doi:10.11959/j.issn.1000-436x.2017061

基于快速高斯变换的不确定数据聚类算法

迟荣华[1]，程媛[2,3]，朱素霞[2]，黄少滨[1]，陈德运[2,3]

（1. 哈尔滨工程大学计算机科学与技术学院，黑龙江 哈尔滨 150001；
2. 哈尔滨理工大学计算机科学与技术学院，黑龙江 哈尔滨 150080；
3. 哈尔滨理工大学计算机科学与技术学院博士后流动站，黑龙江 哈尔滨 150080）

摘 要：数据中不确定性的存在使对其聚类分析时要充分考虑不确定性的影响。针对现有不确定数据聚类算法中构建不确定数据模型以及距离度量时存在的影响结果准确性与聚类性能等问题，提出一种基于快速高斯变换的不确定数据聚类算法。首先在不假设数据分布的前提下，构建符合不确定性分布特征的数据模型；然后结合不确定对象的 2 个重要特征：属性特征与表示不确定数据分布特征的概率密度函数，度量不确定数据对象间的相似性；并以此为基础提出不确定数据聚类算法；最后在 UCI 以及真实数据集上的实验结果表明，所提算法在运行效率和聚类准确性方面均能取得较好效果。

关键词：聚类分析；不确定数据；概率密度函数；快速高斯变换；核密度估计

中图分类号：TP399 **文献标识码**：A

Uncertain data analysis algorithm based on fast Gaussian transform

CHI Rong-hua[1], CHENG Yuan[2,3], ZHU Su-xia[2], HUANG Shao-bin[1], CHEN De-yun[2,3]

(1. College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China;
2. College of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China;
3. Postdoctoral Research Station, College of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China)

Abstract: The effect of the uncertainties needs to be taken full advantage during uncertain data clustering. An uncertain data clustering algorithm based on fast Gaussian transform was proposed, to solve the problems about the impact on the accuracy of clustering results and the clustering efficiency caused by the uncertainties, during the construction of uncertain data models and the distance measurement, which existed in the current researches. First, the data model according to the characteristic of the uncertainty distribution was constructed, without the premise of assuming the data distribution. And the similarity between uncertain data objects was measured by combining the two important features of uncertain objects, attribute features and the probability density function representing the characteristic of uncertainty distribution. And then the uncertain data clustering algorithm was proposed. Finally, the experiment results on UCI and real datasets indicate the better efficiency and accuracy of proposed algorithm.

Key words: clustering analysis, uncertain data, probability density function, fast Gaussian transform, kernel density estimation

1 引言

由于原始数据不准确、采用粗粒度数据集合、出于隐私保护的特殊目的，以及数据集成等原因，经济、军事、电信等众多领域中普遍包含不确定数据[1]。虽然数据预处理技术可以提高数据质量，如对数据进行插值处理或删除含缺失值的记录等[2]，但是这些方法极可能改变数据的自身特性，如改变

收稿日期：2016-07-28；**修回日期**：2016-09-15
通信作者：程媛，changuang7@sina.com
基金项目：国家自然科学青年基金资助项目（No.61502123）；黑龙江省青年科学基金资助项目（No.QC2015084）；黑龙江省普通高等学校青年学术骨干支持计划基金资助项目（No.1253G017）

Foundation Items: The National Natural Science Foundation of China for Youths (No.61502123), Heilongjiang Province Science Foundation for Youths (No.QC2015084), Heilongjiang Province Support Program for Youth Academic Backbones in Regular Institutions of Higher Education (No.1253G017)

数据原始分布特征等[2]。显然数据中不确定性的存在影响了传统方法对数据进行管理与分析的有效性，因此，不确定数据管理技术的研究大多通过对不确定数据建模以保留数据原始特征，从而以此为基础进行有效管理与分析。针对数据对象中存在不确定因素的 2 种情况：对象存在的不确定性以及属性值的不确定性[2]，通常利用可能世界模型[3]对前者进行建模，利用概率密度函数[4]描述后者对应的不确定数据；据此对不确定数据进行管理分析，包括不确定数据的存储与查询[5~8]，以及挖掘不确定数据中蕴涵知识的数据挖掘算法等[9]。

作为数据挖掘的重要技术之一，面向不确定数据的聚类分析得到了广泛关注。由于数据对象本身的不确定性，会干扰数据对象间的距离度量，使其同样具有不确定性，从而可能进一步对聚类结果产生影响。因此，相关研究多是在构建表示不确定数据对象的模型，以及定义不确定数据间距离的度量方式后，提出面向不确定数据特征的聚类算法[10~15]。这些算法虽然在一定程度上考虑了数据原始特征中的不确定性，但仍存在一定问题。首先，现有研究几乎均在假设不确定数据服从某种分布的前提下，获取描述不确定数据的概率密度函数[16]，虽然没有简单地忽略不确定性，造成数据信息的损失，但这种对不确定性进行简单假设的处理方式，并不一定符合实际数据的分布情况，会造成对数据的误解。其次，不确定数据间的距离度量也多采用期望距离或基于距离的概率密度函数进行计算，然而期望距离将可用概率描述的距离线性地整合为一个标量值，基于概率密度函数的方式则需要预先假设数据的分布，可见，无论哪种方式都难以反映数据间真实的差异。

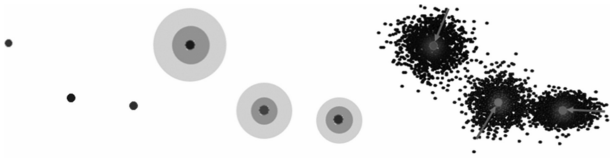
虽然文献[17,18]针对这些问题的提出不再基于对不确定对象构建的模型进行聚类，而是通过对不确定对象的多种可能状态所构成的可能世界进行代表性聚类，但对于不确定信息所构成的巨大状态空间，采用抽样技术提取代表性状态，无法尽量保留原始信息，而且其仍默认所有不确定对象服从同一分布。文献[19,20]虽强调了不同不确定对象的概率密度函数对距离度量影响的重要性，但基于直方图交叉核[21]的方法使距离计算结果受子区间宽度的影响较大，而且计算对象间距离时，直接基于表示不确定对象的样本点，使距离度量的效率过分依赖于样本点的个数，即样本点越多，越能体现不确定对象的实际分布，反而影响了距离度量的效率。

提高不确定数据聚类准确性的首要问题是为不同不确定对象构建体现其自身分布特征的模型，而无参估计方法无需假设数据分布密度，可基于数据特点直接获取其概率密度函数。本文主要面向连续型属性，提出一种基于改进式快速高斯变换核密度估计的不确定数据聚类算法，利用基于快速高斯变换的核密度估计方法，获取能够体现不确定数据真实分布的概率密度函数；据此计算不确定数据对象间的相似度，因为除了属性特征，表示不确定数据分布特征的概率密度函数也是体现不确定对象差异的一个重要特征；进而以此为基础，提出不确定数据聚类算法；通过在 UCI 以及真实数据集上的实验验证所提算法的有效性，最后得出本文结论。

2 基于改进式快速高斯变换的不确定数据模型

2.1 不确定数据对象

一个确定性对象可表示为特征空间中的一个点；而一个不确定性对象在面对由于数据采集、采用粗粒度数据集合等原因引起的属性值的不确定性时，需由一个区域来表示其可能性。以图 1 中的二维数据对象为例，确定性对象对应于二维空间中的一个点（如图 1(a)所示），而不确定性对象则需一个平面图形表示该对象的所有可能取值（如图 1(b)所示）；图形内的每个点均对应于不确定性对象以一定概率所处的可能位置，并且所有可能位置的概率之和为 1，即该平面图形体现了不确定性对象值的概率分布；为了得到其概率分布特点，可从不确定区域中获取一定数量的样本点（如图 1(c)所示），并基于样本点数据分析对象的不确定性特征。



(a) 确定性对象 (b) 不确定性对象 (c) 不确定特征提取

图 1 确定性对象与不确定性对象对比

可见构建数据模型的主要目标是获取尽可能符合不确定数据实际概率分布的模型；而概率密度函数是分析概率分布的重要方法，因此，构建不确定数据模型的根本在于为其构建接近其实际分布的概率密度函数。特别是在数据分布未知的情况下，

为了获取每个不确定对象的分布特征，利用非参估计方法估计不确定对象的概率密度函数是较合适的方法。因为非参估计方法可在不假设数据分布的前提下，基于样本点数据对不确定对象构建概率密度函数，而基于如式(1)所示的核密度估计方法求得的概率密度函数能够在样本点与目标数据量足够大时，收敛到任意一种密度函数[22]。

$$G(x_j)=\sum_{i=1}^{N}q_i\mathrm{e}^{-\frac{\|x_j-s_i\|^2}{h^2}} \tag{1}$$

其中，q_i 为权重系数，h 为核函数的平滑参数——带宽(bandwidth)，$\{s_i\}_{i=1,\cdots,N}$ 为服从某种未知分布的样本点。

可由式(1)估计该分布的概率密度函数。但其面临的一个主要问题是计算复杂度较高，在 N 个样本点上为 M 个目标数据估计密度函数，时间复杂度为 $O(MN)$ 。特别是实际应用领域中的数据以多变量形式居多，此时若将快速高斯变换扩展至多个维度时，所得概率密度函数为各维度之积[22]，此时时间复杂度则会迅速增长，达到 $O((MN)^d)$，其中，d 为数据维度，当 d 增加时，将呈指数级增长。

虽然针对该问题，文献[21]提出基于直方图变换计算概率密度函数，但基于核密度估计方法所得函数在光滑度与收敛到实际概率密度函数的速度方面比直方图变换方法更适合于数值型变量[23]。为了提高基于核密度估计计算概率密度函数的有效性，文献[24]提出一种快速高斯变换方法,利用 Hermite 以及 Taylor 展开式计算核函数，但其仍面临时间复杂度随数据维度增加而呈指数级增长的问题。因此，文献[25]以此为基础，仍利用 Taylor 展开式计算概率密度函数，只是将该过程直接应用于高维数据，而非如文献[24]先求每个维度的概率密度函数，再扩展至多个维度，从而降低总体时间复杂度。基于这种改进式快速高斯变换进行核密度估计，能够在保证时间复杂度的情况下，获得可以有效拟合实际分布的概率密度函数。因此，为保证模型构建的准确性与有效性，本文基于文献[25]提出改进快速高斯变换的核密度估计方法，构建不确定数据模型。

2.2　模型介绍

本文采用文献[24]中的多元标记法(multi-index notation)简化多维数据的书写方式。多元指数 $\alpha=(\alpha_1,\cdots,\alpha_d)$ 用于表示一组非负的 d 维整数，且其长度和阶乘分别表示为 $|\alpha|=\alpha_1+\cdots+\alpha_d$ 和 $\alpha!=\alpha_1!\alpha_2!\cdots\alpha_d!$，因此，可用于简化多维数据幂计算以及多项式的公式。此时一个 d 维变量 $x=(x_1,\cdots,x_d)$ 的幂计算可表示为 $x^\alpha=(x_1^{\alpha_1},\cdots,x_d^{\alpha_d})$；而多项式的幂计算利用多元标记法的方式表示如式(2)所示，即 α 各维度上的整数分别对应多项式中各项的次数。

$$\begin{aligned}&(x_1+x_2+x_3)^m\\&=\sum_{|\alpha|=m}\frac{m!}{\alpha!}(x_1+x_2+x_3)\\&=\sum_{i+j+k=m}x_1^ix_2^jx_3^k\end{aligned} \tag{2}$$

为了避免对不确定数据的分布进行预先假设，本文借助样本点构建表示不确定对象的区域，然后根据样本点获取体现每个不确定对象可能取值分布的概率密度函数。

令包含 N 个样本点的集合 $S_j=\{s_{j1},\cdots,s_{jN}\}$ 表示 d 维不确定对象 x_j 的不确定区域，其中样本点用于描述 x_j 的可能取值，并用于根据改进式快速高斯变换估计能够体现 x_j 不确定性分布的概率密度函数。

记样本点集合 $\{s_{ji}\}(1\leqslant i\leqslant N)$ 以 s_* 为聚集中心。对式(1)中的 $\mathrm{e}^{-\frac{\|x_j-s_i\|^2}{h^2}}$ 进行因式分解：$\mathrm{e}^{-\frac{\|x_j-s_i\|^2}{h^2}}=\mathrm{e}^{-\frac{\|x_j-s_*\|^2}{h^2}}\cdot\mathrm{e}^{-\frac{\|s_i-s_*\|^2}{h^2}}\cdot\mathrm{e}^{\frac{2(x_j-s_*)(s_i-s_*)}{h^2}}$；其中，前 2 个乘数可分别基于不确定数据 x_j 和样本点数据 s_i 直接计算；但第 3 个乘数的计算则涉及到任意 $x_j(1\leqslant j\leqslant M)$ 和 $s_i(1\leqslant i\leqslant N)$ 的乘积，特别在面向高维数据时，会导致较高的计算复杂度；利用 Taylor 展开式可减少其计算量，以多元标记法记为 $\mathrm{e}^{\frac{2(x_j-s_*)(s_i-s_*)}{h^2}}=\sum_{\alpha\geqslant0}\frac{2^{|\alpha|}}{\alpha!}\left(\frac{x_j-s_*}{h}\right)^\alpha\left(\frac{s_i-s_*}{h}\right)^\alpha$ [25]。此时式(1)可表示为如式(3)所示的以 s_* 为中心的多元泰勒展开式。

$$\begin{aligned}G(x_j)&=\sum_{i=1}^{N}q_i\mathrm{e}^{-\frac{\|x_j-s_*\|^2}{h^2}}\mathrm{e}^{-\frac{\|s_i-s_*\|^2}{h^2}}\cdot\\&\sum_{\alpha\geqslant0}\frac{2^{|\alpha|}}{\alpha!}\left(\frac{x_j-s_*}{h}\right)^\alpha\left(\frac{s_i-s_*}{h}\right)^\alpha\\&=\sum_{\alpha\geqslant0}C_\alpha\mathrm{e}^{-\frac{\|x_j-s_*\|^2}{h^2}}\left(\frac{x_j-s_*}{h}\right)^\alpha\end{aligned} \tag{3}$$

其中，$C_\alpha=\frac{2^{|\alpha|}}{\alpha!}\sum_{i=1}^{N}q_i\mathrm{e}^{-\frac{\|s_i-s_*\|^2}{h^2}}\left(\frac{s_i-s_*}{h}\right)^\alpha$。

使用上述基于样本点中心的方法可减少式(1)的计算量，因为原始方法中需基于描述对象不确定性的所有样本点计算其概率密度函数。而实际应用中 $\{s_{ji}\}(1\leqslant i\leqslant N)$ 内的样本点可能分布于多个聚簇，并形成多个聚簇中心。为了进一步提高式(3)的计算效率，在保证一定精度的前提下，对于泰勒展开式，可保留其中多项式次数大于 p 的项，此时，参与计算式(3)中 d 维泰勒展开式的项数为 $r_{pd}=\binom{p+d}{d}=\frac{(p+d)!}{d!p!}$ [22]；此时，不仅无需逐项求得泰勒展开式，而且 r_{pd} 远小于快速高斯方法中先求每个维度的概率密度函数再扩展至多个维度所需的 p^d；另外，为了进一步减少参与计算的样本点数量，可以忽略与目标对象 x_j 相距较远的样本点的影响，即仅考虑与不确定对象相邻聚簇中的样本点。基于上述考虑，由式(3)获取任意不确定对象 $x_j(1\leqslant j\leqslant M)$ 的概率密度函数可由式(4)估计得到。

$$\begin{aligned}G(x_j)&\simeq\sum_{sim(x_j-c_k)\geqslant h_\rho}\sum_{|\alpha|\leqslant p}C_\alpha^k\mathrm{e}^{-\frac{\|x_j-c_k\|^2}{h^2}}\left(\frac{x_j-c_k}{h}\right)^\alpha\\&=\sum_{sim(x_j-c_k)\geqslant h_\rho}\sum_{(|\alpha_1|+\cdots+|\alpha_d|)\leqslant p}C_\alpha^k\mathrm{e}^{-\frac{\|x_j-c_k\|^2}{h^2}}\cdot\\&\quad\left(\left(\frac{x_j-c_k}{h}\right)_1^{\alpha_1},\ldots,\left(\frac{x_j-c_k}{h}\right)_d^{\alpha_d}\right)\end{aligned}\tag{4}$$

其中，$C_\alpha^k=\frac{2^{|\alpha|}}{\alpha!}\sum_{s_i\in S_k}q_i\mathrm{e}^{-\frac{\|s_i-c_k\|^2}{h^2}}\left(\frac{s_i-c_k}{h}\right)^\alpha$；相邻聚簇通过计算与 x_j 的相似度不小于 h_ρ 的中心 c_k 所在聚簇 S_k 获得。

基于上述思想构建不确定数据模型，实为根据非参估计获取能够表示对象不确定性分布的概率密度函数,并基于改进式快速高斯变换方法提高模型构建效率。不确定数据模型构造的算法步骤如算法 1 所述。

算法 1 UncertainDataModel(D,S,h,h_ρ,p)

输入 不确定数据集 $D=\{x_1,x_2,\cdots,x_M\}$；
样本点集合 $SS_i=\{s_{i1},\cdots,s_{in}\},1\leqslant i\leqslant M$；
带宽 h；
相似度阈值 h_ρ；
多项式保留项的次数阈值 p；

输出 x_i 的概率密度函数 $G(x_i)(1\leqslant i\leqslant M)$；

1) for $i:=1$ to M do

2) 对不确定对象 x_i 的样本点集合 SS_i 进行聚类，形成聚簇 S_k 及其聚簇中心 $c_k(1\leqslant k\leqslant K)$；

3) 根据式(4)计算 x_i 的概率密度函数，即不确定数据模型 $G(x_i)$。

该算法中获取每个 d 维不确定对象 x_i 的概率密度函数时，忽略了与其相距较远的样本点。式(4)的计算量为 $O(n'r_{pd}n)$，其中，n'为与 x_i 相邻的聚簇个数，即与其相似性大于 h_ρ 的中心点的个数；r_{pd} 为式中保留的多项式的项数；n 为最大的聚簇中对象的个数。那么求得 M 个对象的不确定数据模型的总时间复杂度为 $O(M(n'r_{pd}n))$，其中，$n'n\ll N$，$r_{pd}\ll p^d$，使其远小于之前的 $O((MN)^d)$，同时，还能无需预先假设数据分布，获得符合对象不确定性分布的数据模型。

3 不确定对象间相似性度量

如前所述，利用期望距离或基于距离的概率密度函数是计算不确定数据间的距离度量的常见方式；但前者将可用概率描述的距离线性地整合为一个标量值，后者却也需要预先假设数据的分布，难以反映数据间真实的差异。而文献[21]基于直方图交叉核方法使距离计算结果受子区间宽度的影响较大，事实上不确定对象之间由于其不确定性影响，不仅在属性值范围上可能存在差异，在描述其不确定性分布的概率密度函数上还可能存在差异。因此，本文结合对象的属性特征以及不确定性分布特征 2 个方面度量不确定对象间的相似性。

考虑到不确定对象的属性值存在不确定性，而对象由值的可能分布区域来表示，那么不确定对象关于属性的相似性通过计算表示不确定对象可能取值的不确定区域的交集而获得。另外，对象间关于不确定性的相似性基于描述对象不确定性分布特征的概率密度函数进行计算。如果将对象基于属性的相似性和基于概率密度函数的相似性分别表示为向量 $\boldsymbol{r}$ 和 $\boldsymbol{p}$，那么根据式(5)，通过余弦相似度将两者相结合，以计算不确定性对象间的相似性。

$$sim(x_1,x_2)=\frac{\boldsymbol{r}\cdot\boldsymbol{p}}{\|\boldsymbol{r}\|\|\boldsymbol{p}\|},\|\boldsymbol{r}\|=(r_\mathrm{s},r_\mathrm{d}),\|\boldsymbol{p}\|=(p_\mathrm{s},p_\mathrm{d})\tag{5}$$

其中，$0\leqslant r_\mathrm{s},r_\mathrm{d},p_\mathrm{s},p_\mathrm{d}\leqslant 1$，且 $r_\mathrm{s}+r_\mathrm{d}=1$，$p_\mathrm{s}+p_\mathrm{d}=1$。该式表明当描述对象属性值可能分布的区

域在交集以及概率分布上均具有较高相似性时，不确定对象之间也更相似；但只要有一个相似性较低，即 r_s 或 p_s 较小时，对象间的差异就较明显，即 $sim(x_1,x_2)$ 较小。例如，如果表示 2 个对象属性可能取值的概率分布不同，即使属性取值的范围相似，也只能说明 2 个对象关于不确定性的分布并无相似，只是属性值的范围恰好相似，因而它们的相似性度量结果自然较小。

式(5)中的 r_s 为不确定性对象关于属性的相似性，假设 2 个 d 维不确定对象 x_1、x_2 在每个维度上的重合率分别为 $s_1,s_2,\cdots,s_d$，其中，$0\leqslant s_i\leqslant 1$, $1\leqslant i\leqslant d$，则 $r_s=\sqrt{\dfrac{s_1^2+\cdots+s_d^2}{d}}$。另外，不确定性对象基于概率分布的相似性 p_s 基于式(4)中的多项式进行计算。如果将 2 个多项式的系数分别表示为向量 $\boldsymbol{v}_1=(v_{11},v_{12},\cdots,v_{1m})$，$\boldsymbol{v}_2=(v_{21},v_{22},\cdots,v_{2m})$，如式(6)所示，通过计算两者的余弦相似度即可求得 p_s。其中，向量维度 m 取表示 $f(x_1)$ 和 $f(x_2)$ 的多项式中项数较大者。

$$p_s=sim(f(x_1),f(x_2))=\frac{\boldsymbol{v}_1\cdot\boldsymbol{v}_2}{\|\boldsymbol{v}_1\|\ \|\boldsymbol{v}_2\|} \tag{6}$$

已有研究中基于期望距离的度量公式[10]使计算效率受表示不确定对象的样本点数量的影响较大，即样本点数量越多，计算距离的效率越低，进而影响聚类效率；而本文利用改进式快速高斯变换方法获取描述对象不确定性分布特征的概率密度函数，首先保证了样本点越多所得密度函数越能符合对象的实际分布；其次，上述所提对象间相似性计算方法依赖于所得数据模型结果，而非像期望距离一样需直接依赖于样本点数据进行计算，因此，样本点的数量不会过多地影响相似性的计算效率，相反，足够多数量的样本点还会提高度量的准确性。

4　不确定数据聚类算法

不确定数据对象聚类的目标是在尽量保留不确定数据对象原始特征的基础上，按照相似度将其分为多个聚簇，使同一聚簇中的数据对象相似性较高，不同聚簇中的对象相似性较低。基于改进式快速高斯变换方法描述数据的不确定性特征在有效性方面保证了不确定数据模型的构建，据此度量不确定数据对象的相似性，进而进行不确定数据聚类，旨在从基本的数据模型、不确定数据对象相似性度量到聚类过程等多个环节提高不确定数据聚类的准确性。

在不确定性数据模型构建完成与对象间相似性度量后，不确定性数据聚类的 2 个重要基础已经具备。本文对不确定性数据聚类算法的研究，针对不确定性数据对象的特点，将传统聚类算法的基本思路扩展至不确定性领域。

考虑到 *K*-means 适于处理数值型属性，以及具有线性时间复杂度适于处理大数据集的特点，本文利用 *K*-means 算法对不确定数据对象进行聚类。在面向不确定性数据对象时，对算法中涉及的数据对象以及聚簇的不确定性进行处理。针对不确定对象的表示模型以及不确定对象相似性的度量方式，聚类过程中的聚簇均值表示为式(7)中的形式，其中，K 为聚簇个数。

$$\begin{aligned} &mean_i=(m_1,\cdots,m_d),1\leqslant i\leqslant K\\ &m_j=(\frac{1}{n_i}\sum_{h=1}^{n_i}ml_h,\frac{1}{n_i}\sum_{h=1}^{n_i}ms_h),1\leqslant j\leqslant d \end{aligned} \tag{7}$$

与不确定对象类似，聚簇均值也包含 2 种重要特征：聚簇内所有对象的属性值特征与描述聚簇内对象不确定性的分布特征；前者即为聚簇内对象的平均不确定区域，由式(7)通过计算各属性维度上的属性值区间可得；后者则为聚簇内所有对象的概率密度函数的平均，由式(8)可求。进而每个不确定对象与聚簇均值间的相似性均可通过式(5)求得。

$$pdf(mean_i)=\frac{1}{n_i}\sum_{j=1}^{n_i}G(x_j),1\leqslant i\leqslant K \tag{8}$$

基于上述不确定对象的表示模型、不确定对象的相似性度量方式以及各聚簇均值的定义，面向不确定数据的 *K*-means 聚类算法的步骤如算法 2 所述。

算法 2　IUK-means$(D,G(x_i),K)$

输入　不确定数据集 $D=\{x_1,x_2,\cdots,x_M\}$；

概率密度函数 $G(x_i),1\leqslant i\leqslant M$；

聚簇数目 K；

输出　不确定对象聚类结果 $C=\{C_1,\cdots,C_K\}$；

1) 从 D 中任意选择 K 个对象作为初始聚簇中心;

2) repeat

3)　for　$i:=1$　to M do

4) 根据式(5)计算对象 x_i 到各聚簇均值的相似

性，得 $sim=\{sim_1,sim_2,\cdots,sim_K\}$；

5) 将对象 x_i 分配至最相似(即 $sim_j=\max(sim_1,sim_2,\cdots,sim_K)$)的聚簇 C_j；

6) end for

7) 根据式(7)和式(8)，更新 K 个聚簇的均值 $mean=\{mean_1,\cdots,mean_K\}$，描述聚簇内对象不确定性的分布特征 $pdf(mean_j)(1\leqslant j\leqslant K)$；

8) 计算目标函数 $E=\sum_{j=1}^{K}\sum_{o\in C_j}sim(o,mean_j)^2$；

9) until 目标函数 E 收敛；

在从构建不确定对象的数据模型到聚类的过程中，如前所述，基于 N 个样本点计算 M 个概率密度函数的时间复杂度为 $O(M(n'r_{pd}n))$；另外，基于 K-means 为不确定对象划分聚簇的聚类算法时间复杂度为 $O(tKdM)$，其中，t 为迭代次数，d 为不确定对象的属性维度。

5 实验

为了验证上述面向不确定数据的聚类算法的有效性，并在 UCI 和真实数据集上，与其他算法进行对比。本文的所有算法都是基于 C++编程语言实现，实验在 CPU 4 GHz、16 GB RAM、SSD 存储介质的 PC 上进行。首先与传统核密度估计方法对比获取概率密度函数的效率，说明所提数据模型构建方法即使面向多维不确定数据对象，以及增加表示不确定对象的样本数量时，仍具有较高的效率；进一步与 UK-means[10]、MC[16]与 LMHIK[19]等不确定数据聚类算法比较，说明所提的不确定数据聚类算法具有较高准确性。

表 1 所列数据集为几种常见的验证聚类结果的数据集，通过在这些数据集上与其他不确定聚类算法的比较，说明所提算法的有效性。

表 1 UCI 数据集

数据集	对象个数	属性个数	类别数
Iris	150	4	3
Ecoli	336	8	8
Yeast	1 484	8	10
Abalone	4 177	8	16
Letter	20 000	16	26

所选真实数据集是国内某电信公司在某市的全部基站所采集的服务人数数据。该数据集的采集周期为 2013 年 2 月，共计 28 天，被采集对象为该电信公司的 3G 用户群，记录了在该城市中的每个基站在采集周期内随机观测的服务人数，具体内容包括基站的地理信息、采集时间、服务人数和基站 ID，共计 3 096 个基站以及约 1.3 亿条记录。表 2 为其主要属性以及数据示例，包含了每个基站下每次随机采集的服务用户人数数据。其中，经度和纬度是基站唯一标识，代表该基站物理位置。由于采集时的客观条件，该数据集中可能具有不确定性，因为用户极可能在采集时间点附近由一个基站移动到另一基站覆盖区域。简单通过降维、去噪等数据预处理技术可能会忽略这些数据的实际意义，因为虽然在某一采集时间点上数据具有不确定性，但在更粗粒度的集合中，如 1 h 内某基站下的用户的样本数据，则可能具有体现基站服务模式的意义。本文通过在该数据集上的实验说明所提算法的实用性。

表 2 各基站下服务人数数据集

服务人数	采集时间	经度/°	纬度/°
889	2013/2/4 1:00	126.683 3	45.750 97
888	2013/2/4 1:00	126.683 3	45.750 97
⋮	⋮	⋮	⋮
443	2013/2/4 2:00	126.683 3	45.750 97
441	2013/2/4 2:00	126.683 3	45.750 97
⋮	⋮	⋮	⋮
510	2013/2/4 1:00	128.038 8	45.950 6
⋮	⋮	⋮	⋮

5.1 UCI 数据集

5.1.1 不确定性的生成

为了验证不确定数据聚类算法的有效性，相关研究大多利用人工方式向确定性数据集中添加不确定性[10,13,16]。本文采用文献[16]所提方法为表 1 中的数据集生成不确定性信息，分别基于高斯分布和均匀分布生成样本点数据，即为数据集 D 中的每个 d 维对象 o，利用高斯分布或均匀分布构建体现其不确定性的 s 个样本点，并用变量 e 表示不确定性。

1) 高斯分布：为每个维度选择一个统一的标准差 $\sigma_j\in[0,e](1\leqslant j\leqslant d)$；对于任意 $o\in D$，将该对象作为分布的中心样本点 $g=o$，然后针对每个属性维度，基于 $\mu=g,\sigma=\sigma_j$ 的高斯分布生成其余 $s-1$ 个样本点。

2) 均匀分布：为每个对象基于均匀分布构建描述其不确定性的样本点，样本点即构成一个超矩形 r。为每个维度从 $[0,e]$ 中选择一个统一的范围 $[x,y]_j$，作为该维度上均匀分布的区间范围；对于任意 $o\in D$，将 o 作为超矩形 r 的中心点，然后针对每个属性维度，基于 $[x,y]_j$ 范围上的均匀分布生成其余的 $s-1$ 个样本点。

2 种情况中所生成的 s 个点均可作为描述一个不确定对象的样本点，而且该方法为每个对象所生成的表示不确定信息的分布，可能由于参数的不同而不同，这与实际应用中的数据集中各不确定对象的分布特征可能不同的事实相似。因此，这种利用不同分布为数据集生成不确定性的方式，有助于验证所提算法是否能够针对不同分布的不确定数据进行有效聚类。

5.1.2 对比结果

首先，当增加表示目标对象不确定性的样本点数量时，对比本文所提基于改进式快速高斯变换与传统核密度估计方法构建不确定数据模型时的算法运行效率。图 2 为对表 1 每个数据集中的每个对象基于上述 2 种分布，分别为其生成 100、400 个样本点，以表示每个对象的不确定性，并据此分别采用改进式快速高斯变换与传统核密度估计方法构建不确定数据模型的运行时间的对比结果。其中，每个柱形为当表示不确定性的样本点个数为 100 或 400 时，利用本文所提算法与传统核密度估计方法构建不确定数据模型的时间占两者总数的比例。结果显示运行时间并未随样本点数量的增加而显著增加；另外，表 1 中几个数据集的属性个数不同，可见属性个数也不会成为本文所提算法的瓶颈。相反，使用传统核密度估计方法构建数据模型的时间则因样本点数量的增加呈现显著增加的趋势。这是因为本文基于改进式快速高斯方法构建不确定数据模型，一是基于样本点中心计算且仅考虑与不确定对象相邻聚簇中的样本点的方式减少了参与计算的样本点数量，而原始方法中需基于描述对象不确定性的所有样本点计算其概率密度函数；二是将泰勒展开直接应用于高维数据，相对传统的先求每个维度的概率密度函数再扩展至多个维度的做法，降低了时间复杂度。因此，计算量不会随样本点数量以及属性个数的增加而显著增加，从而保证了模型构建的效率。

然后，通过与 UK-means、MC 与 LMHIK 等不确定聚类算法的对比，进一步验证所提算法的准确性。本文利用常见的外部评价指标 Purity 和 Entropy[26]、NMI 以及 F-measure[27]分析聚类结果有效性。图 3 与图 4 所示为每个不确定对象生成 100 个样本点的聚类有效性对比结果。算法 MC 与 LMHIK 是近年来提出的较典型的不确定数据聚类算法，相较于 UK-means 等早期的不确定聚类算法，在 UCI 数据集上均可取得较准确的聚类结果。从图 3 中的指标值对比可见，本文所提算法 IUK-means 能够在多数数据集上获得与 MC 相当的聚类结果，并且明显优于 UK-means。

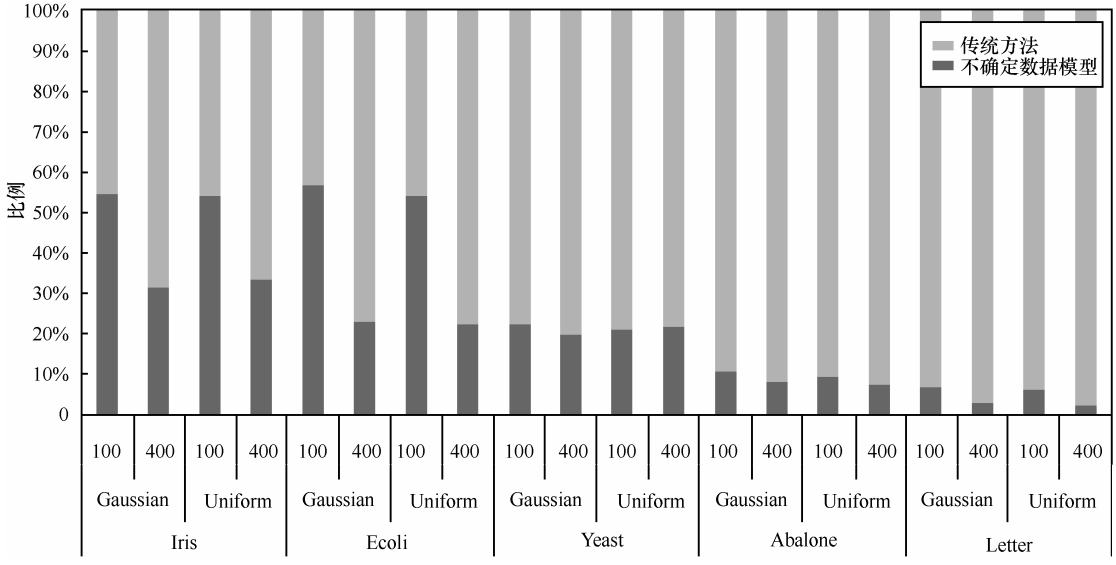


图 2　数据模型构建运行时间相对对比结果

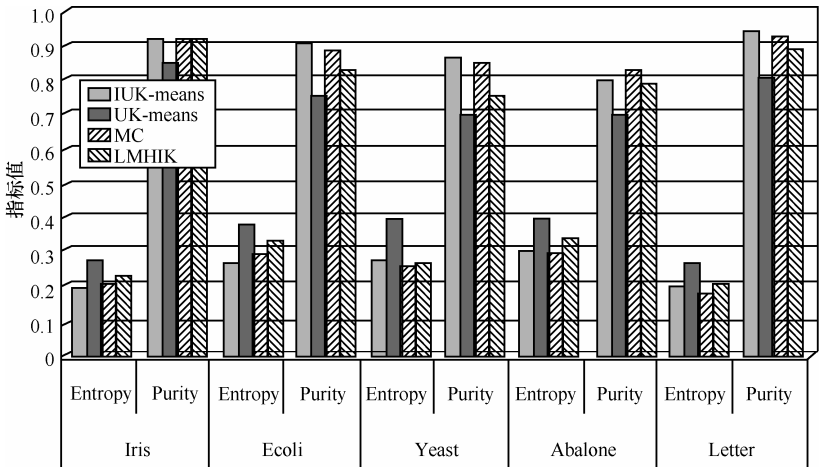


图 3　UCI 数据集上基于 Purity 和 Entropy 的聚类准确性对比结果

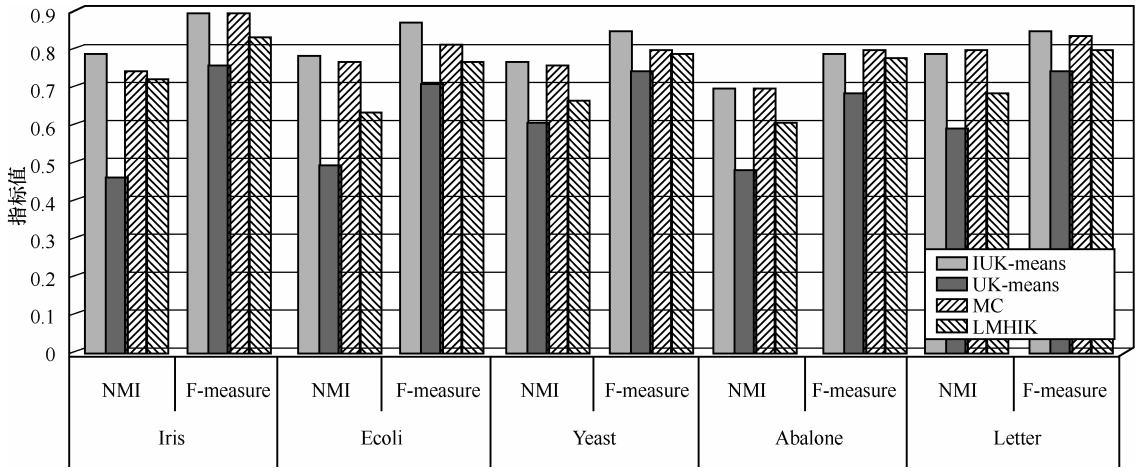


图 4　UCI 数据集上基于 NMI 和 F-measure 的聚类准确性对比结果

另外，基于文献[28]提出的非参数化统计测试方法，对几种算法在表 1 数据集上的聚类准确性进行成对数据 T 检验，以进一步说明对比算法间的差异，表 3 即为关于各指标值的检验结果。通过表 3 中的非参统计检验结果可比较所提算法与对比算法关于聚类结果准确性的差异，当显著性水平为 0.05 时，IUK-means 与 UK-means，以及 LMHIK 间的差异较为显著（p 值均小于 0.05），而与 MC 并不具有显著性差异（p 值大于 0.05）。上述结果得益于所提算法充分利用数据特点构建不确定数据模型，并据此进行客观的对象相似性度量，以两者为基础的聚类效果较好。而 MC 默认不确定对象均服从同一分布，因此，在人工生成的不确定数据集上，能够获得较好的效果。而 LMHIK 算法尽管也考虑了概率密度函数表示的不确定性分布特征对距离度量的影响，但其基于直方图交叉核获取不确定性分布特征，使它的结果过多依赖直方图组距的选择，因此，其准确性略低于 IUK-means。

UCI 数据集上的实验结果说明本文所提算法能够适用于不同的不确定性分布情况；此外，算法效率不仅不受样本点数量以及属性个数的影响，还能取得相对较准确的聚类结果。下面进一步在真实数据集上验证算法的有效性与实用性。

表 3　聚类结果准备性的成对数据 T 检验结果

指标	IUK-means & UK-means	IUK-means & MC	IUK-means & LMHIK
Purity	0.001	0.746	0.039
Entropy	0.001	0.446	0.027
NMI	0.001	0.353	0.002
F-measure	0.001	0.221	0.017

5.2　真实数据集

本文基于表 2 所示的数据集，针对如前所述的数据中的不确定性，将更粗粒度的以小时为单位的数据

对象作为处理对象（即不确定对象），基于每一基站覆盖区域内的服务用户数据分析基站服务的模式。此时，可将表 2 中数据组织成表 4 中的形式，基站的定义和表 2 相同，那么原数据集中每分钟采集周期内观测到的服务用户数据，将作为构成以小时为单位的分析周期内的样本点，并用于分析每小时内每一基站覆盖区域下用户的通话量分布情况，即用于构建粗粒度对象的不确定数据模型。然后对该数据集进行聚类，试图通过用户通话行为的聚集特点区分基站所属区域的类型。

表 4　用于不确定分析的基站下服务人数数据集

服务人数观测值	采集周期	经度/°	纬度/°
889 888 … 888 887	2013/2/4 1:00	126.683 3	45.7509 7
443 441 … 444 442	2013/2/4 1:00	126.683 3	45.7509 7
418 417 … 417 418	2013/2/4 1:00	126.683 3	45.7509 7
…	…	…	…
510 509 … 511 508	2013/2/4 1:00	128.038 8	45.950 6
…	…	…	…

由于实际应用中难于获得不确定对象的原始分类信息，因此，本文利用相对评价指标 Dunn 和 DB[29]衡量真实数据集上聚类的准确性。相对指标度量聚簇内的紧密与聚簇间的分离程度。通过与其他聚类方案的对比指标值来评估聚类结果的相对有效性。Dunn 值越大聚类效果越好，而 DB 值越小则聚类效果越好。

图 5 为真实数据集上关于相对评价指标的聚类准确性对比结果。图中结果显示本文所提不确定数据聚类算法在 h=1.46，保留的多项式次数 p=20 时，得到的结果具有最大 Dunn 值以及最小的 DB 值，即相较于对比算法能获得更准确的聚类结果。

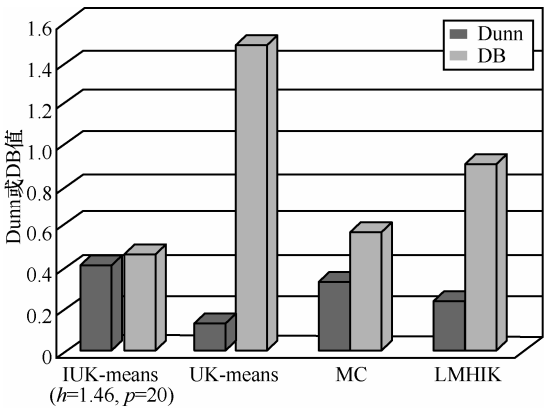


图 5　真实数据集上聚类准确性对比结果

图 6 为在该数据集上运行几种对比算法的时间对比结果。可见本文所提算法能够获得相对较快的运行效率，与同样具有线性时间复杂度的 LMHIK 具有接近的运行时间，并优于另外 2 种对比算法。从上述结果来看，本文所提算法不仅能够具有较高的运行效率，同时也能获得相对较准确的聚类结果。

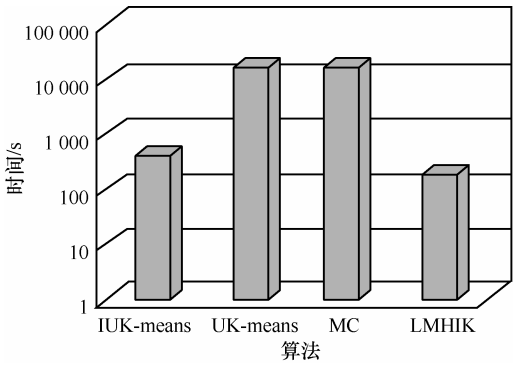


图 6　真实数据集上聚类算法运行时间对比结果

主要原因一是依赖于其前述所提算法的特点，能够在无需对不确定性分布进行假设的前提下，充分利用数据特点构建不确定数据模型，同时，基于改进式快速高斯变换方法保证模型构建的效率；二是 UCI 数据集上的不确定性是由人工生成的，使对象的不确定性服从同一分布，这种方式本身有利于其他聚类算法（如 MC 和 LMHIK）的过程应用；而一旦面对真实数据集，数据的不确定性分布则不再具备服从同一分布的特征。因此，IUK-means 能够基于改进式快速变换的非参估计方法，有效地构建充分体现不确定性分布特征的数据模型，据此度量不确定数据间的相似性也能获得相对客观的结果，进而面向数据的不确定性进行聚类；最终使其相较于对比算法，能够获得相对较准确的聚类结果。

图 8 为聚类后的结果示意，标示了每个聚簇中的基站在 24 h 内服务人数的上限与下限，由此说明了各聚簇间的差异。如其中的聚簇 1 和聚簇 8 在 20:00 后的服务人数相较于 10:00 左右明显下降，符合商业区的人群活动模式，而聚簇 2 和聚簇 3 在 20:00 后的服务人数则无显著变化，甚至还有上升趋势，符合居民区的人群活动模式；可见服务人数的分布特征能够用于区分基站所提供的服务模式，那么基于此聚类结果可进一步为基站针对不同的通话区域类型制定对应的服务策略。

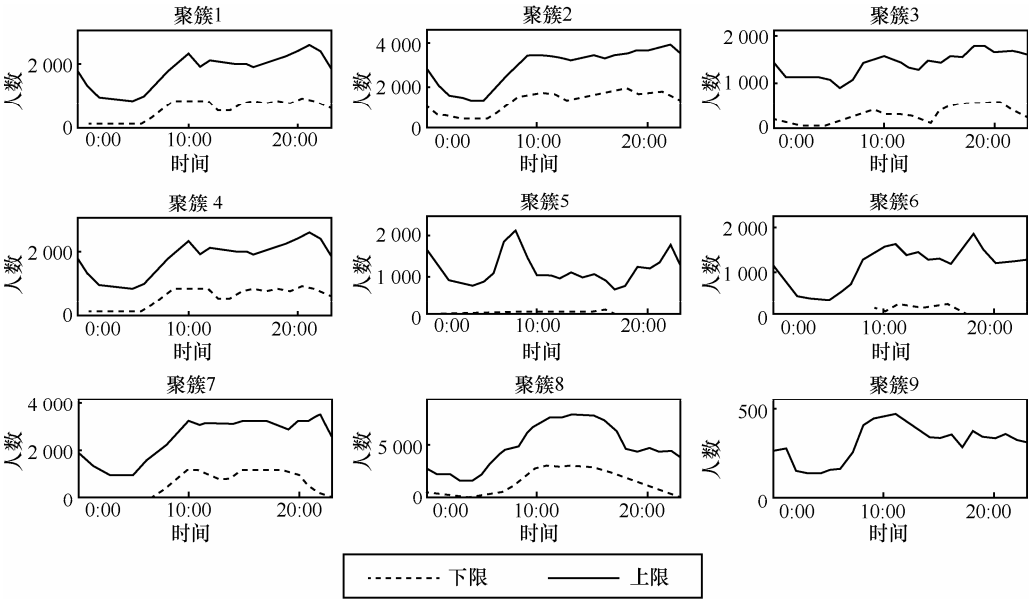


图 7　真实数据集上聚类结果示意

真实数据集上的聚类结果显示了基于本文所提不确定数据聚类算法能够获得相对较准确的聚类结果；另外，所得聚类结果还能为实际应用领域提供决策制定的依据。这些均说明了所提算法的有效性和实用性。

6　结束语

本文针对不确定数据聚类算法中存在的问题，以提高算法效率与准确性为主要目标，提出一种基于快速高斯变换的不确定数据聚类算法。基于改进式快速高斯变换构建不确定数据模型，无需假设数据的不确定性分布，为每个对象构建符合实际特征的不确定数据模型，并且提高了模型构建的计算效率；为了充分利用不确定数据特征，结合数据不确定性特征与属性特征度量对象间相似性，度量结果更加客观；进而面向数据的不确定性，提出不确定聚类算法，使算法在效率与准确性方面均得到改善。在 UCI 以及真实数据集上的对比实验进一步验证了所提算法的准确性、有效性与实用性。

参考文献：

[1] WANG Y, LI X, LI X, et al. A survey of queries over uncertain data[J]. Knowledge and Information Systems, 2013, 37(3):485-530.

[2] SEN P, DESHPANDE A. Representing and querying correlated tuples in probabilistic databases[C]//IEEE 23rd International Conference on Data Engineering. Istanbul, Turkey, IEEE, 2007:596-605.

[3] MUZAMMAL M, RAMAN R. Mining sequential patterns from probabilistic databases[J]. Knowledge and Information Systems, 2015, 44(2):325-358.

[4] SOLIMAN M, ILYAS I, CHANG K. Top-*k* query processing in uncertain databases[C]//IEEE 23rd International Conference on Data Engineering. Istanbul, Turkey, IEEE, 2007:896-905.

[5] AGGARWAL C. Managing and mining uncertain data[M].USA: Springer, 2009: 1-35.

[6] BARBARÁ D, GARCIA-MOLINA H, PORTER D. The management of probabilistic data[J]. IEEE Transactions on Knowledge and Data Engineering, 1992, 4(5):487-502.

[7] ANTOVA L, JANSEN T, KOCH C, et al. Fast and simple relational processing of uncertain data[C]//IEEE 24th International Conference on Data Engineering. Cancun, Mexico, IEEE, 2008:983-992.

[8] TANG R, CHENG R, WU H, et al. A framework for conditioning uncertain relational data[J]. Database and Expert Systems Applications, 2012, 37(3):71-87.

[9] TASKAR B, SEGAL E, KOLLER D. Probabilistic classification and clustering in relational data[C]//The Seventeenth International Joint Conference on Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers Inc. 2001: 870-878.

[10] NGAI W K, KAO B, CHUI C K, et al. Efficient clustering of uncertain data[C]//The Sixth International Conference on Data Mining. 2006: 436-445.

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

[12] KRIEGEL H, PFEIFLE M. Hierarchical density-based clustering of uncertain data[C]//Fifth IEEE International Conference on Data Mining.Houston, USA, IEEE, 2005: 689-692.

[13] 李云飞，王丽珍，周丽华. 不确定数据的高效聚类算法[J].广西师范大学学报：自然科学版, 2011, 29(2): 161-166.
LI Y F, WANG L Z, ZHOU L H. More efficient clustering algorithm over uncertain data[J]. Journal of Guangxi Normal University (Natural Science Edition), 2011, 29(2): 161-166.

[14] 金萍, 宗瑜, 屈世超, 等. 面向不确定数据的近似骨架启发式聚类算法[J]. 南京大学学报 (自然科学), 2015, 51(1): 197-205.
JIN P, ZONG Y, QU S C, et al. Approximate backbone guided heuristic clustering algorithm for uncertain data[J]. Journal of Nanjing University (Natural Sciences), 2015, 51(1): 197-205.

[15] 肖宇鹏, 何云斌, 万静, 等. 基于模糊 C-均值的空间不确定数据聚类[J]. 计算机工程, 2015, 41(10): 47-52.
XIAO Y P, HE Y B, WAN J, et al. Clustering of space uncertain data based on fuzzy C-means[J]. Computer Engineering, 2015, 41(10): 47-52.

[16] CORMODE G, MCGREGOR A. Approximation algorithms for clustering uncertain data[C]//The 27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Vancouver, Canada, ACM, 2008: 191-200.

[17] ZÜFLE A, EMRICH T, SCHMID K A, et al. Representative clustering of uncertain data[C]//The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.New York, USA, ACM, 2014:243-252.

[18] SCHUBERT E, KOOS A, EMRICH T, et al. A framework for clustering uncertain data[C]//The 41st International Conference on Very Large Data Bases.Hawaii, USA, VLDB Endowment, 2015: 1976-1979.

[19] XU L, HU Q, HUNG E, et al. Large margin clustering on uncertain data by considering probability distribution similarity[J]. Neurocomputing, 2015, 158: 81-89.

[20] JIANG B, PEI J, TAO Y, et al. Clustering uncertain data based on probability distribution similarity[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(4):751-763.

[21] LÓPEZ-RUBIO E, JOSE. Probability densty function estimation with the frequency polygon transform[J]. Information Science, 2015, 298: 136-158.

[22] ELGAMMAL A, DURAISWAMI R, DAVIS L S. Efficient kernel density estimation using the fast gauss transform with applications to color modeling and tracking[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2003, 25(11): 1499-1504.

[23] SCOTT D. On optimal and data-based histograms[J].Biometrika, 1979, 66(3): 605-610.

[24] GREENGARD L, STRAIN J. The fast Gauss transform[J].SIAM Journal on Scientific and Statistical Computing, 1991, 12(1):79-94.

[25] YANG C, DURAISWAMI R, GUMEROV N, et al. Improved fast gauss transform and efficient kernel density estimation[C]//Ninth IEEE International Conference on Computer Vision.Nice, France: IEEE, 2003: 664-671.

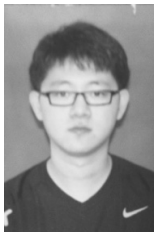
[26] ZHAO Y, KARYPIS G. Criterion functions for document clustering: experiment and analysis[R]. Technical Report TR 01-40,Department of Computer Science, University of Minnesota, USA, 2001.

[27] MANNING C D, RAGHAVAN P, SCHUTZE H, 著. 王斌, 译. 信息检索导论[M]. 北京:人民邮电出版社, 2010.
MANNING C D, RAGHAVAN P, SCHUTZE H. WANG B, translate. Introduction to information retrieval[M]. Beijing: PTPRESS 2010.

[28] DEMSAR J. Statistical comparisons of classifiers over multiple data sets[J]. Journal of Machine Learning Research, 2006(7): 1-30.

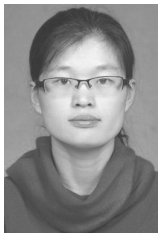
[29] ARBELAITZ O, GURRUTXAGA I, MUGUERZA J, et al. An extensive comparative study of cluster validity indices[J]. Pattern Recognition, 2013, 46(1):243-256.

作者简介：



迟荣华（1981-），男，黑龙江哈尔滨人，哈尔滨工程大学博士生，主要研究方向为复杂网络、不确定性研究等。

程媛（1985-），女，黑龙江哈尔滨人，哈尔滨理工大学讲师，主要研究方向为数据挖掘、不确定性研究等。



朱素霞（1978-），女，山东寿光人，哈尔滨理工大学副教授，主要研究方向为高性能体系结构、并行计算。

黄少滨（1965-），男，黑龙江哈尔滨人，哈尔滨工程大学教授、博士生导师，主要研究方向为分布式计算与仿真、模型检测、数据集成等。

陈德运（1962-），男，黑龙江哈尔滨人，哈尔滨理工大学教授、博士生导师，主要研究方向为图像处理、探测和成像技术。