DOI:10.13203/j.whugis20150760 文章编号:1671-8860(2019)01-0153-06

数据质量聚类算法

李 延[1] 王大魁[2] 耿 晶[1] 王树良[1]

1 北京理工大学软件学院,北京,100081
2 中国科学院信息工程研究所,北京,100093

摘 要:在聚类算法中,聚类中心决定聚类的最终结果,而传统的分割聚类算法不能准确定位聚类中心。根据数据场提出了数据质量聚类中心的新概念,给出数据质量聚类算法,能够一次定位聚类中心,无需迭代,也无需预置聚类个数。7 组对比实验表明,提出的方法能够准确定位聚类中心,获得良好的聚类结果和稳定性,优于传统的分割聚类算法和峰值密度聚类算法。

关键词:数据场;聚类;数据质量;聚类中心

中图分类号:P208 **文献标志码**:A

聚类是数据挖掘的基础技术,有广泛的应用前景[1-2]。聚类算法主要分为层次聚类法、网格聚类法、分割聚类法和密度聚类法[3]。其中,分割聚类法简单、快速,广泛应用于各个领域,典型的分割聚类法是 K-means 算法和 K-medoids 算法。在实际应用中,这两种算法由于需要用户输入聚类个数,聚类结果与初始点选择有关等缺点,不能很好地满足用户的需要[4]。《Science》中提出的峰值密度聚类算法虽然解决了上述问题,但存在阈值需要人为输入的问题[5]。

本文根据数据场,提出了数据质量聚类中心的概念。数据场将物质粒子间的相互作用及场描述方法引入到抽象的数域空间,实现数据对象或者样本点间相互作用的形式化描述[6]和计算。数据场将数据所具有的固有属性定义为数据的质量,并根据实际挖掘视角的不同,表示数据不同的属性。本文中,数据质量将代表数据的密集程度,并以此确定聚类中心,该方法无需用户输入聚类个数,也无需选择初始点,更无需人为设定阈值。

1 数据质量聚类

在物理场中,物体的质量是不能改变的,是物体固有的属性。同理,在数据场中,数据的质量也代表了每个数据自身的固有属性。所不同的是,在数据场中,数据并不是实际存在的物体,可以这样认为,n 维数据集构成了一个 n 维的数据空间,数据集中每一个数据就是存在于这个 n 维空间中的“物体”,其各种属性都遵从于这个 n 维空间自身的特点。

定义:设数据集 α 含有 N 个数据点,$\alpha=\{x_1,x_2\cdots x_n\}$,其中 $x_i=\{x_{i1},x_{i2}\cdots x_{ip}\}$,组成一个 P 维空间 Ω,在空间 Ω 中的数据点 x_i 所固有的属性 τ,称之为点 x_i 在数据集 α 中的数据质量。

需要注意的是,定义中数据质量代表的是数据在数据集中的固有属性,这个固有属性会随着数据挖掘视角的不同而改变。一个数据点在数据集中可能会具有多种不同的固有属性,应当根据当前的挖掘任务赋予数据相应的属性。因此,数据场中数据质量具有集群性,即只在数据集中具有质量;空间唯一性,即相关的属性只在对应的数据集中存在;可变性,即根据需求不同代表的数据属性也不同。

聚类算法的目的是让类内相似度最高,类间相似度最低。反映在数据集的空间分布上,就是相似度高的数据分布在同一个类簇中,不同的类簇代表了不同的类别。因此,在聚类分析中,一般取数据密集程度这一属性作为数据的质量。此时,数据场中的数据质量本质上是反映数据集中数据的密集程度,处于密集区域的数据具有较大

收稿日期:2017-05-26

项目资助:国家自然科学基金(61472039);高等学校博士学科点专项科研基金(20121101110036)。

第一作者:李延,博士生,主要从事数据挖掘方面的研究。liy_007@126.com

通讯作者:王树良,博士,教授。slwang2011@bit.edu.cn

的数据质量，处于稀疏区域的数据具有较小的数据质量。

图1所示的红色点标出的是数据集中质量较大的点，与所描述的数据质量概念一致，这些点都处于数据集中的密集区域。在聚类分析中，处于密集区域的点都有可能成为聚类中心。图1中所示的数据集含有5 000个点，而质量较大的点约有1 000个，显然，只根据数据的质量不能确定数据集的聚类中心。

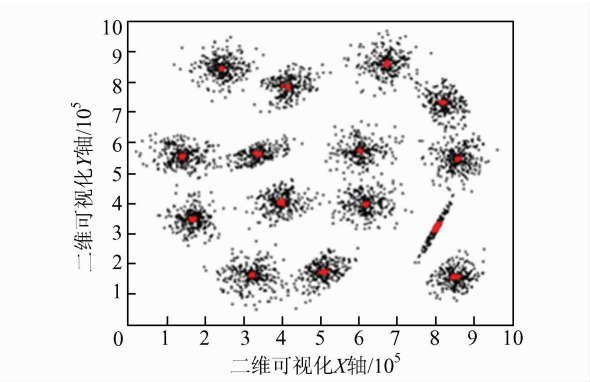

图1 具有较大数据质量的点

Fig. 1 Points with Big Mass

类比于物理场中的引力，聚类中心应当具有较大的质量，能够吸引其他质量较小的点在其周围形成一个类簇。同时，各个聚类中心应当相距较远，从而使聚类中心之间的作用力很小，直至可以忽略，这样，类簇与类簇间的相互关系就很弱，而类簇内的相互关系就很强，满足了最基本的聚类思想。

因此，数据质量聚类算法使用数据质量和数据之间的距离两个属性共同确定一个聚类中心。其中，数据之间的距离属性定义为：在数据集$\{x_1, x_2 \cdots x_n\}$中，所有比x_i质量大的点到x_i距离的最小值；如果点x_i是数据集中质量最大的点，那么其距离属性就为数据集中其他点$x_j(j \neq i)$到x_i距离的最大值。

数据距离属性的计算式为：

$$\delta_i = \begin{cases} \min\limits_{j:m_j > m_i} (d_{ij}), & \exists\ m_i < m_j \\ \max\limits_{j=1,2,\cdots,n} (d_{ij}), & \nexists\ m_i < m_j \end{cases} \tag{1}$$

式中，m表示数据的质量，d_{ij}表示两点间的距离。当数据集$\{x_1, x_2 \cdots x_n\}$中存在比x_i数据质量大的点x_j，即$m_i < m_j$时，数据之间的距离为所有比x_i质量大的点到x_i距离的最小值；如果不存在比x_i数据质量大的点x_j，即x_i是数据集中质量最大的点，那么其距离属性就为数据集中其他点$x_j(j \neq i)$到x_i距离的最大值。所以点x_i的m_i和δ_i都较大时，可以确定是聚类中心。在实际操作中，为了便于准确找到数据集中同时具有较大数据质量和较大距离属性的点，用数据集中每个数据点的质量属性作为横坐标、距离属性作为纵坐标绘制的决策图来确定聚类中心。在决策图中，同时具有较大横坐标和纵坐标数值的点会脱离其他只具有1个较大属性的点或者不具有较大属性的点，从而可以将这些脱离出来的点作为聚类中心。

图2所示为数据集的决策图，可以发现，只有少数几个点的两个属性都较大，这些点用红色标出，作为备选聚类中心。

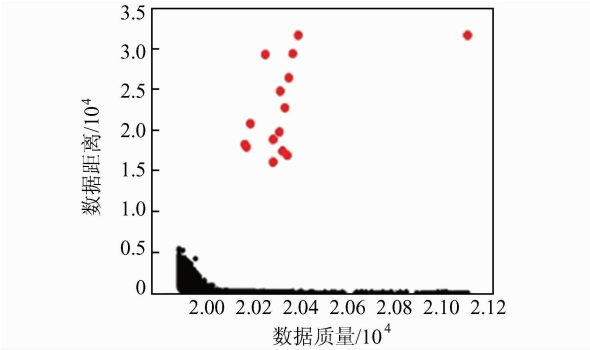

图2 聚类中心

Fig. 2 Clustering Centers

2 数据质量聚类算法实验验证

2.1 实验数据

数据质量聚类算法的核心是确定聚类中心，涉及数据的质量和距离两个属性。其中，距离属性计算使用欧氏距离，质量的计算采用参考文献[7]中的方法。在确定聚类中心后，先进行数据类别的划分，即将剩余点划入与其最近的聚类中心，形成一个个类簇，然后根据用户需要输出聚类结果。算法流程如图3所示。

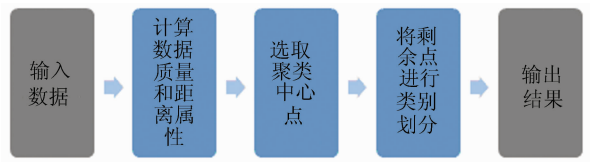

图3 算法流程图

Fig. 3 Algorithm Flow

通过一系列的对比实验验证数据质量聚类算法的聚类效果，并与传统的K-means算法、K-medoids算法和文献[1]中的峰值密度聚类算法进行了对比。

在对比实验中，采用7个数据集进行实验。数据集$A1$、$A2$、$A3$分别含有3 000个点和20个

类簇、5 250 个点和 35 个类簇、7 500 个点和 50 个类簇,并且 3 个数据集中类簇内点的个数均为 150 个。数据集 $S1$、$S2$、$S3$、$S4$ 都含有 5 000 个点和 15 个类簇,但是每个数据集中类簇的扩展程度不一样,而且 4 个数据集中每个类簇的中心是已知的[8]。这 7 个数据集的二维可视图如图 4 和图 5所示,图 4 和图 5 中的横、纵坐标分别为数据集二维可视图的 X 轴和 Y 轴。

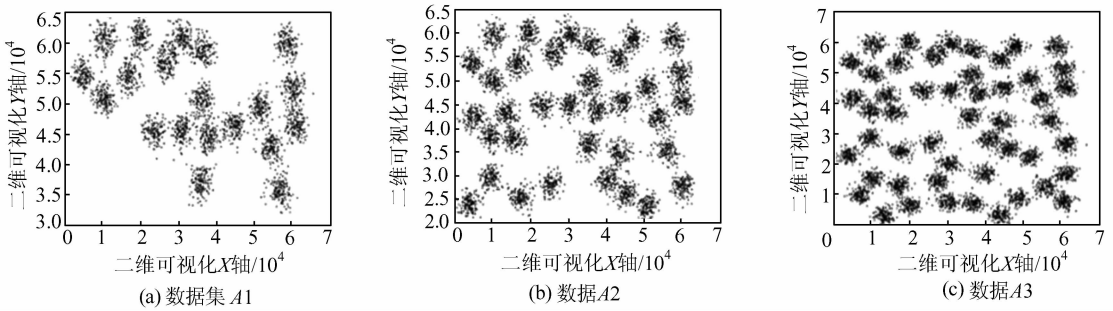

图 4　数据集 $A1$、$A2$、$A3$

Fig. 4　Datasets of $A1$, $A2$, $A3$

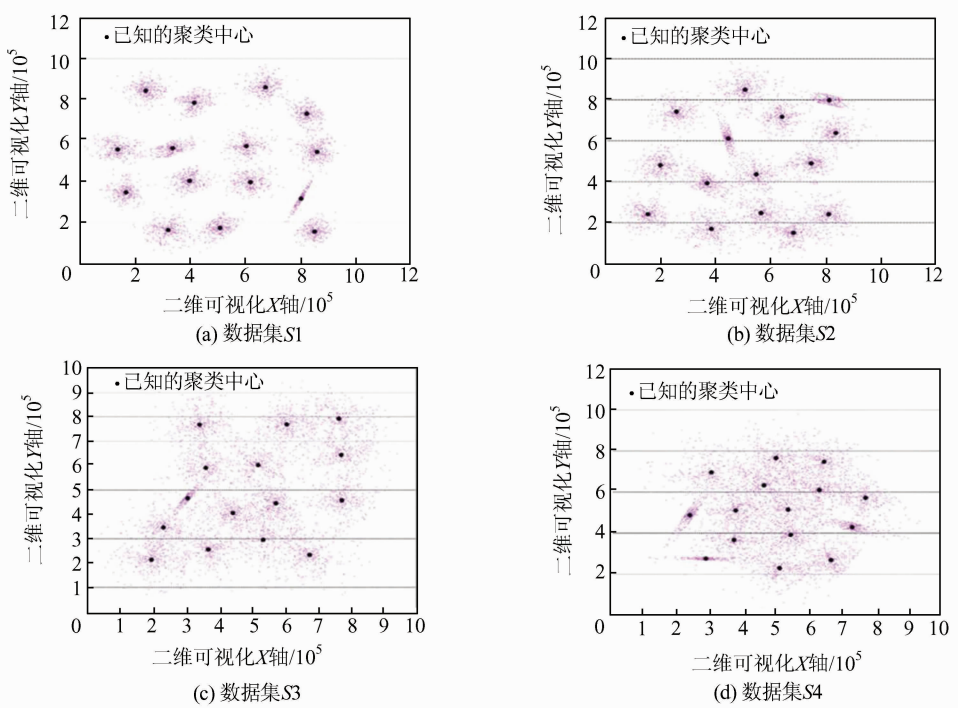

图 5　聚类中心数据集 $S1$、$S2$、$S3$、$S4$

Fig. 5　Clustering Centers Datasets of $S1$, $S2$, $S3$, $S4$

2.2　对比实验

首先对数据集 $A1$, $A2$, $A3$ 分别使用数据质量聚类算法和 K-means 算法、K-medoids 算法和峰值密度聚类算法进行聚类。将得到的聚类结果进行二维可视化展示,同时,对每个数据集中聚类结果进行统计,记录每种算法在每个类簇中聚集的点个数,与数据集实际每个类簇中应有点的个数进行对比,计算出准确率。

因 K-means 算法和 K-medoids 算法需要输入聚类个数,故按照数据集实际情况输入。数据质量聚类算法使用决策图确定聚类中心,如图 6 所示为数据集 $A1$、$A2$ 和 $A3$ 通过决策图选出的聚类中心。图 6 中彩色点为聚类中心,即横坐标和纵坐标都较大的点。所选出的聚类中心个数在数据集 $A1$ 中为 20,在 $A2$ 中为 35,在 $A3$ 中为 50,这与数据集原有的类簇个数相同。

图 7 是 4 种聚类算法的结果图,从图 7 中可以发现,数据质量聚类算法和峰值密度聚类算法都有较好的聚类效果。对于聚类算法的准确率统计每一个数据集中 4 种算法对每一个类簇聚类的准确率,即类簇内点的个数和实际每个类内点的个数比值。统计结果如表 1 所示。

从表 1 的统计结果中可以发现,数据质量的聚类算法具有最高的平均准确率,相比于传统的

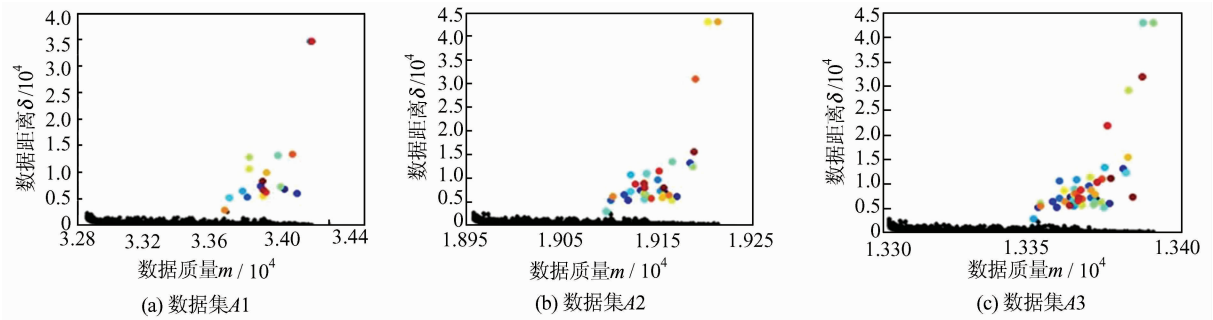

图6 数据集 $A1$、$A2$、$A3$ 的聚类中心

Fig. 6 Clustering Centers Datasets of $A1$, $A2$, $A3$

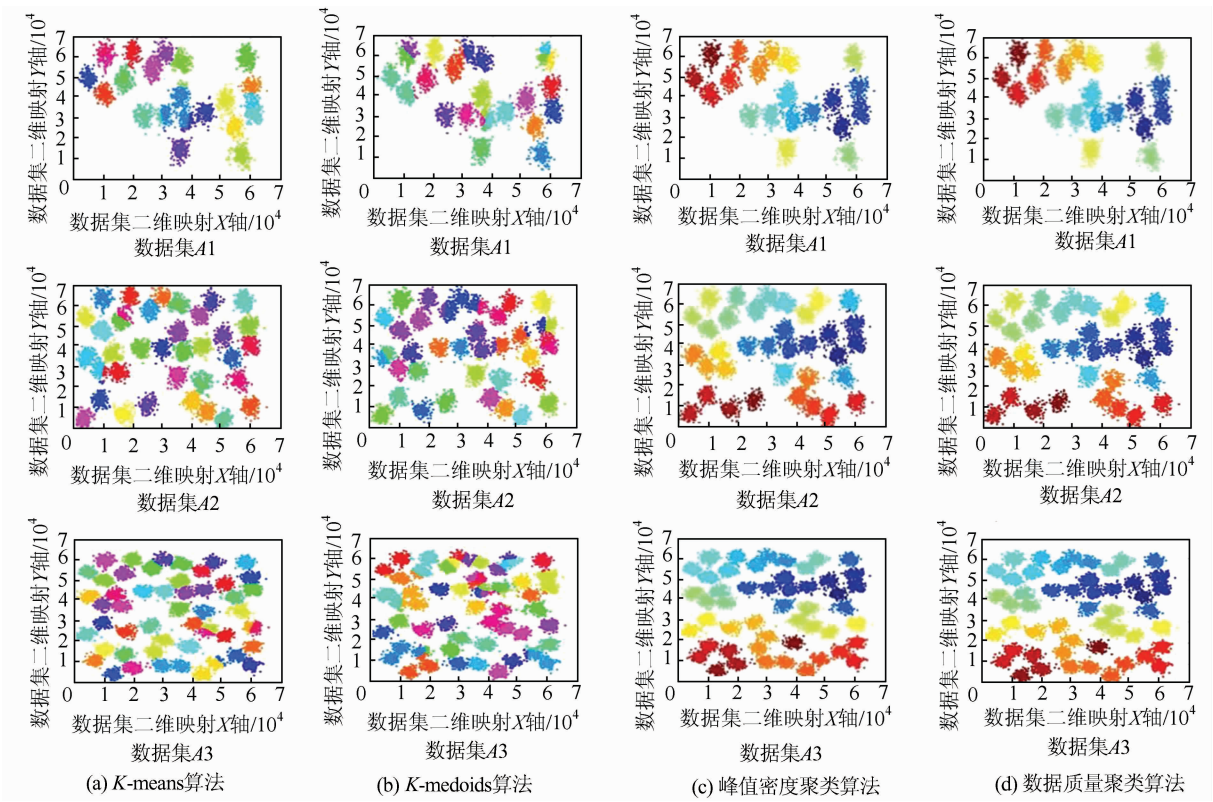

图7 数据集 $A1$、$A2$、$A3$ 聚类结果比较

Fig. 7 Comparison of Clustering Results on Datasets $A1$, $A2$, $A3$

表1 数据集 A1、A2、A3 实验平均准确率统计表/%

Tab. 1 Clustering Accuracies of Datasets $A1$, $A2$, $A3$/%

数据集	K-means 算法	K-medoids 算法	峰值密度聚类	数据质量聚类
$A1$	86.87	70.33	95.33	96.00
$A2$	76.84	79.73	96.65	96.91
$A3$	79.81	61.17	96.17	97.49

K-mean 算法和 K-medoids 算法分割聚类算法，在准确率上提高了很多，同时，与最新的峰值密度聚类算法相比，准确率也有所提高。

在数据集 $S1$、$S2$、$S3$、$S4$ 中，每个类簇的中心是已知的，通过比较 4 种算法得到的聚类中心与实际中心的偏差量，对比每种算法确定聚类中心的效果。使用决策图确定数据质量聚类算法的聚类中心。K-means 算法与 K-medoids 算法依然输入真实的类簇个数，4 种算法聚类结果二维可视图如图 8 所示。

在图 8 中，数据质量聚类算法和峰值密度聚类算法的聚类效果直观上要优于 K-means 算法和 K-medoids 算法。在对比聚类效果后，统计 4 种聚类算法所确定的聚类中心与实际中心位置的误差率。具体计算式为：

$$\gamma_i = \frac{1}{2}\left(\frac{x_i - a_i}{a_i} + \frac{y_i - b_i}{b_i}\right) \tag{2}$$

式中，x_i 和 y_i 为实验中得到的聚类中心的坐标；a_i 和 b_i 为数据集类簇实际的坐标。γ_i 值越小，说明越接近实际的类簇中心。每个数据集中的平均误差率统计结果如表 2 所示。

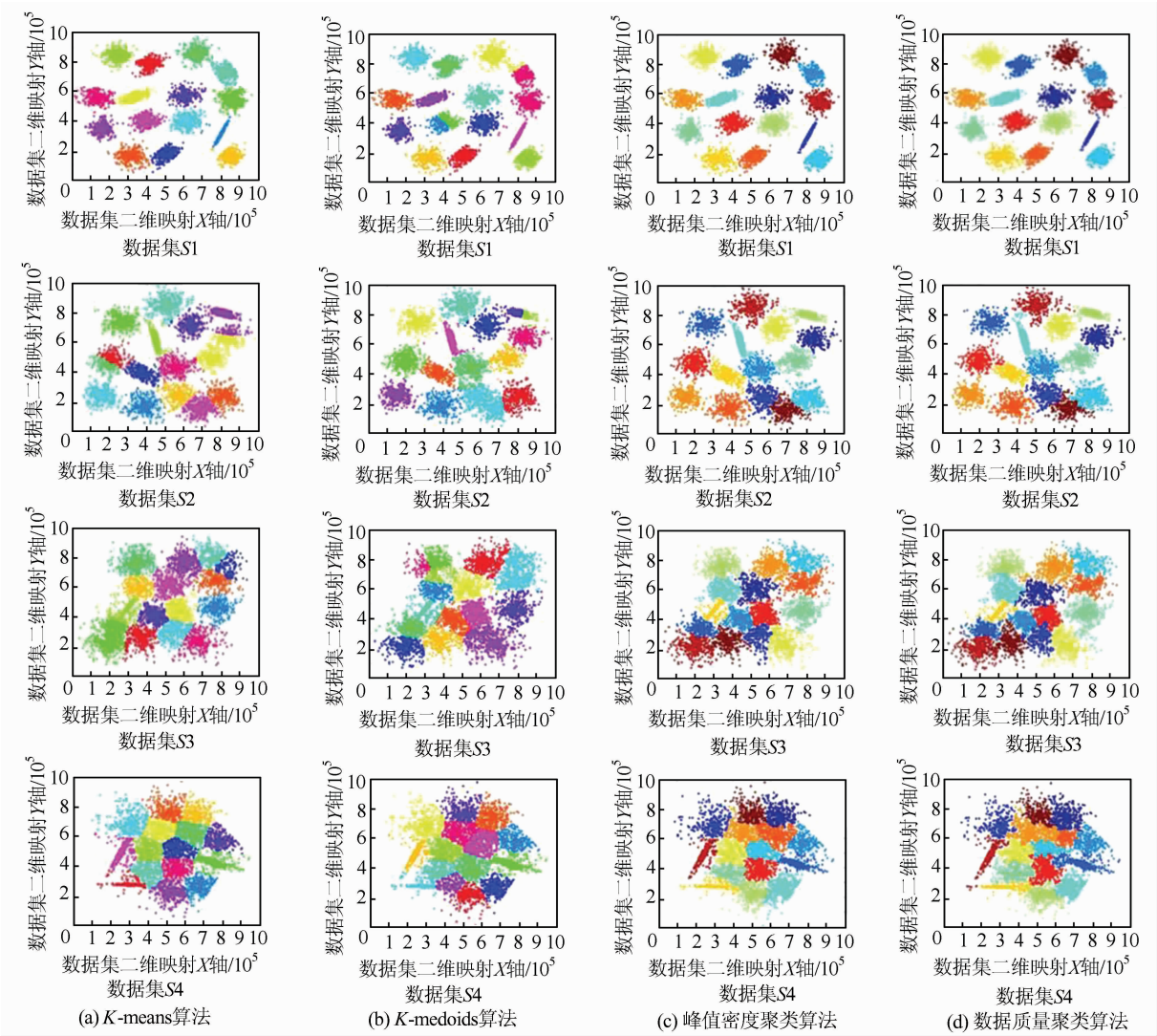

图 8　数据集 S1、S2、S3、S4 聚类结果比较

Fig. 8　Comparison of Clustering Results on Datasets S1, S2, S3, S4

表 2　数据集 *S1*、*S2*、*S3*、*S4* 聚类中心平均误差率统计/%

Tab. 2　Error Rate of Clustering Centers for Datasets S1, S2, S3, S4/%

数据集	*K*-means 算法	*K*-medoids 算法	峰值密度聚类	数据质量聚类
S1	0.37	0.49	2.81	0.14
S2	0.53	0.74	0.31	0.11
S3	0.98	1.55	0.66	0.15
S4	1.39	1.71	0.46	0.14

从表 2 中可以看出,数据质量聚类算法所确定的聚类中心与实际聚类中心的误差率最小,几乎与实际中心重合,明显优于 *K*-means 算法、*K*-medoids算法和峰值密度聚类算法。

综合数据集 *A*1、*A*2、*A*3 和数据集 *S*1、*S*2、*S*3、*S*4 的实验结果,可以认为数据质量聚类算法比传统的分割聚类算法和峰值密度聚类算法有更好的聚类效果。

2.3　实验结果分析

上述实验结果说明,数据质量聚类算法不仅可以准确提取出聚类中心的个数,而且在剩余点的划分上也有很高的准确率,对于数据集 *A*1、*A*2、*A*3 平均准确率分别达到了96.00%、96.91%和 97.49%。在确定聚类中心上,本文方法也有很高的准确率,对于数据集 *S*1、*S*2、*S*3、*S*4,聚类中心的平均误差率分别为 0.14%、0.11%、0.15%和 0.14%。数据质量聚类算法不仅在各项指标上明显优于传统的 *K*-means 算法和 *K*-medoids 算法,而且优于峰值密度聚类算法。

对于数据集 *A*1、*A*2、*A*3,数据质量聚类算法比峰值密度聚类算法在平均准确率上分别提高了 0.67、0.26 和 1.32 个百分点,而对于数据集 *S*1、*S*2、*S*3、*S*4,聚类中心的平均误差率分别降低了 20.07、2.82、4.40 和 3.29 倍。综合以上实验结果,可以证明数据质量聚类算法能够准确确定聚

类中心,并能够得到准确的聚类结果。

3 结 语

传统的中心聚类算法虽然简单快速,但是需要用户输入较多参数,并且具有球形偏差,在实际应用中有较多限制。本文提出了数据质量的概念,即代表了数据场中数据的固有属性,并且根据挖掘视角的不同,数据质量所代表的属性也不同。在本文中,赋予数据质量数据密集程度的属性,结合物理场中引力的概念,提出一种确定聚类中心的新方法,即具有较大质量和较大距离属性的点可以视为聚类中心。本文方法解决了需要用户输入参数、聚类结果受初始点影响等问题,减少了中心聚类算法在实际应用中的限制。实验结果证明,数据质量聚类算法能够准确找到数据集的聚类中心,并具有较为准确的聚类结果。

数据质量聚类算法虽然较为准确,但在实际应用中需要提高算法的效率,可以采取分布式计算的方式,这将是下一步研究的方向。

参 考 文 献

[1] Rodriguez A, Laio A. Clustering by Fast Search and Find of Density Peaks [J]. *Science*, 2014, 344 (6 191): 1 492-1 496

[2] Wang S L, Yuan H N. Spatial Data Mining: A Perspective of Big Data [J]. *International Journal of Data Warehousing and Mining*, 2014, 10(4): 50-70

[3] Wang S L, Chen Y. HASTA: A Hierarchical-Grid Clustering Algorithm with Data Field [J]. *International Journal of Data Warehousing and Mining*, 2014, 10(2): 39-54

[4] Aggarwal C C, Reddy C K. Data Clustering: Algorithms and Applications [M]. UK: Chapman & Hall/CRC, 2013

[5] Wang S L, Wang D K, Li Y, et al. Clustering by Fast Search and Find of Density Peaks with Data Field [J]. *Chinese Journal of Electronics*, 2016, 25(3): 397-402

[6] Li D R, Wang S L, Li D Y. Spatial Data Mining: Theory and Application [M]. Berlin: Springer, 2013

[7] Wang S L, Gan W, Li D Y, et al. Data Field for Hierarchical Clustering[J]. *International Journal of Data Warehousing and Mining*, 2011, 7(4): 43-63

[8] Fränti P, Virmajoki O. Iterative Shrinking Method for Clustering Problems[J]. *Pattern Recognition*, 2006, 39 (5): 761-765

Clustering Data with Mass

LI Yan[1] *WANG Dakui*[2] *GENG Jing*[1] *WANG Shuliang*[1]

1 School of Software, Beijing Institute of Technology, Beijing 100081, China

2 Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China

Abstract: The clustering center has a great effect on the clustering result. In this paper, a new concept of the data mass is proposed. The mass of data represents one of the inherent attributes of the data. With different view angles of data mining, the data mass maybe different. Based on the concept of data mass, a new clustering algorithm, which is clustering data with mass, is put forward. This new algorithm finds the clustering centers based on two attributes of data: the data mass and the data distance. And it can complete the clustering process with only one pass of the whole dataset. Experimental results show that the proposed algorithm can find the clustering center accurately and can get better clustering result than the same typical clustering algorithms, such as *K*-means, *K*-medoids and clustering by fast search and find of density peaks.

Key words: data field; cluster; data mass; clustering center

First author: LI Yan, PhD candidate, major in spatial data mining. E-mail: liy_007@126.com

Corresponding author: WANG Shuliang, PhD, professor. E-mail: slwang2011@bit.edu.cn

Foundation support: The National Natural Science Foundation of China, No. 61472039; the Specialized Research Fund for the Doctoral Program of Higher Education, No. 20121101110036.