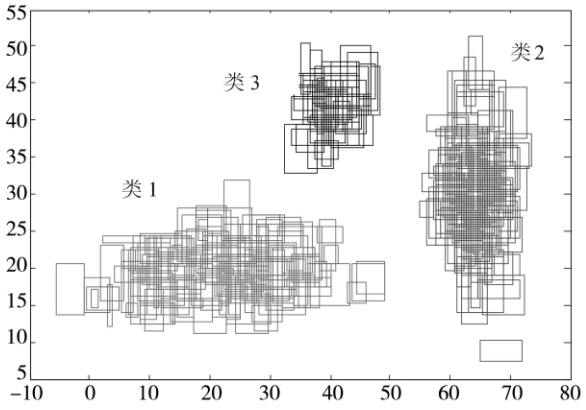

图 1　随机区间数据集的示意图

Fig. 1 The random interval data

3) 产生区间数内散点

为模拟类似一般分布的真实情况,在生成区间型符号数据集后,本试验在每个随机区间数内再分别产生服从正态分布的 50 个随机二维实数点数据.

正态曲线下横轴上一定区间的面积反映该区间的例数占总例数的百分率,横轴与正态曲线之间的面积恒等于 1,但正态曲线两端趋近于无穷,因此在理论上无法生成某一区间内的正态分布区间数. 而横轴区间$(\mu - 2.58\sigma, \mu + 2.58\sigma)$内的面积为 99.73%,即随机散点落在此范围外的概率仅为 0.27%,这是一个小概率事件,通常在一次试验中不易发生. 于是可以通过以下公式得到随机二维实数点数据: $N(x, r_1/2.58)$ 和 $N(y, r_2/2.58)$.

在本试验中,为了使模拟产生的区间数据集具有广泛的代表性,r_1 和 r_3 的生成范围分别选取 [1,4], [1,8], [1,12], [1,16], [1,20] 5 种情况. 根据蒙特卡洛模拟思想,对这 5 种情况的每一种再分别按照上述步骤,以随机生成的点数据集为中心,各随机产生 60 次区间数据集,即每次试验共基于 350 组区间数据进行,每个区间数内随机生成 50 个服从正态分布的二维点数据.

得到随机模拟区间数据集后,对其实施 K 均值聚类分析方法,并根据一般分布和均匀分布的不同,分别使用不同的距离度量和更新聚类中心的方式: 在一般分布中,使用式(5)所定义的距离度量公式,并根据文中所定义的经验均值和经验方差更新聚类中心,过程详见算法 1; 在均匀分布假设下,则采用普通的 Hausdorff 距离度量,并按传统的方式[13]更新聚类中心. 最后,利用 CR 指数分析比较通过聚类分析得到的划分与已知的先验划分之间的差异.

3.3　试验结果分析

对于随机模拟的每一种 r_1 和 r_2 的生成情况,分别求得 60 次实验结果平均值,最终得到的正态分布区间数据 K 均值聚类分析的平均 CR 指数值如表 1 所示.

表 1　正态分布区间数 K 均值聚类分析的平均 CR 指数值表

Table 1 The average CR values of K-means clustering of normally distributed interval symbolic data

r_1 和 r_2 生成范围	一般分布	均匀分布
[1,4]	0.925 772 67	0.783 692 47
[1,8]	0.924 783 18	0.775 969 05
[1,12]	0.919 936 31	0.775 379 41
[1,16]	0.920 510 24	0.779 131 74
[1,20]	0.920 249 46	0.773 328 38

对比两种分布情况下的聚类效果,可以明显看出一般分布前提条件下的 K 均值聚类有效性在所有情况下都比均匀分布假设下的 K 均值聚类有效性更好,特别是在扩展半径较小的情况下其有效性更为突出. 这样的分析结果说明一般分布的模拟更反映真实情况.

图 2 通过 CR 指数变化情况形象说明了在某次半径 r 范围下一般分布和均匀分布的区间型符号函数一次完整的 K 均值聚类过程中聚类结果的迭代变化. 从图中可以看出在一次迭代过程中两种分布下的聚类迭代结果都是从较小值开始,逐渐增加到较为稳定的值,即本次聚类结束的最终 CR 指数值. 说明本文所描述的区间符号数据 K 均值聚类方法具有较好的收敛性和稳定性.

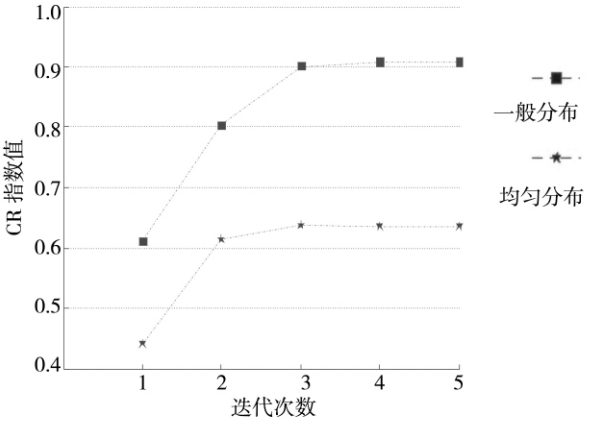

图 2　一般分布与均匀分布区间型符号数据 K 均值聚类迭代过程对比图

Fig. 2 The iterative process of K-means clustering of generally and uniformly distributed interval symbolic data

4 应用举例

为说明本文所提出的聚类方法在实际应用中的有效性，现以汽车样本为例进行聚类分析. 选取 22 个具有先验类别属性的不同型号车辆构成符号对象样本，其先验类别分别为跑车、豪华车、中型车 3 种，其中属于跑车的 8 个车型，豪华车 7 个车型，中型车 7 个车型. 选取厂商指导价格、最大功率、标称 0 – 100 加速、标称最高车速、长、宽、高等 7 个聚类变量. 由于同一车型可能有多种不同的配置，因而同一车型在同一聚类变量上可能有多种取值. 收集各种车型在各个变量上的不同配置下的点数据，并对其打包形成区间型符号数据，如表 2 所示.

由于有 3 个先验类别，故定义 $k = 3$，分别采用文中的 K 均值聚类方法和假设各区间变量服从均匀分布的传统 K 均值聚类算法，对表 2 的样本数据进行聚类分析，并将聚类结果与先验类别对比，形成聚类正确率表，如表 3 所示.

表 2 汽车样本的区间型符号数据表

Table 2 Interval symbolic data of the cars sample

车辆型号	厂商指导价格 / 万元	最大功率 /hp	标称 0 – 100 加速时间 /s	标称最高车速 /km·h^{-1}	长 /mm	宽 /mm	高 /mm	车辆类别
保时捷 911	[129,191.3]	[325,517]	[4.2,5.8]	[275,310]	[4 427,4 491]	[1 770,1 852]	[1 275,1 310]	跑车
奔驰 SL 级	[119.8,256.8]	[231,525]	[4.6,7.8]	[250,250]	[4 535,4 605]	[1 815,1 835]	[1 298,1 303]	跑车
宝马 Z4	[57.8,92.8]	[184,340]	[4.8,7.3]	[232,250]	[4 239,4 244]	[1 790,1 790]	[1 284,1 291]	跑车
宝马 3 系敞篷	[58,79.8]	[152,306]	[5.7,10.9]	[214,250]	[4 580,4 612]	[1 782,1 782]	[1 384,1 384]	跑车
奥迪 TT	[50.9,70.8]	[200,272]	[5.2,6.4]	[237,250]	[4 178,4 198]	[1 842,1 842]	[1 345,1 358]	跑车
宝马 3 系双门	[49.1,69.8]	[152,306]	[5.4,9.7]	[218,250]	[4 580,4 612]	[1 782,1 782]	[1 395,1 395]	跑车
保时捷 Cayman	[72.8,114.8]	[245,331]	[4.9,7]	[253,285]	[4 341,4 347]	[1 801,1 801]	[1 304,1 305]	跑车
保时捷 Boxster	[68.8,106]	[245,320]	[5,7]	[251,274]	[4 329,4 342]	[1 801,1 801]	[1 292,1 231]	跑车
大众朗逸	[11.28,16.28]	[105,131]	[9.8,12.8]	[174,200]	[4 608,4 608]	[1 743,1 743]	[1 465,1 465]	中型车
大众速腾	[13.28,18.58]	[101,161]	[8.8,16]	[180,220]	[4 544,4 544]	[1 760,1 760]	[1 461,1 464]	中型车
别克英朗 GT	[13.77,18.97]	[121,184]	[9,13.2]	[180,220]	[4 671,4 671]	[1 815,1 815]	[1 478,1 478]	中型车
斯柯达明锐	[12.34,18.05]	[105,161]	[7.9,14]	[176,213]	[4 569,4 572]	[1 769,1 769]	[1 462,1 462]	中型车
大众迈腾	[19.28,43.98]	[131,250]	[6.9,14.5]	[193,250]	[4 765,4 865]	[1 820,1 820]	[1 472,1 475]	中型车
雪铁龙 C5	[17.69,40.38]	[140,220]	[8.6,10.9]	[200,230]	[4 745,4 805]	[1 780,1 860]	[1 458,1 476]	中型车
雪佛兰科鲁兹	[10.89,15.99]	[117,184]	[8.7,13.9]	[180,225]	[4 598,4 598]	[1 797,1 797]	[1 477,1 477]	中型车
奥迪 A6L	[35.5,85.33]	[170,350]	[6.4,9.3]	[220,265]	[5 012,5 035]	[1 855,1 855]	[1 485,1 485]	豪华车
宝马 5 系	[41.26,79.76]	[156,320]	[6.2,11.8]	[210,250]	[4 981,5 039]	[1 846,1 860]	[1 471,1 477]	豪华车
奔驰 E 级长轴距	[46.5,71]	[184,245]	[8.5,9.1]	[230,245]	[5 012,5 012]	[1 855,1 855]	[1 464,1 466]	豪华车
大众辉腾	[73.45,241.2]	[241,450]	[6.1,9.7]	[239,250]	[5 175,5 175]	[1 903,1 903]	[1 450,1 450]	豪华车
英菲尼迪 G	[38.8,73.41]	[235,351]	[5.6,8.9]	[210,249]	[4 653,4 780]	[1 773,1 852]	[1 394,1 455]	豪华车
宝马 7 系	[89.8,329.8]	[259,544]	[4.6,7.8]	[245,250]	[5 179,5 212]	[1 902,1 902]	[1 484,1 478]	豪华车
奔驰 S 级	[93,259.8]	[231,517]	[4.6,8.3]	[244,250]	[5 206,5 230]	[1 871,1 871]	[1 473,1 485]	豪华车

表 3　聚类正确率对比表

Table 3 The comparison of precision of the two methods

聚类结果	基于 $\mu\sigma$ 距离的 K 均值聚类方法	服从均匀分布的传统 K 均值聚类方法
误断样本	宝马 3 系敞篷、宝马 3 系双门、英菲尼迪 G	宝马 3 系敞篷、宝马 3 系双门、宝马 5 系、奔驰 E 级长轴距、英菲尼迪 G
误判个数	3	5
正确个数	19	17
判断正确率	86.36%	77.27%

对比两种聚类方法对实际数据应用的效果，可以看出，相对于均匀分布假设下的传统 K 均值聚类方法，本文所提出的一般分布区间型符号数据的 K 均值聚类方法，更具有效性.

5　结束语

本文放弃以往文献中常用的均匀分布区间数的假定，在定义了一般分布区间型符号数据的 $\mu\sigma$ 距离基础上，充分考虑区间数据内散点数据的信息，提出了一般分布区间型符号数据的 K 均值聚类算法. 随机模拟试验和实际应用举例均表明，文中方法与以往只考虑区间端点信息的均匀分布区间数聚类方法相比，能更准确、客观地反映实际情况，更具有效性. 将来的研究，可在本文方法基础上，进一步研究一般分布区间型符号数据的其它多元分析方法.

参 考 文 献：

[1] 胡　艳，王惠文. 一种海量数据的分析技术——符号数据分析及应用 [J]. 北京航空航天大学学报(社会科学版)，2004，17(2)：40－44.
Hu Yan，Wang Huiwen. A new data mining method based on huge data and its application [J]. Journal of Beijing University of Aeronautics and Astronautics(Social Sciences Edition)，2004，17(2)：40－44.(in Chinese)

[2] Bock H H，Diday E. Analysis of Symbolic Data [M]. New York：Springer-Verlag，2000.

[3] 李汶华，郭均鹏. 区间型符号数据回归分析及其应用 [J]. 管理科学学报，2010，13(4)：38－43.
Li Wenhua，Guo Junpeng. Methodology and application of regression analysis of interval-type symbolic data [J]. Journal of Management Sciences in China，2010，13(4)：38－43.(in Chinese)

[4] Diday E，Brito M P. Symbolic cluster analysis [C] // Conceptual and Numerical Analysis of Data(Eds. Opitz O)，Heidelberg：Springer-Verlag，1989：45－84.

[5] De Carvalho F A T，Csernel M，Lechevallier Y. Clustering constrained symbolic data [J]. Pattern Recognition Letters，2009，30(11)：1037－1045.

[6] De Carvalho F A T，Brito P，Bock H H. Dynamic clustering for interval data based on L2 distance [J]. Computational Statistics，2006，21 (2)：231－250.

[7] Tenorio C P，De Carvalho F A T，Pimentel J T. A partitioning fuzzy clustering algorithm for symbolic interval data based on adaptive mahalanobis distances [C] // Proceedings of 7th International Conference on Hybrid Intelligent Systems，2007：174－179.

[8] De Carvalho F A T，Tenorio C P. Fuzzy *K*-means clustering algorithms for interval-valued data based on adaptive quadratic distances [J]. Fuzzy Sets and Systems，2010，161(23)：2978－2999.

[9] De Carvalho F A T，Lechevallier Y. Partitional clustering algorithms for symbolic interval data based on single adaptive distances [J]. Pattern Recognition，2009，42(7)：1223－1236.

[10] Irpino A，Verde R. Dynamic clustering of interval data using a Wasserstein-based distance [J]. Pattern Recognition，2008，29(11)：1648－1658.

[11] 任世锦，吕俊怀. 基于遗传算法的区间数核模糊聚类方法[J]. 系统工程学报，2009，24(2)：226－230.
Ren Shijin，Lü Junhuai. Genetic algorithm-based kernel function FCM clustering algorithm for interval numbers[J]. Journal of System Engineering，2009，24(2)：226－230.(in Chinese)

[12] 郭均鹏，李汶华，高 峰. 一般分布区间型符号数据的描述统计与分析[J]. 系统工程理论与实践，2011，31(12)：2367－2372.
Guo Junpeng，Li Wenhua，Gao Feng. Descriptive statistics and analysis of interval symbolic data with general distribution[J]. Systems Engineering－Theory & Practice，2011，31(12)：2367－2372.(in Chinese)

[13] 樊 博. 基于空间聚类挖掘的城市应急救援机构选址研究[J]. 管理科学学报，2008，11(3)：16－28.
Fan Bo. Spatial clustering mining method for site selection problem of emergency response center[J]. Journal of Management Sciences in China，2008，11(3)：16－28.(in Chinese)

[14] De Carvalho F A T，De Souza R M C R，Chavent M，et al. Adaptive Hausdorff distances and dynamic clustering of symbolic interval data[J]. Pattern Recognition，2006，27(3)：167－179.

K-means clustering of generally distributed interval symbolic data

GUO Jun-peng，CHEN Ying，LI Wen-hua
College of Management and Economics，Tianjin University，Tianjin 300072，China

Abstract: The existed clustering methods of interval data mostly supposed that the data are uniformly distributed across the interval. However，this is not always practical. Taking this into account，this paper aims to research the *k*-means clustering method of interval data with a general distribution. The definition of generally distributed interval data is proposed，and descriptive statistics was researched based on empirical distribution theory. On the basis of Hausdorff distance，the paper puts forward a new distance for interval data，which considers the point data contained in the intervals. Based on this，we present a algorithm of *k*-means clustering of generally distributed interval symbolic data. A simulation experiment is conducted to evaluate the validity of our method. The results show that，compared with analysis methods of uniform interval symbolic data，the analysis methods of generally distributed interval symbolic data are more effective under all the conditions designed in our experiment. Finally，the method is illustrated by an example of real-case data which shows the advantages of our method in the practical application.

Key words: interval symbolic data；general distribution；symbolic data analysis；clustering analysis