不确定数据聚类的U-PAM算法和UM-PAM算法的研究

何云斌 张志超 万 静 李 松
(哈尔滨理工大学计算机科学与技术学院 哈尔滨 150080)

摘 要 UK-means算法在处理不确定数据时对孤立点非常敏感,而且事先必须已知不确定数据的分布函数或概率密度,然而这在实际中往往很难获得。因此,针对UK-means在处理不确定测量数据时的不足,首先提出了基于区间数的PAM不确定聚类算法——U-PAM,该算法用区间数和标准差合理地描述了不确定测量数据的不确定性,进而完成有效的聚类;其次,针对海量不确定测量数据难以聚类的问题,基于U-PAM聚类算法,采用抽样技术提出了处理海量不确定测量数据的算法——UM-PAM算法,该算法先抽样,对样本数据聚类,然后再总体聚类;最后,基于U-PAM算法和CH聚类的有效性指标函数对聚类结果进行分析,以确定最佳聚类数。实验理论表明,所提算法聚类效果明显。 **关键词** 不确定数据,区间数,聚类算法,PAM
中图法分类号 TP311.13 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.6.052

Research for Uncertain Data Clustering Algorithm: U-PAM and UM-PAM Algorithm

HE Yun-bin ZHANG Zhi-chao WAN Jing LI Song
(School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China)

Abstract UK-means algorithm is very sensitive to outliers in dealing with uncertain data, and the probability density or distribution function of uncertain data must be acquired in advance. However, it is often difficult to obtain in practice. For the shortage of UK-means in dealing with uncertainty measurement data, this paper firstly proposed a new algorithm namely U-PAM, based on PAM algorithm and intervals. It describes the uncertainty of measurement data with intervals reasonably and standard deviation so as to complete clustering effectively. Secondly, it is often difficult to cluster for the massive of data. For this regard, according to sampling techniques, this paper proposed the UM-PAM algorithm so as to deal with massive of uncertainty measurement data efficiently. It primary clusters sample data, and then clusters overall. Finally, the U-PAM algorithm can analyze the clustering result by combining with the CH validity index to determine the optimal clustering number. Experimental results show that the proposed algorithm can give effective clustering result obviously.

Keywords Uncertain data, Intervals, Clustering, PAM

1 前言

近些年,随着不确定数据在传感器网络、数据集成和互联网上不断出现,不确定数据的研究越来越受到人们的重视。不确定数据的特点是各数据点的属性不是固定不变的,而是以某个概率值出现[1]。聚类分析是数据挖掘中很重要的一个分支,随着不确定数据的日益增多,不确定数据聚类的研究已成为数据挖掘领域的热点之一。

聚类分析是在没有学习集的情况下按照某种准则将数据集划分成为若干个组,使同一组内的数据对象有很高的相似度,组间的数据对象有较低的相似度[2]。已有的聚类分析算法主要分为基于划分的、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,在研究不确定聚类算法时,主要借鉴传统聚类算法的思想,将聚类算法应用到不确定数据中。其中Chau等人提出了经典的UK-Means不确定数据聚类算法[4],该算法利用一个概率密度函数来表示数据的不确定性,提高了对不确定数据聚类的准确度,然而这种方法使用了大量的积分运算,时间复杂度较高。Lee等人将UK-means算法中的期望距离简化为数据点之间的距离,提出了CK-means聚类算法[5]。该算法相对于UK-means算法降低了算法的时间复杂度,但是存在任意初始值选择的不足,会降低聚类效果,并且对孤立点较为敏感。彭宇等人利用区间数和标准差来表示不确定测量数据,提出了UIDK-means聚类算法[6],该算法不需要预知数据的概率分布即可对数据进行聚类,而且时间复杂度相对于UK-means算法的要低,但是,该算法会"继承"K-means算法任意初始值和孤立点敏感的不足,可能降低聚类质量。任培花等人提出了基于DKC改进的K-means的U2d-Kmeans算法[7],该算法对数据集进行预处

到稿日期:2015-05-13 返修日期:2015-08-05 本文受黑龙江省教育厅科学技术研究项目(12511100),黑龙江省自然科学基金项目(F201302,F201134)资助。

何云斌(1972—),男,教授,硕士生导师,主要研究方向为数据库理论与应用、时空数据库、数据挖掘,E-mail:hybha@163.com;**张志超**(1988—),男,硕士,主要研究方向为空间数据挖掘;**万 静**(1972—),女,博士,教授,硕士生导师,主要研究方向为数据库理论及应用;**李 松**(1977—),男,博士,副教授,主要研究方向为空间数据库理论及应用。

理，采用累积距离的方法确定初始中心点，避免了算法对初始中心点敏感的问题，但是数据中引入不确定因素会增加算法的时间复杂度。Kao B 等人[8]提出一种基于 Voronoi 图和 R-tree 的剪枝策略，减小了衡量样本相似度的计算量，但该策略在构造 Voronoi 图和 R-tree 时会产生较大时间开销。Jian H 等人提出了高维不确定聚类算法即 HDUDEC 算法[9]，该算法可以有效过滤噪声数据和解决维度灾难问题，但是该算法需要设定阈值，如果阈值设置不准确则会影响聚类质量。

Gullo 等人提出了 UK-mediods 算法[10]，该算法以不确定性数据对象作为簇中心，避免了孤立点对算法的影响；然而，该算法采用局部搜索技术设计启发式搜索，仍然存在初始解敏感的问题，而且事先必须已知不确定数据的概率分布，然而这在实际中往往很难获得。对此，本文针对 UK-mediods 在处理不确定测量数据的不足，提出了基于区间数的 PAM 不确定聚类算法：U-PAM，该算法不需要计算概率密度函数值，因此减小了积分的运算量，而且在数据概率密度或分布缺失的情况下也可对数据进行有效聚类，对孤立点也不敏感。在此算法的基础上提出了 UM-PAM 算法，该算法采用抽样划分技术，可以有效的处理海量不确定数据集。最后，针对划分算法的 k 值必须事先确定的问题，提出了 U-PAM 算法和 CH 聚类有效性指标函数相结合的方法对聚类结果进行分析，最终确定最佳聚类数 k。

2 基础知识

2.1 区间数

定义 1（区间数）[11] 给定 $A_L, A_R \in R^d$ 且 $A_R \geqslant A_L$，称集合：$A=[A_L, A_R] \triangleq \{u \mid A_L \leqslant A_R\}$ 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 $A_L = A_R$ 时，即上下界相等时，区间数变为精确数。

定义 2（区间数的中点和半径）[11] 给定区间数 $A=[A_L, A_R]$，令 $\alpha_A = (A_L - A_R)/2$，$m_A = (A_L + A_R)/2$，则有：$A_L = m_A - \alpha_A$，$A_R = m_A + \alpha_A$ 其中，m_A 为区间数 A 的中点，α_A 是区间数 A 的半径，因此区间数也可表示为 $[m_A - \alpha_A, m_A + \alpha_A]$。

定义 3（区间数的距离）[11] 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：

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

由以上定义可知，区间数是用一个区间表示的数，这样，区间数就可以用区间的形式来表示数据的一种不确定性，继而可以应用到不确定数据的聚类算法中。由定义 3 可知道，当用区间数去表示两个不确定数据对象时，不确定数据对象之间的距离就可用区间数之间的距离来表示，这样，对不确定数据对象进行聚类时，即可用区间数之间的距离来表示不确定数据对象的相似度，区间数之间的距离越大，相似度越小；区间数之间的距离越小，相似度越大。

区间数还可以表示多维数据模型，当数据模型是一维时，区间数模型为数轴上的一条线段；当数据模型为二维时，区间数模型为二维平面的矩形框；当数据模型为三维时，区间数模型为立方体模型；当数据对象为多维时，数据模型为一个超几何体。

2.2 PAM 算法

PAM 算法[12]是对经典 k-means 算法的一个改进，改进了 K-means 算法对数据孤立点和输入噪声敏感的问题，该算法的结果是输出 k 个簇，使簇内相似度达到最小，簇间相似度达到最大。设聚类的数据集为 $O_1, O_2, O_3, \cdots, O_n$，聚类中心为 $\{k_1, k_2, k_3, \cdots, k_k\}$，聚类得到的 k 个簇为 $\{C_1, C_2, C_3, \cdots, C_k\}$，PAM 算法的准则函数如下：

$$E = \sum_{i=1}^{k} \sum_{p \in c_j} |p - o_i|^2$$

PAM 算法的思想：首先随机选择 k 个对象作为初始中心点，其次将剩余的对象中与中心点的距离最小的对象分配给各中心点。然后反复地用非代表对象来替换代表对象，以提高聚类的质量，聚类质量用准则函数来评估，该函数度量一个非代表对象是否是当前一个代表对象的好的代替，如果是就进行替换，否则不替换，最后输出 k 个更新的簇。

PAM 算法改善了 K-means 算法对孤立点敏感而容易使聚类结果陷入局部最优的问题。然而该算法处理的数据是确定数据集，当输入数据不确定时，该算法无法完成有效的聚类。故本文在 PAM 算法的基础上结合区间数和统计学知识提出了 U-PAM 算法，该算法主要处理不确定测量数据集，而且在数据集的分布函数或概率密度未知的情况下，仍可对其进行有效的聚类。

3 U-PAM 聚类算法

第 2 节介绍了区间数的定义和区间数距离计算的一些概念，区间数可以用区间的形式表示数据的不确定性；而区间数的距离可以表示不确定数据之间的距离，这样就可以用区间数的距离作为不确定测量数据的相似度准则，进而对不确定数据进行聚类算法。

本节针对不确定测量数据主要提出了 U-PAM 聚类算法，该算法基于统计学中的一些知识，采用区间数和标准差表示不确定测量数据，在不确定测量数据的概率密度未知的情况下即可对数据进行聚类，同时还解决了传统 UK-means 算法对孤立点敏感的问题。U-PAM 聚类算法的基本思想：用区间数和标准差表示不确定数据，进而转化为对确定的区间数进行聚类。首先用区间数分别表示 N 个不确定测量数据，其次随机选出 K 个区间数作为聚类中心点，并计算剩余的区间数分别到中心点的距离，并将距离最小的分配给中心点。用非中心点去代替中心点，计算代价函数，如果代价函数小于 0，则用非中心点代替中心点，一直循环，直到簇不发生变化。为了更好地描述 U-PAM 算法，本节首先给出定义 4、定义 5。

定义 4[13] 设 $O_1, O_2, O_3, \cdots, O_n$ 为 n 个 m 维不确定测量数据，$\overline{\varphi(O_i)}$ 表示第 i 个不确定测量数据的误差向量，则第 i 个不确定测量数据 O_i 落入区间 $[O_i - k\overline{\varphi(O_i)}, O_i + k\overline{\varphi(O_i)}]$ 上的概率为：

$$P(k=1) = 68.3\%$$

$$P(k=2) = 95.4\%$$

$$P(k=3) = 99.7\%$$

由此可知，可以用 k 值去控制不确定测量数据用区间数表示的精确度，一般取 $k=2$。

定义 5 设 X 和 Y 为两个不确定测量数据对象，X 表示为 $[a,b]$，Y 表示 $[c,d]$，$d(x,y)$ 为两者之间的距离，则

$$d(x,y) = \sqrt{(a-c)^2 + (b-d)^2}$$

根据定义 4 和定义 5，本节提出了 U-PAM 算法，该算法在不确定数据的函数分布或概率密度未知的情况下完成对数据的聚类，算法描述如下。

算法 1　U-PAM 算法

输入：n 个不确定测量数据 $O_1,O_2,O_3,\cdots,O_n$

输出：k 个最优聚类结果簇

begin：

1. 用 $Q_1,Q_2,Q_3,\cdots,Q_n$ 分别表示 $O_1,O_2,O_3,\cdots,O_n$，其中 $Q_i=[O_i-k\overline{\varphi(O_i)},O_i+k\overline{\varphi(O_i)}]$；
2. center ← select k from $\{Q_1,Q_2,Q_3,\cdots,Q_n\}$，得初始中心点 $\{w_1,w_2,w_3,\cdots,w_k\}$；
3. for each remain Object do{
4. comput $d(Q_j,w_i)$，$k<j<n$，$0<i<k$；/ * 剩余区间数到中心点的距离 * /
5. $w_i\leftarrow$Min $d(Q_j,w_i)$；/ * 将离中心点距离最近的对象分别分配给各中心点 * /
6. }；
7. 得到 k 个簇 $C=\{C_1,C_2,C_3,\cdots,C_k\}$；
8. do{
9. select w_i from center；/ * w_i 为一个中心点 * /
10. do{
11. select Q_j from C；/ * Q_j 为一个非中心点 * /
12. $w_i\leftarrow Q_j$；/ * 用非中心点代替中心点 * /
13. compute $S(O_j)$，$S(w_i)$；/ * $S(x_i)=\sum_{i=1}^{k}\sum_{x\in q}\|x_i-q_i\|^2$ 表示簇的误差平方和 * /
14. if($S(O_j)<S(w_i)$){
15. $w_i\leftarrow Q_j$；
16. return 3；
17. }/ * Q_j 是 w_i 的好的代替 * /
18. else if($S(O_j)>S(w_i)$)
19. return 11；
20. }while(all w_i selected)；/ * 所有的中心点被选择过，内层循环的终止 * /
21. while(all Q_j selected)；/ * 所有的非中心点都被选过，外层循环的终止 * /
22. 循环结束，最终得到不确定测量数据对象的新的 k 个簇 $\{C_1,C_2,C_3,\cdots,C_k\}$

end

在 U-PAM 算法中，首先将 n 个不确定测量数据对象用区间数的形式表示；其次从 n 个区间数中随机选出 k 个对象作为中心点，并计算剩余区间数到中心点的距离（$d(Q_j,w_i)$），且将离中心点最近的区间数分配给 k 个中心点（$w_i\leftarrow$ Min $d(Q_j,w_i)$）；再次用非中心点的区间数对象代替中心点并计算总代价，总代价用代替前、后的误差平方和的差表示，若总代价小于零，则找到了一个更好的替换，对中心点进行更新，进而得到 k 个簇。

该算法基于传统的 PAM 算法并结合区间数和标准差来对不确定测量数据进行聚类，解决了当不确定数据概率密度或分布函数缺失的条件下传统 UK-means 无法对其聚类的问题，而且还改善了对孤立点敏感的影响，该算法的时间复杂度与传统的 PAM 算法的相同，为 $O(n(n-k)^2)$，其中 n 为不确定测量数据的个数，k 为聚类簇的个数；但当对不确定数据聚类时，利用了区间数的一些性质，避免了大量的积分运算，有效降低了算法的时间复杂度。

4　UM-PAM 聚类算法

第 3 节提出了 U-PAM 算法，该算法可以在数据的分布函数或概率密度缺失的情况下对不确定测量数据进行有效的聚类，而且对孤立点不敏感。但是当处理海量不确定数据时，U-PAM 算法可能执行得很缓慢或根本就不执行，对此，本节提出了 UM-PAM 算法，该算法“继承”了 U-PAM 算法的优点，而且可以有效地处理海量不确定测量数据。

UM-PAM 算法中，为了有效处理海量不确定测量数据，首先对不确定测量数据对象随机抽样；对抽样数据随机选取 k 个中心点，将抽样数据的其他对象按照距离准则分配给 k 个中心点，得到 k 个初始簇，再在簇中用非代表对象去代替中心点，计算代价函数，如果代价函数小于 0，则用非中心点代替中心点，一直循环，直到簇不发生变化；得到抽样样本的 k 个聚类中心后，以抽样样本的聚类中心作为总体的聚类中心，对总体进行聚类；进而有效完成对海量不确定测量数据的聚类。为了较好地对算法进行描述，给出定义。

定义 6[14]　设不确定测量数据 $O_1,O_2,O_3,\cdots,O_n$，n 取值很大，如果对该数据进行随机抽样，抽样样本容量 S 满足以下公式时，抽样样本可以很好地表示总体的特征。

$$s=f\times n+\frac{n}{n_i}\times\log\left(\frac{1}{\delta}\right)+\frac{n}{n_i}\sqrt{\log\left(\frac{1}{\delta}\right)^2+2\times f\times n_i\times\log\left(\frac{1}{\delta}\right)}$$

其中，f 是抽取到指定数据的比例，$0\leqslant f\leqslant 1$，n 为数据规模，n_i 为簇 C_i 的规模。

算法 2　UM-PAM 算法

输入：n 个不确定测量数据 $O_1,O_2,O_3,\cdots,O_n$

输出：k 个最优聚类结果簇

begin：

1. 用 $Q_1,Q_2,Q_3,\cdots,Q_n$ 分别表示 $O_1,O_2,O_3,\cdots,O_n$，其中 $Q_i=[O_i-k\overline{\varphi(O_i)},O_i+k\overline{\varphi(O_i)}]$
2. M← Sample m from $Q_1,Q_2,Q_3,\cdots,Q_n$；/ * 对数据进行抽样，M 为含有 m 个对象，m 按照定义 6 计算 * /
3. get sample data $M=\{m_1,m_2,m_3,\cdots,m_m\}$；
4. center ← select k from $\{m_1,m_2,m_3,\cdots,m_m\}$，得样本初始中心点 $\{k_1,k_2,k_3,\cdots,k_k\}$；
5. for each remain Object of M do{
6. compute $d(m_j,k_i)$，$k<j<m$，$0<i<k$；/ * 抽样后剩余对象到中心点的距离 * /
7. $k_i\leftarrow$Min $d(m_j,k_i)$；/ * 将离中心点距离最近的对象分别分配给各中心点 * /
8. }
9. 得到 k 个簇 $C_1,C_2,C_3,\cdots,C_k$；
10. repeat{
11. select k_i from center/ * 选取一个中心点 * /
12. repeat {
13. select m_j from C/ * 选取一个非中心点 * /
14. $k_j\leftarrow$ represent(m_j)/
15. if($S(m_j)>S(k_i)$){
16. $k_j\leftarrow$ represent(m_j)；
17. return 5}；/ * 假如非中心点是中心点的一个好的替换，则替换 * /
18. }while(all m_j selected)；/ * 所有中心点被选择内层循环终止 * /
19. }while(all k_i selected)；/ * 所有非中心点被选择外层循环终止 * /
20. 得到 k 个更新后的 center=$\{k_1,k_2,k_3,\cdots,k_k\}$；/ * 抽样样本的中心点 * /
21. for each Object of Q_i do{
22. D←d(k_i,O_i) / * 计算非中心点到中心点的距离，存入集合 D 中 * /

23. k_i←Min d(k_i,O_i) of Object in D/*集合 D 中离中心点最近的点分配给该中心点*/

24. }

25. k←FinalClusterNumber(C_i);/*总体聚类得到 k 个簇*/

end

在 UM-PAM 算法中，首先对不确定测量数据用区间数的形式表示，其次对不确定测量数据进行随机抽样，在样本数据中随机选出 k 个对象作为中心点，并计算剩余区间数到中心点的距离($d(m_j,k_i)$)，且将离中心点最近的区间数分配给 k 个中心点(k_i←Min $d(m_j,k_i)$)；再次用非中心点的区间数对象代替中心点并计算总代价，总代价用代替前、后的误差平方和的差表示，若总代价小于零，则找到了一个更好的替换，对中心点进行更新，进而得到样本的 k 个聚类中心($center=\{k_1,k_2,k_3,\cdots,k_k\}$)，然后计算总体的对象离样本中心点的距离($d(k_i,O_i)$)，并将离中心点最近的点分配给聚类中心点($k_i$←Min $d(k_i,O_i)$)，进而完成对海量不确定测量数据的聚类。

UM-PAM 算法是对 U-PAM 算法的一种改进，旨在解决 U-PAM 算法难以处理海量不确定测量数据的问题。该算法“继承”了 U-PAM 算法的优点，即在概率密度或分布函数缺失的情况下对不确定测量数据进行聚类，且对孤立点不敏感，当不确定测量数据的数据量很大时仍可以有效地完成对不确定测量数据的聚类。

5 基于 U-PAM 和 CH 有效性指标优化

对于最佳聚类数 k 值的选取，目前主要的思想是针对预先已知的数据集，先确定聚类数 k 的搜索范围$[k_{\min},k_{\max}]$，从中选取出合适的 k 值，然后对样本数据集进行聚类，最后利用合适的有效性评价函数对聚类结果进行评测，通过评测结果找出最优聚类结果所对应的 k 值，并将其作为最优的聚类数 k 值，记为 k_{opt}。在实际情况中，当 $k_{\min}=1$ 时，表明样本均匀分布，没有明显特征差异。因此通常聚类数 k 最小为 2，即 $k_{\min}=2$[15]，同时最佳聚类数 k_{opt} 应该远小于样本对象个数，即 $k_{opt}\ll n$，这样要确定 k_{opt}，使得 $k_{opt}\leqslant k_{\max}$，接下来的计算就会简单一些。然而如何确定 $k_{\max}$ 目前尚没有明确的理论指导，大多数采用的方法是通过经验规则来确定聚类数搜索范围的上限 $k_{\max}$，即 $k_{\max}\leqslant\sqrt{n}$。文献[16]给出了确定 $k_{\max}$ 的一种方法，该方法证明了规则 $k_{\max}\leqslant\sqrt{n}$ 具有一定的理论依据，并在文献中验证了此方法的有效性。文献[17]提出了一种运用距离代价函数确定最优的聚类数 k 值，同时还给出了 k 值的最优解 k_{opt} 及其上界 k_{opt} 的条件，并证明了规则的合理性。因此本文在优化的过程中选取的最佳聚类数搜索范围为$[k_{\min}=2,k_{\max}=\sqrt{n}]$

5.1 聚类有效性指标函数

当确定了聚类数目 k 值的搜索范围之后，选择一个合适的聚类有效性指标函数很关键。聚类有效性是指对聚类结果进行评价以确定最适合特定数据集的划分和评判所得结果是否是有效的、正确的。常用的聚类有效性评价方法有外部评价法、内部评价法和相对评价法。外部评价法和内部评价法均基于统计测试，具有较高的计算复杂性，通过度量一个数据集与预先已知结构的相似程度进行判断。相对评价法寻求一个聚类算法在一定假设和参数下能定义的最好聚类结果。通过对不同 k 值下的聚类有效性指标的计算，将最优聚类结果对应的聚类数目作为最佳聚类数，从而最终确定有效的 k 值。传统 PAM 算法使用簇内误差平方和作为评价指标，对每个簇的误差平方进行求和，会导致将密度不均匀的较大类簇拆成两个簇的结果，从而导致局部最优，进而得不到正确的聚类数 k 值，最终影响了聚类的效果。因此本文选用评价指标 CH 来确定最佳聚类数 k。

定义 7 给定一个 p 维不确定测量数据集 $X=\{x_1,x_2,x_3,\cdots,x_n\}$，其中 $x=\{x^1,x^2,x^3,\cdots,x^p\}$ 为一个数据对象，n 为数据集 X 中数据对象的个数。一个聚类算法将 X 划分为 NC 个子集的集合$\{C_1,C_2,C_3,\cdots,C_{NC}\}$，子集 C_i 称为 X 的子类。c 表示数据集 X 的中心点，c_i 表示数据集 X 的 k 个聚类簇 C_i 的中心点，n_i 表示聚类簇 C_i 中数据对象的个数，$d(x_i,x_j)$表示数据对象间的期望距离。其中 $i=\{1,2,3,\cdots,k\}$。

CH(Calinski-Harabasz)指标定义如下：

$$CH(NC)=\frac{\frac{1}{NC-1}\sum_{i=1}^{NC}n_i d^2(c_i,c)}{\frac{1}{n-NC}\sum_{i=1}^{NC}\sum_{x\in C_i}d^2(x,c_i)}$$

其中，$\frac{1}{NC-1}\sum_{i=1}^{NC}n_i d^2(c_i,c)$表示聚类簇中各个类的中心与该类的中心的距离的平方和，记为紧密度(Tightness)，$\frac{1}{n-NC}\sum_{i=1}^{NC}\sum_{x\in C_i}d^2(x,c_i)$表示聚类簇各类中心点与数据集的中心点的距离平方和，记为分离度(Resolution)，CH 指标由分离度与紧密度的比值得到。因此，CH 越大代表着类自身越紧密，类与类之间越分散，即聚类结果更优。

5.2 基于 U-PAM 和 CH 有效性指标的 k 值优化

本节针对 U-PAM 聚类算法对不确定测量数据进行聚类时必须预先给定聚类数 k 值的问题，提出了基于 U-PAM 聚类算法和 CH 聚类有效性指标的 k 值优化算法，其基本思想如下：在最佳聚类数 k 值的搜索范围$[k_{\min},k_{\max}]$内，针对不同的 k 值，首先分别对不确定数据进行 U-PAM 聚类，接下来评估聚类结果，计算每次聚类结果的 CH 有效性指标，最后对聚类结果进行分析，根据 CH 聚类指标得到最佳聚类数 k_{opt} 值。算法描述如下。

算法 3 U-PAM 算法结合 CH 指标函数确定最佳聚类数算法

输入：n 个不确定测量数据 $O_1,O_2,O_3,\cdots,O_n$

输出：CH 指标最大时所对应的最佳聚类数 k_{opt}

begin:

1. 取 $k_{\min}=2,k_{\max}=\sqrt{n}$;
2. for $k_{\min}$ to $k_{\max}$ do{
3. 用 $Q_1,Q_2,Q_3,\cdots,Q_n$ 分别表示 $O_1,O_2,O_3,\cdots,O_n$，其中 $Q_i=[O_i-k\overline{\varphi(O_i)},O_i+k\overline{\varphi(O_i)}]$
4. center ← select k from$\{Q_1,Q_2,Q_3,\cdots,Q_n\}$，得初始中心点$\{w_1,w_2,w_3,\cdots,w_k\}$
5. compute d(Q_j,w_i),k<j<n,0<i<k;/*剩余区间数到中心点的距离*/
6. w_i←Mind(Q_j,w_i);/*将离中心点距离最近的对象分别分配给各中心点*/}
7. 得到 k 个簇 $C=\{C_1,C_2,C_3,\cdots,C_k\}$;
8. do{
9. select w_i from center/* w_i 为一个中心点*/
10. do{

```
11.    select Qj from C/* Qj 为一个非中心点*/
12.    wi← represent(Qj)
13.    comput S(Oj),S(wi);/*S=∑_{i=1}^{k} ∑_{x∈q} ‖xi−qi‖²
14.    if(S(Oj)<S(wi)){
15.      wi← represent(Qj)
16.      return 3
17.      }/*Qj 是 wi 的好的代替*/
18.    else if(S(Oj)>S(wi))
19.      return 11;
20.  }while(all wi selected);/* 内层循环的终止条件,即所有的中心点被选择过
21.  }while(all Qj selected);/*外层循环的终止条件
22. 循环结束,最终得到不确定测量数据对象的新的 k 个簇{C1,C2,C3,…,Ck};
23.  M ← compute CH;/*根据定义 6 计算平均 CH 聚类指标值,并记录*/}
24. kopt←max(M);/*比较平均 CH 指标值,将平均 CH 指标值最大所对应的 k 值即是最佳聚类数*/
25. output kopt;
end
```

针对 U-PAM 算法对不确定数据进行聚类时必须事先确定 k 值的问题，本节提出了基于 U-PAM 聚类算法结合 CH 聚类有效性指标函数确定最佳聚类数的算法，CH 聚类有效性指标通过分离度与紧密度来反映聚类结果中类内各点与聚类中心点的紧密性以及各类之间的分离程度，由此来判断聚类结果的准确性，从而得到期望的聚类数 k，以解决 k 值未知而需要预先指定的问题。

6 实验

本文详细研究了在不确定测量数据的概率密度或分布函数未知时基于 PAM 聚类算法的 U-PAM、UM-PAM 算法，此外还提出了基于 U-PAM 算法结合 CH 有效性指标函数优化聚类数目的算法，本节在配置为 2.60Hz Intel i5 4210M CPU 和 8GB 内存的 PC 机上进行实验。操作系统为 Windows 8，程序用 C++、R 语言编写。

为了验证本文提出的 U-PAM 算法的有效性，本节采用 UCI 数据集中的鸢尾花数据集 Iris，该数据集包含 4 个属性，共有 150 个数据对象，可以分为 3 类。为了构造数据的不确定性，分别在数据的每个维度添加一个均值为 0、方差为 1 的噪声。用 U-PAM 算法对该数据集进行聚类，聚类的效果如图 1 所示。

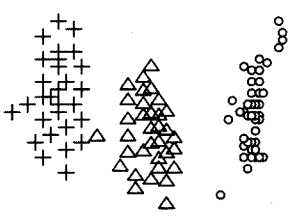


图 1 U-PAM 算法对 Iris 数据集的聚类图

由图 1 可知，用 U-PAM 算法对上述数据集进行聚类，分别用圆形、三角形和十字形代表了鸢尾花的 3 种类别，图中显示不同类别的相似度很低，同一类别对象的相似度很高，而且数据对象之间几乎没有孤立点，由此可知 U-PAM 算法在不确定测量数据的概率密度或分布函数未知的情况下的聚类效果明显而且对孤立点不敏感。

为了验证本文提出的 U-PAM 算法有效性，将其与文献[18]中所提出的 VDBiP 算法进行比较，所采用的测试数据集是 UCI 中的 Iris、Glass、Wine。且为了表示测量数据的不确定性，分别在测试数据集 UCI 中的 Iris、Glass、Wine 的每一维度添加了 3 种不同分布的噪声，分别为 normal、uniform、binormial 分布。利用本文提出的 U-PAM 算法和文献[18]中提出的 VDBiP 算法对以上数据集进行聚类。实验结果如表 1 所列。

表 1 U-PAM 算法与 VDBiP 算法的聚类结果比较

数据集	噪声分布	聚类精度(%)	
		VDBiP 算法	U-PAM 算法
Iris	normal	80.14	84.97
	uniform	75.78	80.36
	binomial	79.65	85.44
Glass	normal	81.32	82.11
	uniform	76.28	80.43
	binomial	81.27	86.54
Wine	normal	83.25	85.14
	uniform	74.09	79.87
	binomial	78.33	80.77

由表 1 可知，U-PAM 算法和 VDBiP 算法分别对 3 种不确定数据集进行聚类时，U-PAM 算法在 3 种数据集上的聚类精度优于 VDBiP 算法，说明 U-PAM 算法具有很高的聚类精度；U-PAM 算法采用区间数和标准差来表示数据的不确定性，将其减少了大量积分运算，提高了算法的效率。

为了验证本文提出的 UM-PAM 算法在处理海量不确定测量数据的有效性，将其与文献[18]中提出的 VDBiP 算法和文献[19]中所提的 M-FDBSCAN 算法进行比较，所采用的数据集是人工合成数据集 DataSet1-8，聚类结果如表 2 和图 2 所示。

表 2 UM-PAM 算法与 VDBiP 算法比较

数据集	样本数目	VDBiP 算法	UM-PAM 算法
		准确率(%)	准确率(%)
DataSet1	1000	81.21	85.89
DataSet2	5000	76.78	89.16
DataSet3	10000	79.66	80.45
DataSet4	15000	75.95	81.59
DataSet5	20000	82.91	87.05
DataSet6	25000	76.98	80.21
DataSet7	30000	78.67	83.10
DataSet8	35000	80.25	82.47

由表 2 可知，UM-PAM 算法和 VDBiP 算法分别对以上人工数据集进行聚类，在不同样本数目下，UM-PAM 算法的准确率都稍高于 VDBiP 算法，由此可知 UM-PAM 聚类算法具有很高的聚类精度。

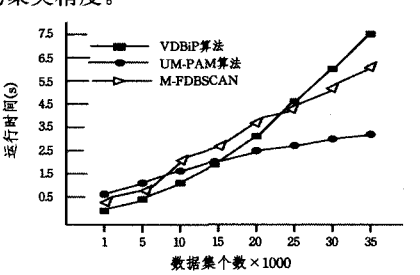


图 2 UM-PAM 算法、VDBiP 算法和 M-FDBSCAN 算法的运行时间与数据集数目的关系

由图 2 所示，当不确定测量数据的数据量较小时，UM-PAM 算法比 VDBiP 和 M-FDBSCAN 算法的运行时间稍长，这是因为 UM-PAM 算法有一个抽样的过程，随着不确定数据量的不断增加，VDBiP 算法和 M-FDBSCAN 算法的运行时间大致呈指数递增，而 UM-PAM 算法却缓慢地增加；由此看出，当不确定测量数据的数目较大时，UM-PAM 算法优于 VDBiP 算法和 M-FDBSCAN 算法。

为了验证基于 U-PAM 算法和 CH 有效性指标对 k 值优化的有效性，采用人工数据集 DS1 和 DS2，其中 DS1 是维度为 2、类别数为 3、样本为 48 的数据集，DS2 是维度 2、类别数为 4、样本数为 62 的数据集。

根据数据集中的样本数据数确定聚类数 k 的搜索范围，数据集 DS1 的聚类数 k 的范围为$[2,\sqrt{48}]$，数据集 DS2 的聚类数 k 的范围为$[2,\sqrt{62}]$。基于 U-PAM 算法分别对数据集 DS1、DS2 进行聚类，其聚类结果与其 CH 有效性指标值的关系如图 3、图 4 所示。样本数据的 CH 有效性指标值越大代表着类内数据越紧密，类与类之间越分散，聚类效果越好。因此，当 CH 有效性指标值的平均值最大时，聚类结果所对应的聚类数即为最佳聚类数 k_{opt}。通过图 3 可以看出，当聚类数 $k=3$ 时，CH 有效性函数指标值的平均值为 $avg_{CH}(3)=0.612$，此时 CH 有效性指标的平均值最大，由此得到最佳聚类数 $k_{opt}=3$，同时 DS1 的类别数为 3，所以数据集 DS1 得到正确的聚类结果。同样地，通过图 4 可以看出，数据集 DS2 的聚类结果中，当聚类数 $k=4$ 时，$avg_{CH}(4)=0.635$ 为最大 CH 有效性指标平均值，所以数据集 DS2 的最佳聚类数 $k_{opt}=4$，可知数据集 DS2 也得到了正确的聚类结果。

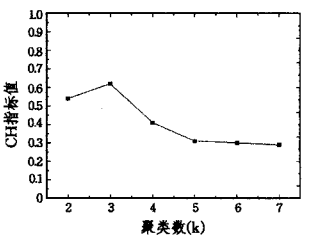


图 3　基于 U-PAM 算法的数据集 DS1 聚类数 k 与 CH 的关系

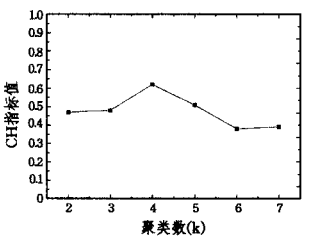
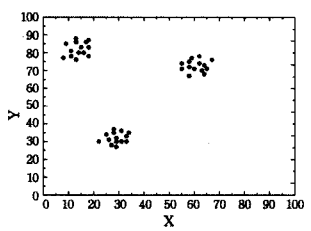


图 4　基于 U-PAM 算法的数据集 DS2 聚类数 k 与 CH 的关系

根据上述实验分析得出的最佳聚类数 k_{opt}，分别对数据集 DS1、DS2 进行 U-PAM 聚类的聚类结果如图 5、图 6 所示。



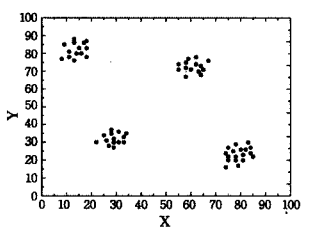


图 5　数据集 DS1 的聚类结果图　图 6　数据集 DS2 的聚类结果图

综上实验分析所述，基于 U-PAM 和 CH 有效性指标函数可以得到正确的聚类数 k，并且有效地提高了聚类的效率。

结束语　针对概率密度函数或者概率分布函数等信息在很多应用场合中很难得到，并且采用该信息来表示不确定性会带来聚类算法计算复杂度较高的问题，本文利用区间数和标准差来表示测量数据属性的不确定性，并通过设计新的不确定性数据间距离计算方法，提出了多维不确定性数据聚类算法 U-PAM，并且用实验验证了其有效性，通过与其他划分聚类算法做比较分析，表明本文提出的 U-PAM 算法有很高的聚类精度。当不确定测量数据的数据量很大时，针对 U-PAM 算法执行很缓慢的问题，提出了 UM-PAM 算法，该算法先对数据进行抽样，对抽样数据进行局部聚类，然后再总体聚类，实验证明其显著提升了聚类的时效性。针对划分算法的 k 值必须事先确定的问题，基于 U-PAM 算法和 CH 聚类指标函数相结合的方法，对聚类结果进行分析，最终确定最佳聚类数。

本文今后的研究重点在于对任意形状的不确定测量数据进行有效的聚类的问题。

参 考 文 献

[1] Xing Chang-zheng, Wen Pei. Uncertain data streams clustering algorithm based on grid density and force[J]. Applicaton Research of Computer, 2015, 32(1): 98-101(in Chinese)
邢长征，温培. 基于网格密度和引力的不确定数据流聚类算法[J]. 计算机应用研究，2015，32(1)：98-101

[2] Zhou Tao, Lu Hui-ling. Clustering algorithm research advances on data mining[J]. Computer Engineering and Applications, 2012, 48(12): 100-111(in Chinese)
周涛，陆惠玲. 数据挖掘中聚类算法研究进展[J]. 计算机工程与应用，2012，48(12)：100-111

[3] Sun J G, Liu J, Zhao L Y. Clustering algorithms research[J]. Journal of Software, 2008, 19(1): 48-61(in Chinese)
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[M]// Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2006: 199-204

[5] Lee S D, Kao B, Cheng R. Reducing UK-means to K-means [C]// Seventh IEEE International Conference on Data Mining Workshops, 2007. ICDM Workshops 2007. IEEE, 2007: 483-488

[6] Peng Yu, Luo Qing-hua, Peng Xi-yuan. A multi-dimensional unertain measurement data clustering algorithm[J]. Chinese Journal of Scientific Instrument, 2011, 32(6): 1201-1207(in Chinese)
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

[7] Ren Pei-hua, Wang Li-zhen. Improved K-means Clustering Algorithm Based on DKC in Uncertain Region Environment[J]. Computer Science, 2013, 40(4): 181-184(in Chinese)
任培花，王丽珍. 不确定域环境下基于 DKC 值改进的 K-means 聚类算法[J]. 计算机科学，2013，40(4)：181-184

[8] Kao B, Lee S D, Lee F K F, et al. Clustering uncertain data using voronoi diagrams and r-tree index[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(9): 1219-1233

[9] H Jian, S Shu-bin, M Yi-min, et al. High Dimensional Uncertain Data Efficient Clustering Algorithm[J]. Computer Knowledge & Technology, 2014(4)

[10] Gullo F, Ponti G, Tagarelli A. Clustering uncertain data via k-medoids[M]// Scalable Uncertainty Management. Springer Berlin Heidelberg, 2008: 229-242

[11] Xie Xiao-lu, Li Lei. Research on Multi-attribute Group Decision Under Interval Number Information[J]. Computer Engineering, 2014, 40(10): 210-213(in Chinese)
谢小璐,李磊. 区间数信息下的多属性群决策研究[J]. 计算机工程,2014,40(10):210-213

[12] Reynolds P A, Richards G J, Rayward-smith V. The Application of K-Medoids and PAM to the Clustering of Rules[J]. Lecture Notes in Computer Science, 2004, 3177: 173-178

[13] Aggarwal C C. Yu P S. A survey of uncertain dataalgorithms and applieations[J]. IEEE Transactions Onknowledge and Data Engineering, 2009, 21(5): 609-623

[14] Lu Zhi-mao, Feng Jin-gong, Fan Dong-mei, et al. New clustering algorithms for large data processing[J]. System Engineering and Electronics, 2014(5): 1010-1015(in Chinese)
卢志茂,冯进玫,范冬梅,等. 面向大数据处理的划分聚类新方法[J]. 系统工程与电子技术,2014(5):1010-1015

[15] Zhou Shi-bing, Xu Zhen-yuan, Tang Xu-qing. New method for determining optimal number of clusters in K-means clustering algorithm[J]. Computer Engineering and Applications, 2010, 46(16): 27-31(in Chinese)
周世兵,徐振源,唐旭清. 新的 K-均值算法最佳聚类数确定方法[J]. 计算机工程与应用,2010,46(16):27-31

[16] Yu Jian, Cheng Qian-sheng. Search range of the Optimal clustering number in fuzzy clustering algorithms[J]. Science in China: Series E, 2002, 32(2): 274-280(in Chinese)
于剑,程乾生. 模糊聚类方法中的最佳聚类数的搜索范围[J]. 中国科学:E辑,2002,32(2):274-280

[17] Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset[J]. Genome Biology, 2002, 3(7): 1-21

[18] Kao B, Lee S, Lee F, et al. Clustering Uncertain Data Using Voronoi Diagrams and R-Tree Index. [J]. Knowledge & Data Engineering IEEE Transactions on, 2010, 22(9): 1219-1233

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

(上接第 217 页)

[7] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space [C]// ICLR 2013. 2013

[8] Joachims T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization [M]. Springer US, 1997: 143-151

[9] Hinton G E. Learning distributed representations of concepts [C]// Proceedings of CogSci. 1986: 1-12

[10] Socher R, Bauer J, Manning C D, et al. Parsing with Compositional Vector Grammars [C]// Meeting of the Association for Computational Linguistics. 2013: 455-465

[11] Socher R, Perelygin A, Wu J Y, et al. Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2013: 1631-1642

[12] Sun Y, Lin L, Yang N, et al. Radical-Enhanced Chinese Character Embedding [J]. Lecture Notes in Computer Science, 2014, 8835: 279-286

[13] Mansur M, Pei W, Chang B. Feature-based Neural Language Model and Chinese Word Segmentation [C]// IJCNLP. 2013: 1271-1277

[14] Zheng X, Chen H, Xu T. Deep Learning for Chinese Word Segmentation and POS Tagging [C]// EMNLP. 2013: 647-657

[15] Tang D, Wei F, Yang N, et al. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification [C]// ACL. 2014: 1555-1565

[16] Zhang M, Zhang Y, Che W, et al. Chinese Parsing Exploiting Characters [C]// ACL. 2013: 125-134

[17] Xing C, Wang D, Zhang X, et al. Document Classification with Distributions of Word Vectors [C]// 2014 Annual Summit and Conference Asia-Pacific Signal and Information Processing Association(APSIPA). IEEE, 2014: 1-5

[18] Kim H K, Kim H, Cho S. Bag-of-Concepts: Comprehending Document Representation through Clustering Words in Distributed Representation[OL]. http://dm. snu. ac. kr/static/docs/TR/SNUDM-TR-2015-05. pdf

[19] Le Q V, Mikolov T. Distributed Representations of Sentences and Documents [J]. Eprint Arxiv, 2014, 4: 1188-1196

[20] Morin F, Bengio Y. Hierarchical Probabilistic Neural Network Language Model [J]. Aistats. 2005, 5: 246-252

[21] Mnih A, Hinton G E. A Scalable Hierarchical Distributed Language Model [C]// Advances in Neural Information Processing Systems. 2009: 1081-1088

[22] Mikolov T, Yih W, Zweig G. Linguistic Regularities in Continuous Space Word Representations [C]// HLT-NAACL. 2013: 746-751

[23] Santana L E A, De Oliveira D F, Canuto A M P, et al. A Comparative Analysis of Feature Selection Mmethods for Ensembles with Different Combination Methods [C]// International Joint Conference on Neural Networks, 2007 (IJCNN 2007). IEEE, 2007: 643-648

[24] Forman G. An Extensive Empirical Study of Feature Selection Metrics for Text Classification [J]. The Journal of Machine Learning Research, 2003, 3: 1289-1305

[25] 搜狗. 文本分类语料库[OL]. http://www. sogou. com/labs/dl/c. html

[26] Gensim. Topic Modelling for Humans [OL]. http://radimrehurek. com/gensim

[27] Bengio Y, Schwenk H, Senécal J S, et al. Neural Probabilistic Language Models [M]// Innovations in Machine Learning. Springer Berlin Heidelberg, 2006

[28] Mnih A, Hinton G. Three New Graphical Models for Statistical Language Modelling [C]// Proceedings of the 24th International Conference on Machine Learning. ACM, 2007: 641-648