基于区间数的多维不确定性数据 UID-DBSCAN 聚类算法

魏方圆 黄德才

（浙江工业大学计算机科学与技术学院 杭州 310023）

摘 要 不确定性数据聚类方法的研究日益受到广泛关注，其中 UIDK-means 算法与 U-PAM 算法继承了基于划分算法无法识别任意形状簇和对噪声点敏感的缺陷。FDBSCAN 算法事先假定不确定性数据的概率分布函数或概率密度函数是已知的，然而这些信息在实际应用中往往难以获取。针对上述算法的不足，提出一种基于区间数的多维不确定性数据聚类 UID-DBSCAN 算法。该算法利用区间数结合数据的统计信息合理地表示不确定性数据，采用低计算复杂度的区间数距离函数衡量不确定性数据对象间的相似度，首次提出区间数的密度、密度可达与密度相连等概念，并将其用于扩展簇中，同时结合数据集的统计特征自适应地选取算法的密度参数来实现自动聚类。实验结果表明，UID-DBSCAN 算法能够有效识别噪声，处理任意形状簇，具有较高的聚类精度和较低的计算复杂度。

关键词 不确定性数据，区间数，聚类算法，DBSCAN

中图法分类号 TP311.13 **文献标识码** A

UID-DBSCAN Clustering Algorithm of Multi-dimensional Uncertain Data Based on Interval Number

WEI Fang-yuan HUANG De-cai

(College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract The researches on clustering methods of uncertain data have been paid more and more attention, among them, the UIDK-means algorithm and U-PAM algorithm inherit the partition-based algorithm defects that can not identify any shape clusters and is sensitive to noise. FDBSCAN algorithm assumes that the probability distribution function or probability density function of uncertain data is known, however this information is hard to acquire. For the shortage of the above algorithms, a new multi-dimensional uncertain data clustering algorithm namely UID-DBSCAN based on interval numbers was proposed. It uses interval data combined with statistic information to describe uncertain data reasonably. And it utilizes the intervals distance function of low computing complexity to measure the similarity of different uncertain data. The concepts of interval density, interval density-reachable and interval density connected were firstly proposed and applied to expand clusters. Meanwhile in order to realize automatic clustering, combining with statistical features of the data, the parameters of density can be adaptively selected. Experiment results show that UID-DBSCAN algorithm can identify noise effectively, process arbitrary shape clusters and obtain better clustering precision with low computing complexity.

Keywords Uncertain data, Interval number, Clustering algorithm, DBSCAN

1 引言

近年来，随着无线通信技术和网络信息技术的快速发展，在信息的采集、传输和处理过程中由于测量误差、环境干扰、数据缺失或人为因素等产生了不确定性数据。不确定性数据具有不固定性、不可预知性、随机性、不规则性、易变性等特点[1]。根据不确定性数据的表现形式可将其分为元组存在级不确定性和元组属性级不确定性。元组存在级不确定性是指元组是否存在于数据库中具有不确定性，这种不确定性又分为元组之间相互依赖和元组之间相互独立两种情况。一般可采用点概率模型来表示一个元组存在的不确定性。在该模型中，元组的属性值是确定的，采用一个[0,1]之间的概率值表示其存在的可能性。元组属性级不确定性是指元组属性的取值具有不确定性。这种不确定通常采用概率密度函数或其他统计参数来描述特定属性的不确定性。

如果不能准确描述或充分考虑数据的不确定性，那么得到的数据挖掘结果往往是不可信甚至是错误的[2]。因此数据挖掘方法只有充分考虑数据的不确定性，才能对数据“去伪存真”，得到基于不确定数据的准确反映客观事物的潜在有用的知识。不确定性数据的聚类分析作为数据挖掘中一个重要的分支，日益受到众多学者的关注与研究。

确定性数据的聚类分析是在没有任何先验知识的前提下对数据进行自动划分，以达到相同簇内数据具有最大相似度、不同簇间数据具有最大相异度的一种无监督学习。已有的聚类分析方法主要可分为基于划分的、基于层次的、基于密度的、基于网格的和基于模型的算法[3]。

目前，在不确定数据聚类算法的研究中，主要将传统聚类算法思想扩展到不确定数据中。文献[4-7]提出了以 UK-means，CK-means，UK-medoid 为代表的基于划分的不确定性数据聚类算法，普遍采用概率密度函数或概率分布函数来表

本文受水利部公益性行业科研专项(201401044)资助。

魏方圆(1992－)，女，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仓库与数据挖掘、决策方法等，E-mail：hdc@zjut.edu.cn(通信作者)。

示不确定性数据，并通过积分计算数据对象间的距离的期望作为数据对象间的相似性判断，从而进行聚类分析。在基于密度的不确定性数据聚类算法的研究中，Kriegel 等人提出的 FDBSCAN 算法[8]利用模糊距离函数计算数据间距离，仅简单地通过判断计算得到的可达概率是否大于 0.5 作为数据对象和核心对象密度是否可达的标准，因此聚类的准确度和效率都无法得到保证。在此基础上，许华杰等人提出了 PDBSCAN 算法[9]，该算法以两个不确定对象间距离的最大、最小值为限定范围，同时通过建立概率阈值索引和 R 树索引来提高效率，并允许用户通过设置概率阈值 p 在计算精度和运行效率之间进行权衡。但该算法没有对数据进行预处理，而是直接生成 R 树，造成大的 I/O 消耗，同时仍采用积分计算，并没有从根本上减小计算量。胡春安等人提出的 HPDBSCAN 算法[10]利用 Hilbert 编码技术将多维数据映射到一维数据空间，通过 Hilbert-R 树索引对不确定性数据进行聚类，有效降低了 PDBSCAN 算法的时间复杂度与空间复杂度。Wang 等人提出了 ENDBSCAN 算法[11]，该算法根据簇心的概率大小对聚类时的半径参数值进行调整，提高了算法的聚类质量。Atakan 等人提出的 M-FDBSCAN 算法[12]将二维模糊数据对象集划分为 C 个子数据对象集，对每一个子数据对象集应用 FDBSCAN 算法，之后合并子数据集得到最终的簇，该算法具有较高的运行效率。Bin Jiang 等人提出的基于概率分布相似性的 DB-KL 算法[13]根据数据的概率分布特征，利用 KL 散度衡量不确定对象间的相似度，但数据间的距离仍需使用积分运算，算法的计算量较大。然而在实际应用系统中，数据的概率密度函数或概率分布函数往往很难得到。针对不确定信息难以获取、计算复杂等问题，彭宇等人利用区间数和统计值表示不确定数据，提出 UIDK-means 算法[14]，该算法不需要预先获得数据的概率分布即可对数据进行聚类，同时具有较低的计算复杂度，但该算法采用随机采样的方法对簇中心点进行初始化，降低了算法的灵活性，而且聚类结果对簇中心点初始化和噪声敏感。何云斌等人提出的 U-PAM 算法[15]、UM-PAM 算法[15]采用区间数结合标准差来表示不确定数据，利用 CH 指标确定最佳聚类个数，在一定程度上提高了聚类效果；但该算法仍会继承基于划分算法无法识别噪声以及不能发现任意形状簇等缺点[18]。

可以看出目前在不确定性数据聚类研究中存在着不确定性信息难以获取、不确定性表示过于理想化、聚类准确度与算法效率较低等问题。但是从统计学角度出发，不确定性数据的范围是较容易获得的。同时通过统计计算可得到数据的均值和标准差等统计信息，无需较多的假设和先验知识[14]。因此，可以采用区间数结合数据相关统计信息的方法科学合理地表示不确定性数据。

鉴于上述分析，本文提出一种基于区间数的多维不确定性数据聚类算法——UID-DBSCAN。该算法利用区间数结合数据相关的统计信息合理地表示不确定性数据，采用低计算复杂度的区间数距离函数衡量不确定性数据对象间的相似度，首次提出区间数的密度、密度可达、密度相连等概念并将其用于扩展簇中，为避免人工干预，结合数据集的统计特征自适应选取算法的密度参数。实验结果表明，UID-DBSCAN 算法能够有效识别噪声，处理任意形状簇，与其他相关的不确定性数据密度聚类算法相比，UID-DBSCAN 算法的聚类精度平均提高了 15.51%，运行时间平均降低 15 倍。

2　相关定义

本节对所提算法涉及到的相关概念进行定义。

2.1　不确定性数据对象

定义 1（不确定性数据对象）　对象 $O_i=\{O_{i1},O_{i2},\cdots,O_{is}\}$ 是一个由 s 个 m 维样本点数据集构成的不确定性数据对象，其中 $O_{ip}=\{O_{ip1},O_{ip2},\cdots,O_{ipm}\}(p\in\{1,2,\cdots,s\})$。

定义 2（不确定数据集）　不确定数据集 $O=\{O_1,O_2,\cdots,O_n\}$ 是一个由 n 个不确定性数据对象构成的不确定数据集。

2.2　区间数

定义 3（区间数[16]）　给定 $A_L,A_R\in R^m$ 且 $A_R\geqslant A_L$，称集合 $A=[A_L,A_R]=\{\mu\mid A_L\leqslant\mu\leqslant A_R\}$ 为一个区间数，其中 A_L 为区间数的下界，A_R 为区间数的上界。当 $A_L=A_R$，即上下界相等时，区间数为一个精确数。

定义 4（区间数的中点和半径[14]）　给定区间数 $A=[A_L,A_R]$，令 $\alpha=(A_R-A_L)/2$，$m=(A_R+A_L)/2$，则有：

$$A_L=m_A-\alpha_A,A_R=m_A+\alpha_A$$

其中，称 m_A 为区间数的中点，α_A 为区间数的半径，因此区间数也可表示为 $[m_A-\alpha_A,m_A+\alpha_A]$。

区间数具有上界和下界，当数据模型为一维空间数据时，区间数为实数轴上的一条线段；当数据模型为二维空间数据时，区间数为二维平面中的一块矩形区域；当数据模型为三维空间数据时，区间数为一个立方体空间；当数据模型为多维空间数据时，区间数为一个超几何体。

定义 5（区间数的均值[16]）　对于给定的区间数 $A_1=[A_{1L},A_{1R}],A_2=[A_{2L},A_{2R}],\cdots,A_n=[A_{nL},A_{nR}]$，它们的均值为：

$$\bar{A}=\left[\frac{A_{1L}+A_{2L}+\cdots+A_{nL}}{n},\frac{A_{1R}+A_{2R}+\cdots+A_{nR}}{n}\right]$$

定义 6（区间数的距离[16]）　对于给定的区间数 $X=[X_L,X_R],Y=[Y_L,Y_R]$，它们之间的距离为：

$$d(X,Y)=\|X-Y\|=\sqrt{|X_L-Y_L|^2+|X_R-Y_R|^2}$$

根据上述定义可知，区间数的距离为一个确定实数。而两个区间数均表示数据的范围，因此区间数之间的距离也应该为一个范围。显然如果仅用一个确定实数来表示两个区间数之间的距离将不够细致，同时也容易丢失一些有效信息，从而影响聚类效果。因此，需要结合区间在空间位置上的关系提出更适合刻画两个区间数间距离的定义。

定义 7（多维区间数之间的距离[16]）　对于给定的区间数 $X=[m_x-\alpha_x,m_x+\alpha_x],Y=[m_y-\alpha_y,m_y+\alpha_y]$。其中 $m_x,m_y,\alpha_x,\alpha_y\in R^m$，在任意维度 $j(1\leqslant j\leqslant m)$ 上，两个区间数的位置在数轴上可表现为重合、相接、相交、包含、相离 5 种位置关系。在维度 j 上，当它们重合时，其距离的最小值和最大值均为 0；当它们相接时，其距离的最小值为 0，距离的最大值为 $|m_{xj}-m_{yj}|+\alpha_{xj}+\alpha_{yj}$；当它们相交时，其距离的最小值为 0，距离的最大值为 $|m_{xj}-m_{yj}|+\alpha_{xj}+\alpha_{yj}$；当它们包含时，其距离的最小值为 0，距离的最大值为 $|m_{xj}-m_{yj}|+\alpha_{xj}+\alpha_{yj}$；当它们相离时，其距离的最小值为 $|m_{xj}-m_{yj}|-\alpha_{xj}-\alpha_{yj}$，距离的最大值为 $|m_{xj}-m_{yj}|+\alpha_{xj}+\alpha_{yj}$。

由上述分析可得，在任意维度 j 上，两个区间数之间的距离可定义为：

$$D_{\min}=\sqrt{\sum_{j=1}^{d}D_{j\min}^2},D_{\max}=\sqrt{\sum_{j=1}^{d}D_{j\max}^2}$$

$$D=[D_{\min},D_{\max}]=\left[\sqrt{\sum_{j=1}^{d}D_{j\min}^2},\sqrt{\sum_{j=1}^{d}D_{j\max}^2}\right]$$

其中，$D_{j\max}=|m_{xj}-m_{yj}|+\alpha_{xj}+\alpha_{yj}$。

$$D_{j\min}=\begin{cases}|m_{xj}-m_{yj}|-\alpha_{xj}-\alpha_{yj}, & |m_{xj}-m_{yj}|-\alpha_{xj}-\alpha_{yj}\geqslant 0\\ 0, & |m_{xj}-m_{yj}|-\alpha_{xj}-\alpha_{yj}<0\end{cases}$$

由此得到的区间数之间的距离仍为一个区间数，更加合理全面地表示了距离。

2.3 密度相关概念

定义 8(区间距离分布矩阵) 区间数之间的距离分布矩阵可表示为：

$$Dist_{n\times n}=\{dist(O_i,O_j)\mid 1\leqslant O_i\leqslant n,1\leqslant O_j\leqslant n\}$$

其中，$n=|D|$表示不确定数据集中的对象个数。$Dist_{n\times n}$是 n 行 n 列的实对称矩阵，每个元素表示不确定数据集中第 i 个对象到第 j 个对象的距离。

定义 9(区间半径 Eps) 对 $Dist_{n\times n}$中每行的值从小到大进行排序，记 $Dist_{n\times i}$为排序后距不确定数据对象 O_n 最近的第 i 个距离值。由于 $Dist_{n\times i}$中所有数据对象的第 i 个最近距离值在数轴上服从泊松分布，运用统计学中的极大似然估计法对所有第 i 个最近距离值的泊松分布进行参数估计，得到区间半径为：

$$Eps=\frac{1}{n}\sum_{i=1}^{n}dist(O_i,O_k),1\leqslant k\leqslant n$$

定义 10(区间阈值 $Minpts$) 在 Eps 确定的情况下，统计不确定数据集中每个数据对象区间半径 Eps 邻域内点的邻居个数，之后对整个不确定数据集中每个数据对象半径 Eps 邻域内的邻居个数求数学期望并取整，得到：

$$Minpts=[\frac{1}{n}\sum_{i=1}^{n}p_i],1\leqslant p\leqslant n$$

其中，p_i 为数据对象 i 的 Eps 邻域内数据对象的个数。

定义 11(核心对象) 在给定密度(Eps，$Minpts$)下对 $\forall O_i\in O(1\leqslant i\leqslant n)$，如果$|Eps(O_i)|\geqslant Minpts|$，则称 O_i 是不确定数据集 O 关于密度(Eps，$Minpts$)的一个核心对象。

定义 12(直接区间密度可达) 对 $\forall O_i,O_j\in O$，如果 O_i 是一个核心对象，且 $O_j\in Eps(O_i)$，则称对象 O_j 从 O_i 出发关于(Eps，$Minpts$)直接区间密度可达。

定义 13(区间密度可达) 如果不确定数据集存在一个对象链 $O_1,O_2,\cdots,O_n$，$O_1=X$，$O_n=Y$ 且从 $O_i(1\leqslant i\leqslant n-1)$到 O_{i+1}直接区间密度可达，则称从 X 到 Y 是区间密度可达的。

定义 14(区间密度相连) 对 $\forall O_j,O_k\in O$，如果 $\exists O_i\in O$，使从 O_i 到 O_j，从 O_i 到 O_k 都是关于(Eps，$Minpts$)密度可达的，则称 O_j 和 O_k 关于(Eps，$Minpts$)是区间密度相连的。

3 UID-DBSCAN 聚类算法

本节主要针对不确定数据提出 UID-DBSCAN 聚类算法。本节首先介绍多维不确定数据的区间数表示方法；然后提出根据区间数距离衡量不确定数据之间相似度的计算方法；最后对 UID-DBSCAN 聚类算法进行描述与分析。

3.1 多维不确定数据的表示

对于 n 个 $m(m\geqslant 2)$维不确定性数据对象 $O_1,O_2,\cdots,O_n$，第 i 个不确定数据对象第 $j(1\leqslant j\leqslant m)$维度的误差为$\overline{\varphi_j(O_i)}$。那么与第 i 个不确定性数据对象相关的 m 维误差向量可表示为$\overline{\varphi(O_i)}$。则第 i 个不确定性数据对象在区间$[O_i-k\overline{\varphi(O_i)}$，$O_i+k\overline{\varphi(O_i)}]$上的概率为[19-20]：$p(k=1)=68.3\%$，$p(k=2)=95.4\%$，$p(k=3)=99.7\%$。一般可根据实际需要选择合适的 k 值来表示数据对象。

由此可得到，n 个 m 维不确定数据对象可表示为$\{[O_1-k\overline{\varphi(O_1)},O_1+k\overline{\varphi(O_1)}],[O_2-k\overline{\varphi(O_2)},O_2+k\overline{\varphi(O_2)}],\cdots,[O_n-k\overline{\varphi(O_n)},O_n+k\overline{\varphi(O_n)}]\}$，其中 $1\leqslant i\leqslant n$ 且$\{k\in R\mid 0\leqslant k\leqslant 3\}$。

3.2 多维不确定数据间的距离

为了能够有效合理地衡量不确定数据对象之间的相似度，引入相关参数 λ，将两个区间数之间距离的最小值与最大值结合起来，其中$\{\lambda\in R\mid 0\leqslant\lambda\leqslant 1\}$。那么对于两个不确定性数据对象 $O_i=[O_i-k\overline{\varphi(O_i)},O_i+k\overline{\varphi(O_i)}]$与 $O_j=[O_j-k\overline{\varphi(O_j)},O_j+k\overline{\varphi(O_j)}]$之间的距离为：

$$D(O_i,O_j)=\sqrt{\lambda D(O_i,O_j)_{\min}^2+(1-\lambda)D(O_i,O_j)_{\max}^2}$$

由此可得不确定性数据对象间的距离具有非负性和对称性。即对于 $\forall O_i,O_j\in O$，可得：

$$0\leqslant D(O_i,O_j)<+\infty,D(O_i,O_j)=D(O_j,O_i)$$

3.3 算法描述

在上述定义的基础上，提出基于密度的多维不确定性数据聚类算法。算法描述如算法 1 所示。

算法 1 UID-DBSCAN 算法

输入：n 个不确定数据对象集 $O=\{O_1,O_2,\cdots,O_n\}$

输出：k 个簇 $C=\{C_1,C_2,\cdots,C_k\}$

Step1 采用区间数结合统计值将不确定数据对象集 O 中的每一个不确定性数据对象 $O_i(1<i<n)$表示为 $O_i=[O_i-k\overline{\varphi(O_i)},O_i+k\overline{\varphi(O_i)}]$。

Step2 对不确定数据集 O 中的每一个数据对象采用区间数之间的距离公式计算每两个不确定对象间的距离 $D(O_i,O_j)(i=i+1,j=j+1,1\leqslant i,j\leqslant n)$，得到 n 行 n 列实对称的区间距离分布矩阵 $Dist_{n\times n}$。

Step3 对区间距离分布矩阵 $Dist_{n\times n}$中的每行元素升序排列，得到每个不确定数据对象 O_i 的 k 近邻距离，计算每个不确定数据对象第 4-近邻距离的期望值，得到区间半径 Eps。

Step4 由 Step3 得到区间半径 Eps 的情况下，计算每个对象 O_i 在 Eps 邻域内邻居个数的期望并取整得到区间阈值 Minpts。

Step5 对每个未处理的对象 O_i，若该对象 Eps 邻域内的邻居个数大于等于 Minpts，则记对象 O_i 为核心对象，并建立新簇 C_i；否则记对象 O_i 为噪声点。

Step6 将核心对象 O_i 区间半径 Eps 邻域内的所有邻居对象，即直接区间密度可达的对象加入簇 C_i 中。

Step7 若对象 O_i 的 Eps 邻域内所有邻居对象均处理完毕，则执行 Step8；否则返回 Step6。

Step8 将与核心对象 O_i 区间密度可达的核心对象 O_j 及其 Eps 邻域内所有区间密度相连的邻居对象加入簇 C_i 中。

Step9 若 i>n，结束；否则返回 Step8。

UID-DBSCAN 算法的特点：1)采用区间数与数据对象的相关统计信息来表示不确定性数据，解决了实际中不确定性数据概率密度函数或概率分布函数不易获取的情况下无法有效进行聚类的问题；2)提出直接区间密度可达与区间密度相连等概念并将其用于扩展簇中，将传统的 DBSCAN 方法扩展到不确定性数据聚类算法中，避免了基于划分的不确定性数据聚类算法对噪声敏感、无法发现任意形状簇的缺陷；3)通过结合数据集的统计特征自适应地选择算法中的相关参数，降低了密度算法对参数的敏感性，提高了算法的聚类质量与计算效率。

4 实验

为了对 UID-DBSCAN 算法的聚类准确度和效率进行评估，并与其他算法进行比较，本节首先分析相关参数 λ 和 k 对

UID-DBSCAN 算法的影响，为 λ 和 k 设置合适的参数值；然后将 UID-DBSCAN 算法与基于划分的采用区间数表示不确定数据的 UIDK-means 算法进行聚类性能比较；最后通过使用聚类内部评价 CH 指标与 DB 指标对 UID-DBSCAN 算法的聚类效果进行评估，并与其他相关基于密度的算法进行对比分析。

UID-DBSCAN 算法的仿真实验的配置为 Intel(R) Core (TM) i5-3210M CPU 2.5GHz，在 4GB 内存的 PC 机上进行，操作系统为 Windows 7 旗舰版，程序采用 Java 语言编写。

为了验证 UID-DBSCAN 算法的有效性，本节采用的测试数据集是形状数据集 Flame，UCI 中的真实数据集 Iris，Glass，Wine。这 4 种数据集的主要特性如表 1 所列。

表 1　实验数据集

数据集	数据项个数	属性个数	类别数	属性类型
Flame	240	2	2	数值型
Iris	150	4	3	数值型
Glass	214	10	6	数值型
Wine	178	13	3	数值型

4.1　参数 λ 的选择

在不确定性数据聚类算法 UID-DBSCAN 中，相关参数需要人工设置，其中 $0\leqslant\lambda\leqslant1$。为了评估参数 λ 对聚类结果的影响，在实验中采用不同的 λ 值，使用 UID-DBSCAN 算法分别对具有不确定性的 Flame，Iris，Glass 和 Wine 4 种数据集进行聚类，其中参数 k 的取值为 1。聚类精度随相关参数 λ 的变化曲线如图 1 所示。

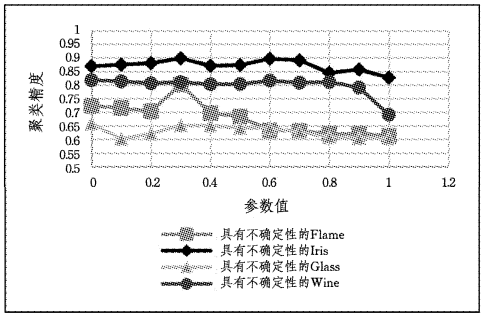


图 1　聚类精度与参数 λ 的关系曲线

从图 1 可以看出，对不同的数据集聚类时，参数 λ 的取值对聚类精度的影响不同。其中，不确定性 Iris 和 Wine 数据集的影响较小，而不确定性 Flame 与 Glass 数据集的聚类精度随参数 λ 的变化较为明显。参数 λ 的取值从 0 到 1 变化，当 λ 值在[0，0.5]之间时，UID-DBSCAN 算法的聚类性能相对比较稳定；当 λ 值大于 0.8 时，UID-DBSCAN 算法的聚类精度相对较低。因此，相关参数 λ 对聚类效果的影响与数据集自身具备的性质有关。在实际应用中，应该通过实验分析对相关参数进行设置。针对上述测试数据集，相关系数 λ 可在[0，0.5]之间取值。

4.2　参数 k 的选择

参数 k 的取值决定了采用区间数表示不确定性数据的范围。当 k 的取值大于 3 时，数据为异常值，因此 k 的取值范围为[0，3]。本文采用不同的参数值，使用 UID-DBSCAN 算法分别对具有不确定性的 Flame，Iris，Glass 和 Wine 4 种数据集进行聚类，其中参数 λ 取值为 0.3。实验得到的聚类精度与参数 k 之间的关系曲线如图 2 所示。从图 2 可以看出，参数 k 对不同数据有不同的影响。除对具有不确定性 Wine 数据集几乎没有影响以外，对其他 3 种具有不确定性的数据集均产生了一定的影响。其中对于具有不确定性的 Iris 和 Flame 数据集而言，当 k 值从 0.5 增加到 1.5 时，聚类精度较高，之后随着 k 值的不断增大，聚类精度逐渐下降；参数 k 对具有不确定性的 Glass 数据集的影响较大，当 k 的取值在[0.5，1]之间时，聚类精度呈上升趋势，之后随着 k 值的不断增加，聚类精度呈不断下降的趋势。通过上述分析可知，针对不同的数据集需要选择不同的参数 k，在聚类算法 UID-DBSCAN 中其值的选取范围为[0.5，1.5]。

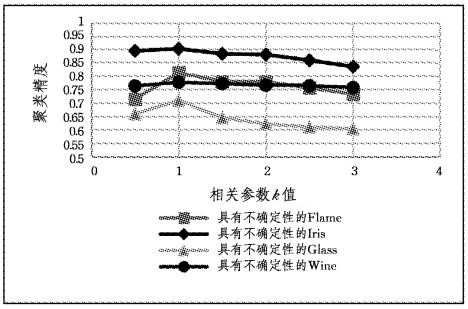


图 2　聚类精度与参数 k 的关系曲线

4.3　UIDK-means 与 UID-DBSCAN 的比较

为了验证本文提出的 UID-DBSCAN 算法的有效性，将其与 UIDK-means 算法进行比较。首先对测试数据集中的数据添加不确定性，对数据集 Flame，Iris，Glass，Wine 的每一维度添加均值为 0、方差为 1 的服从 binormial 分布的噪声。然后使用 UIDK-means 算法和本文提出的 UID-DBSCAN 算法对以上不确定性数据集进行聚类。实验结果如图 3 所示。

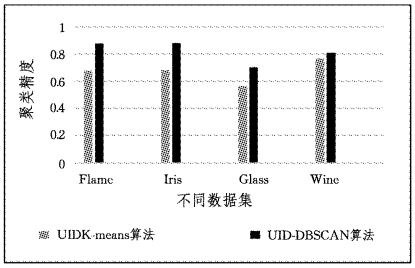


图 3　UIDK-means 算法与 UID-DBSCAN 算法的比较

UIDK-means 算法与 UID-DBSCAN 算法的共同点是都采用区间数的形式表示不确定性数据对象，不同点是 UIDK-means 算法是基于传统的 k-means 算法，其继承了基于划分的聚类算法无法识别噪声，无法发现任意形状簇和对初始簇中心选取敏感的缺点，因此对聚类效果产生了一定的影响，而本文提出的 UID-DBSCAN 算法解决了 UIDK-means 算法由于基于划分所存在的问题，提高了聚类准确度。

4.4　与其他基于密度的不确定性聚类算法的比较

为了评估 UID-DBSCAN 算法的聚类性能，将其与其他基于密度的不确定性聚类算法进行比较。首先为了表示数据的不确定性，分别在测试数据集的每一维度添加 3 种不同分布的噪声，分别为 normal，uniform，binormial 分布；然后使用 FDBSCAN 算法、PDBSCAN 算法与本文提出的 UID-DBSCAN 算法对以上不确定性数据集进行聚类。从聚类准确度、聚类质量与运行时间对算法聚类效果进行评估，相关实验结果如表 2、表 3 以及图 4 所示。

4.4.1　聚类准确度分析

表 2 中的 avg. score 是每一个算法对所有数据集聚类精度的平均值，avg. gain(平均增益)是 UID-DBSCAN 算法相对于其他算法聚类精度值增益的均值。

表 2 FDBSCAN 算法、PDBSCAN 算法与 UID-DBSCAN 算法的比较

数据集	噪声分布	聚类精度		
		FDBSCAN 算法	PDBSCAN 算法	UID-DBSCAN 算法
Flame	normal	0.6362	0.6721	0.8272
	uniform	0.5806	0.6829	0.7323
	binomial	0.6573	0.7021	0.8342
Iris	normal	0.7112	0.8273	0.8939
	uniform	0.8064	0.8071	0.8262
	binomial	0.6169	0.7659	0.8384
Glass	normal	0.5869	0.6671	0.7918
	uniform	0.6051	0.6356	0.8142
	binomial	0.5569	0.7092	0.7605
Wine	normal	0.6025	0.7053	0.7558
	uniform	0.7612	0.7928	0.8168
	binomial	0.7059	0.7087	0.7974
	avg. score	0.6523	0.7230	0.8074
	avg. gain	15.51%	8.44%	—

从表 2 可知，ID-DBSCAN 算法对具有不同分布的不确定性 Flame，Iris，Wine 和 Glass 数据集聚类，聚类的精度值比其他两种基于密度的不确定性数据聚类算法高。通过 avg. gain 可以看出，UID-DBSCAN 算法的精度值比 FDBSCAN 算法高出 15.51%，比 PDBSCAN 算法高出 8.44%，这主要是由于 UID-DBSCAN 算法自适应地选择参数，同时根据实验分析选取较优的相关参数值，从而使聚类准确度得到保证。

4.4.2 聚类质量评估

为了比较不同聚类算法产生的聚类效果，采用相关评估指标对聚类的质量进行评价。目前，对聚类的质量评价已有许多方法可供选择，一般可分为外部质量评价和内部质量评价两类方法[17]。外部质量评价一般假设数据集已经存在某种理想的聚类，并将其作为基准与某种算法的聚类结果进行比较；而内部质量评价是在没有已知外在基准的情况下，利用数据集自身的特性和聚类的量值来进行聚类质量的评价。

本文通过在具有不同分布的不确定性数据集上使用 FDBSCAN 算法、PDBSCAN 算法与 UID-DBSCAN 算法进行聚类，分别采用内部质量评价 CH 指标、DB 指标评估不同算法的聚类效果。CH 指标由聚类的簇间分离度与簇内紧密度的比值得到，CH 值越大意味着聚类中的每个簇自身越紧密，簇与簇之间越分散，即聚类结果更好。DB 指标[17]描述了聚类的簇内分散度与各簇中心的间距。DB 指标值的取值范围通常在[0,1]之间，DB 指标值越小表示簇间相异度越大，聚类效果越好。实验结果如表 3 所列。

表 3 FDBSCAN 算法、PDBSCAN 算法与 UID-DBSCAN 算法指标评估

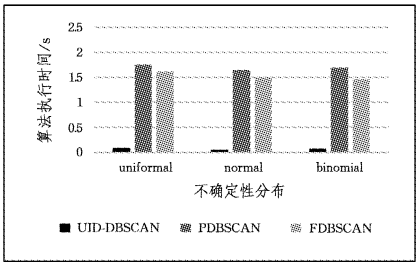
评估指标	噪声分布	FDBSCAN 算法	PDBSCAN 算法	UID-DBSCAN 算法
CH	normal	116.04	118.17	352.85
	uniform	112.81	118.39	225.82
	binomial	44.43	85.9	189.92
	avg. value	91.09	107.49	256.20
	avg. rise	2.80	1.38	—
DB	normal	0.59	0.54	0.42
	uniform	0.60	0.56	0.38
	binomial	0.65	0.45	0.39
	avg. value	0.61	0.52	0.40
	avg. reduce	22%	12%	—

表 3 中的 avg. value 是每一个算法对所有数据集聚类评估指标的平均值，avg. rise 是 UID-DBSCAN 算法相对于其他算法聚类评估指标增益的均值，avg. reduce 是 UID-DBSCAN 算法相对于其他算法聚类评估指标降低的均值。通过实验分析可得，UID-DBSCAN 算法的 CH 指标值高于 FDBSCAN 算法和 PDBSCAN 算法，同时 UID-DBSCAN 算法的 DB 指标值低于 FDBSCAN 算法和 PDBSCAN 算法，因此 UID-DBSCAN 算法与 FDBSCAN 算法和 PDBSCAN 算法相比具有较好的聚类结果。

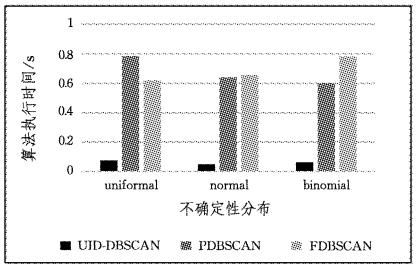
4.4.3 聚类处理时间分析

由于目前一些经典算法及本文提出的 UID-DBSCAN 算法均为基于 DBSCAN 的算法，因此它们的时间复杂度从理论上看是同阶的，为 $O(n^2)$，其中 n 为不确定数据对象的总个数。但 UID-DBSCAN 算法采用区间数结合数据集的统计信息对不确定数据进行聚类，利用区间数的一些特性，避免了大量复杂的积分运算，从而有效降低了算法的计算复杂度。同时由于各聚类算法中对相关参数的优化分析过程均是为提高聚类算法的准确度，因此在比较分析聚类效率时不考虑因参数优化导致的时间消耗。下面对算法的计算复杂度进行分析与比较。

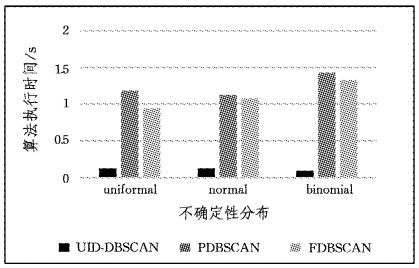
图 4 给出不同算法聚类处理时间的比较。



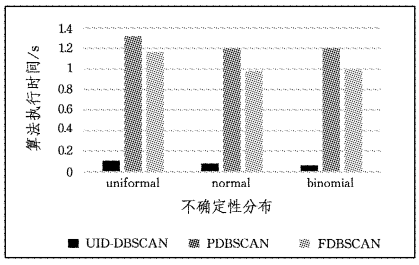
(a)具有不确定性 Flame 数据集上不同算法的处理时间



(b)具有不确定性 Iris 数据集上不同算法的处理时间



(c)具有不确定性 Glass 数据集上不同算法的处理时间



(d)具有不确定性 Wine 数据集上不同算法的处理时间

图 4 不同算法聚类处理时间的比较

由图 4 可看出，对具有不同分布的不确定性数据集进行聚类时，与其他相关算法相比，本文提出的 UID-DBSCAN 算法具有较小的聚类处理时间。在不确定性 Flame 数据集上，UID-DBSCAN 算法的处理时间平均是 PDBSCAN 算法的 1/26，是 FDBSCAN 算法的 1/24；在不确定性 Iris 数据集上，UID-DBSCAN 算法的处理时间平均是 PDBSCAN 算法的 1/11，是 FDBSCAN 算法的 1/12；在不确定性 Wine 数据集上，UID-DBSCAN 算法的处理时间平均是 PDBSCAN 算法的 1/15，是 FDBSCAN 算法的 1/13；在不确定性 Glass 数据集上，UID-DBSCAN 算法的处理时间平均是 PDBSCAN 算法的 1/11，是 FDBSCAN 算法的 1/10。由此可得，UID-DBSCAN 算法的聚类时间平均降低 15 倍，提高了聚类效率。这主要是由于 UID-DBSCAN 算法采用基于区间数的方法计算距离，避免了诸如 PDBSCAN 等算法需要较大计算量的积分运算，从而缩短了聚类处理时间。通过以上分析可知，距离计算方法的计算量往往对不确定性数据聚类算法的效率影响较大。因此，恰当地选择低计算量的距离计算方法对提高不确定性数据聚类算法的效率尤为重要。

结束语　针对概率密度函数或概率分布函数等信息在许多应用场合中往往很难获得，并且采用此信息来表示不确定性会导致聚类算法计算复杂度较高的问题，本文利用区间数结合统计信息来表示数据属性级的不确定性，并通过采用低计算量的区间数之间的距离计算方法来衡量不确定性数据间的相似度，提出了基于密度的多维不确定性数据聚类算法 UID-DBSCAN。同时根据数据自身的特性动态自适应地选取密度参数，避免了人工干预，使聚类过程达到自动化。通过仿真实验分析了相关参数对聚类结果的影响，并根据分析得到其较适合的取值。之后与基于划分的不确定性数据聚类算法 UIDK-means 在聚类精度方面进行比较，在相同的不确定数据集上，UID-DBSCAN 算法的聚类精度高于 UIDK-means 算法；同时 UID-DBSCAN 算法继承了传统的基于密度的聚类算法能够识别噪声并能发现任意形状簇的优点。最后将其与其他相关基于密度的不确定性数据聚类算法在具有不同分布的不确定性数据集上进行聚类，从聚类结果的准确度与聚类的处理时间等方面进行了比较分析。实验结果表明，相对于其他相关的基于密度的不确定性数据聚类算法，UID-DBSCAN 算法的聚类精度值平均提高了 15.51%，运行时间平均降低了 15 倍。

由于基于密度的聚类方法对参数敏感，当数据量增加时对算法效率影响很大，同时该方法主要面向静态的不确定性数据，无法有效地解决高效动态的不确定性数据流聚类的问题，因此今后的研究重点是针对基于层次的不确定性数据、动态的不确定性数据流进行有效聚类。

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] 任世锦. 基于区间数的不确定性数据挖掘及其应用研究[D]. 杭州：浙江大学，2006.

[3] 孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. 软件学报，2008，19(1)：48-61.

[4] 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(PAKDD 2006). Singapore：Springer-Verlag Berlin Heidelberg，2006：199-204.

[5] NGAI W K，KAO B，CHUI C K，et al. Efficient Clustering of Uncertain Data[C]// Proceedings of the 22nd IEEE International Conference on Data Mining(ICDM 2006). Hong Kong：IEEE Computer Society，2006：436-445.

[6] YUN C，YANG J. Reducing UK-means to K-means[C]// Proceedings of the 6th IEEE International Conference on Data Mining(ICDM 2007). Washington：IEEE Computer Society，2007：483-488.

[7] GULLO F，POINT G，TAGAERLLI A. Clustering Uncertain Data Via K-medoids[C]// Proceedings of the 2nd International Confe-rence on Scalable Uncertainty Management. Naples：Springer-Verlag Berlin Heidelberg，2008：229-242.

[8] KRIEGEL H P，PFEIFLE M. Density-based clustering of uncertain data[C]// The 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. Chicago：Illinois，2005：672-677.

[9] 许华杰，李国徽，杨宾，等. 基于密度的不确定性数据概率聚类[J]. 计算机科学，2009，36(5)：68-71.

[10] 胡春安，范丽文，毛伊敏. HPDBSCAN：高效的不确定数据处理算法[J]. 计算机工程与设计，2013，34(3)：1044-1049.

[11] WANG H M，WANG Y Y，WAN S T. A Density-based Clustering Algorithm For Uncertain Data[C]// Proceedings of International Conference on Computer Science and Electronics Engineering(ICCSEE 2012). Hangzhou：IEEE Computer Society，2012：102-105.

[12] ERDEM A，GÜNDEM T İ. M-FDBSCAN：A multicore density-based uncertain data clustering algorithm[J]. Turkish Journal of Electrical Engineering & Computer Sciences，2014，22(1)：143-154.

[13] JIANG B，PEI J，TAO Y F，et al. Clustering Uncertain Data Based on Probability Distribution Similarity[J]. IEEE Transactions on Knowledge and Data Engineering，2013，25(4)：751-763.

[14] 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.

[15] 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.

[16] 刘秀梅，赵克勤. 区间数决策集对分析[M]. 北京：科学出版社，2014：1-28.

[17] 黄德才. 数据仓库与数据挖掘教程[M]. 北京：清华大学出版社，2016.

[18] 戴阳阳，李朝锋，徐华. 初始点优化与参数自适应的密度聚类算法[J]. 计算机工程，2016，42(1)：203-209.

[19] AGGARWAL C C，YU P S. A Survey of Uncertain Data Algorithms and Applications[J]. IEEE Transactions on Knowledge and Data Engineering，2009，21(5)：609-623.

[20] DAVIES D L，BOULDIN D W. A Cluster Separation Measure[J]. Transactions on Pattern Analysis and Machine Intelligence，1979(4)：224-227.