DOI:10.11992.tis.201410021
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.tp.20150930.1557.028.html

一种改进的自适应快速 AF-DBSCAN 聚类算法

周治平,王杰锋,朱书伟,孙子文
(江南大学 物联网工程学院,江苏 无锡 214122)

摘　要:基于密度的 DBSCAN 聚类算法可以识别任意形状簇,但存在全局参数 Eps 与 MinPts 的选择需人工干预,采用的区域查询方式过程复杂且易丢失对象等问题,提出了一种改进的参数自适应以及区域快速查询的密度聚类算法。根据 KNN 分布与数学统计分析自适应计算出最优全局参数 Eps 与 MinPts,避免聚类过程中的人工干预,实现了聚类过程的全自动化。通过改进种子代表对象选取方式进行区域查询,无需漏检操作,有效提高了聚类的效率。对 4 种典型数据集的密度聚类实验结果表明,本文算法使得聚类精度提高了 8.825%,聚类的平均时间减少了0.92 s。

关键词:密度聚类;DBSCAN;区域查询;全局参数;KNN 分布;数学统计分析
中图分类号:TP181　　**文献标志码**:A　　**文章编号**:1673-4785(2016)01-0093-06

中文引用格式:周治平,王杰锋,朱书伟,等.一种改进的自适应快速 AF-DBSCAN 聚类算法[J]. 智能系统学报, 2016, 11(1): 93-98.
英文引用格式:ZHOU Zhiping, WANG Jiefeng, ZHU Shuwei, et al. An improved adaptive and fast AF-DBSCAN clustering algorithm[J]. CAAI Transactions on Intelligent Systems, 2016, 11(1): 93-98.

An improved adaptive and fast AF-DBSCAN clustering algorithm

ZHOU Zhiping, WANG Jiefeng, ZHU Shuwei, SUN Ziwen
(School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China)

Abstract: The density-based DBSCAN clustering algorithm can identify clusters with arbitrary shape, however, the choice of the global parameters Eps and MinPts requires manual intervention, the process of regional query is complex and loses objects easily. Therefore, an improved density clustering algorithm with adaptive parameter for fast regional queries is proposed. Using KNN distribution and mathematical statistical analysis, the optimal global parameters Eps and MinPts are adaptively calculated, so as to avoid manual intervention and enable full automation of the clustering process. The regional query is conducted by improving the selection manner of the object, which is represented by a seed and thus avoiding manual intervention, and so the clustering efficiency is effectively increased. The experiment results looking at density clustering of four typical data sets show that the proposed method effectively improves clustering accuracy by 8.825% and reduces the average time of clustering by 0.92 s.

Keywords: density clustering; DBSCAN; region query; global parameters; KNN distribution; mathematical statistics and analysis

数据挖掘是一种从大量数据中发现感兴趣信息的技术,聚类算法在数据挖掘应用中日益广泛。其中,基于密度的聚类算法可以发现任意形状的簇且能够较好地处理噪声数据,越来越受到广泛的关注。DBSCAN 算法能够发现任意形状的簇,并有效识别离群点,但聚类之前需要人工选择 Eps 和 minPts 2 个参数。当数据量增大时,要求较大的内存支持,I/O消耗也很大;当空间聚类的密度不均匀,聚类间距离相差很大时,聚类质量较差[1-3]。针对 DBSCAN 算法在大型数据库与多密度数据集聚类精度低,计算复杂度高,全局参数人工选取等问题,已有很多学者进行了相关研究:S. Mimaroglu 等[4]提出对位向量使用裁剪技术,H. Jiang 等[5]提出一种基于划分的 DBSCAN 算法,B. Borah 等[6]提出一种改进的基于抽样的 DBSCAN 算法,D. Kellner[7]提出基于格点的 DBSCAN 算法,旨在解决 DBSCAN 算法在内存占用,处理高维数据和密度分布不均数据聚类效果不

收稿日期:2014-10-13.　**网络出版日期**:2015-09-30.
基金项目:国家自然科学基金资助项目(61373126);江苏省产学研联合创新资金-前瞻性联合研究基金资助项目(BY2013015-33).
通信作者:王杰锋. E-mail:18352513420@163.com.

好等问题; H. F. Zhou、S. H. Yue、Y. Ma、S. JAHIRABAPKAR 和 Z. Y. Xiong 等[8-12]基于数据的数学统计特性,确定全局参数;B. Liu[13]提出一种基于密度的快速聚类方法,按照特定维的坐标排序,选择有序的未被标记的在核心对象邻域以外的点作为种子扩展簇。综上所述,基于密度聚类算法的改进点主要集中在全局参数的选择以及提高密度聚类效率等。DBSCAN 全局参数选择根据 k-dist 曲线人工确定,过程繁琐,实用性不高。其他基于统计分析的方法,部分以特定数据分布确定全局参数,而数据分布存在不确定性,以特定分布规定不能准确反映数据的分布特性,使计算出的全局参数不准确;提高密度聚类效率主要集中在区域查询中的代表对象的选择,但是选择的代表对象进行区域查询时存在丢失对象现象,对丢失对象进行查漏操作,一定程度上增加了区域查询的复杂度。

1 DBSCAN 算法及改进算法

DBSCAN 是一种经典的基于密度聚类算法[8],可以自动确定簇的数量,并能够发现任意形状的簇。Eps 近邻表示一个给定对象的 Eps 半径内的近邻称为该对象的 Eps 近邻,表示为 $\mathrm{NEps}(p)$:

$$\mathrm{NEps}(p) = \{q \in D \mid \mathrm{dist}(p,q) \leqslant \mathrm{Eps}\} \quad (1)$$

直接密度可达是指对于给定的 MinPts 和 Eps,从对象 q 可以直接密度可达 p,需要满足的条件为

$$p \in \mathrm{NEps}(q), \mid \mathrm{NEps}(q) \mid \geqslant \mathrm{MinPts} \quad (2)$$

DBSCAN 算法的全局参数 MinPts 和 Eps 的选取依赖于人工干预,对密度分布均匀的数据根据 k-dist 曲线升序排列后,人为选择曲线变化幅度开始陡升的点作为 Eps 参数,并且确定 MinPts 参数为固定常量 4,实施过程繁琐,依赖于人工干预。本文提出一种全局参数自适应选择的方法,根据数据距离空间的统计分布特性,统计出 k-dist 值的分布情况,曲线拟合出分布曲线,通过计算拟合曲线拐点处对应的值,自适应确定出 Eps 参数,并根据数据中每个点 Eps 领域内点数的分布情况,计算出参数 MinPts。

DBSCAN 以核心对象 P 来拓展一个簇,通过对包含在 P 邻域内的点进行区域查询扩展簇。包含在 P 邻域的对象相互交叉,Q 是 P 的邻域内的一个对象,如果它的邻域被 P 中其他对象的邻域所覆盖,那么 Q 的区域查询操作就可以省略,Q 不需要作为种子对象用于类的扩展。因此,用于 Q 的区域查询时间和 Q 作为核心对象的内存占用都可以被省去。而一个核心对象边界的对象更有利于作为候选对象被选为种子,因为内部对象邻域往往会被外部对象的邻域覆盖。因此,抽样种子实际上是选择的代表对象能够准确描绘出核心对象邻域形状的问题。实际上,对于密度聚类,在核心对象邻域内相当一部分种子对象可以被忽略,选择核心对象边界的部分代表对象进行类的扩展,从而达到减少区域查询频度的目的。

为了自适应确定合适的全局参数 MinPts 和 Eps,减少内存占用量和 I/O 消耗,提高 DBSCAN 的计算效率,基于这些分析,本文提出一种改进的自适应快速算法(adaptive and fast density-based spatial clustering of applications with noise,AF-DBSCAN),旨在以自适应方式确定合理的全局参数 MinPts 和 Eps,以及区域查询时选择部分具有代表性的对象作为种子对象进行类扩展。改进算法描述如下:1)自适应确定全局参数 Eps 和 MinPts;2)将所有点分类,分别标记为核心点、边界点和噪声点;3)删除标记处的噪声点;4)连接距离在 Eps 距离内的所有核心点,并归入到同一簇中;5)各个簇中的核心点对应种子代表对象的选择;6)遍历数据集,根据选择的代表对象进行区域查询,将边界点分入与之对应核心点的簇中。如果数据集中所有点都被处理,算法结束。

2 AF-DBSCAN 聚类算法

2.1 参数 Eps 与参数 MinPts 的确定

由于密度衡量指标单一,本文算法数据集主要针对簇密度差异不明显的数据。根据输入数据集 D 计算出距离分布矩阵 $\mathbf{DIST}_{n\times n}$,如式(3)所示:

$$\mathbf{DIST}_{n\times n} = \{\mathrm{dist}(i,j) \mid 1 \leqslant i \leqslant n,\ 1 \leqslant j \leqslant n\} \quad (3)$$

式中:n 为数据集 D 的对象数目;$\mathbf{DIST}_{n\times n}$是一个 n 行和 n 列的实对称矩阵,其中每个元素表示数据集 D 中对象 i 和对象 j 之间的距离。计算 $\mathbf{DIST}_{n\times n}$中的每个元素的值,然后逐行按照升序排列。用 $\mathbf{DIST}_{n\times i}$表示 $\mathbf{DIST}_{n\times n}$中第 i 列的的值,对 $\mathbf{DIST}_{n\times i}$中每一列进行升序排列得到 KNN 分布,如图 1 所示。

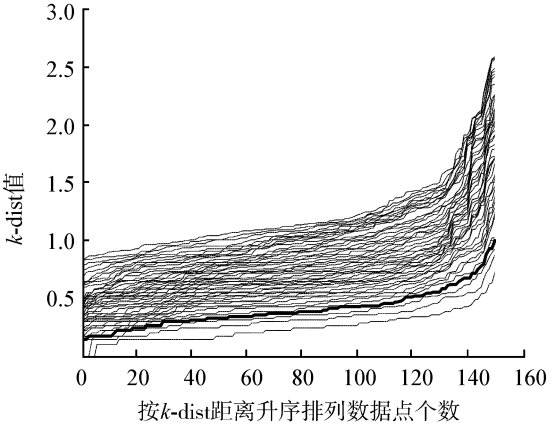


图 1 KNN 分布

Fig.1 KNN distribution

图 1 中,$k=1,2,\cdots,45$,根据 k-dist 分布曲线可以看出,$k=4$ 的 dist_4 曲线可以反映出其他 dist_k 曲线的形状。本文选择 $k=4$ 的 dist_k(k-最近邻距离)的数据进行统计分析,dist_4 的概率分布图 2 所示。

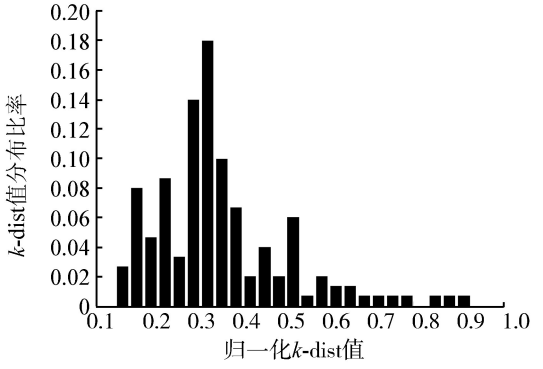


图 2　$\text{dist}_k(k=4)$ 概率分布

Fig.2　Probability distribution of $\text{dist}_k(k=4)$

从图 1 可以看出,任何一条曲线都是在平缓变化后急剧上升,dist_k 中大部分值落在一个比较密集的区域内,因此可以判断 dist_k 中大部分值应落在一个比较密集的区域内(曲线平缓段)。如果可以通过数学方法找出 dist_k 中平缓变化到急剧上升处的点,或者 dist_k 概率分布最为密集的区域,则可确定扫描半径参数 Eps,所以本文选择图 1 中 dist_k 拐点处的值为 Eps。由图 2 可以得到 dist_k 的概率分布情况,假如能够通过数学方法找到分布曲线的峰值,也可以用峰值点所对应的 k-最近邻距离值(横坐标刻度)作为 Eps。

对于概率分布数据,通过分析数据集的统计特性,建立统计模型对数据集进行曲线拟合[14]。本文通过实验对概率分布使用傅里叶、高斯和多项式 3 种典型曲线拟合方法,如图 3 所示。

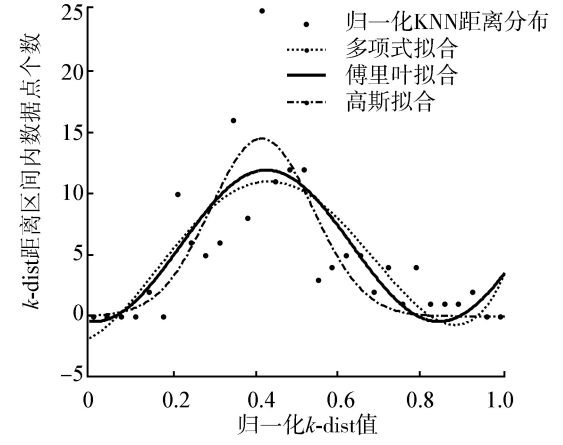


图 3　归一化 KNN 分布拟合曲线

Fig.3　Fitting curves of normalized KNN distribution

其中,高斯曲线拟合方法的效果最佳,但是由于概率分布的拟合精度过低,不可用于全局参数 Eps 的估计。拟合结果为 SSE:312.7,R-square:0.675 5,调整 R-square:0.651 4,RMSE:3.403,参数 SSE 和 RMSE 越接近 0 拟合越准确;R-square 和调整后的 R-square 越接近于 1 越准确;高斯拟合曲线如式(4)所示:

$$f(x)=a\times\exp(-((x-b)/c)^2) \tag{4}$$

根据 KNN 升序排列曲线确定 Eps,对 dist_4 曲线进行拟合。对于升序排列得到 KNN 分布数据,采用 3 种拟合方法进行曲线拟合。实验发现高斯拟合精度高,但拟合阶数高,计算复杂度高;傅里叶拟合精度不高;而多项式拟合不仅拟合阶数低,而且拟合精度高,计算复杂度低,拟合结果为 SSE:0.046 36,R-square:0.988 3,调整 R-square:0.988,RMSE:0.017 88,如图 4 所示。多项式曲线拟合如式(5)所示。

$$f(x)=ax^4+bx^3+cx^2+dx+e \tag{5}$$

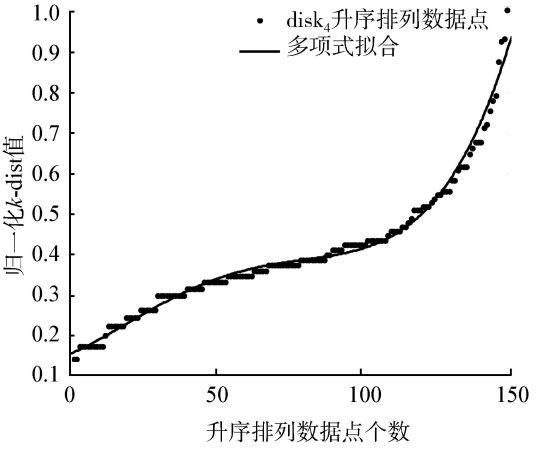


图 4　多项式拟合曲线

Fig.4　Polynomial fitting curves

根据多项式拟合曲线,求解曲线的拐点,对 y 求二次导可得 $f(x)''=12ax^2+6bx+2c$,求解二次导方程得 x 的解为 $x_0=\dfrac{-6b\pm\sqrt{36b^2-96ac}}{24a}$。由于较小的值为靠近边界的点,取 x 解中较大的值,舍去较小的值,将其带入式(5),可以得到 $\text{Eps}=f(x_0)$。Eps 确定后,需要确定 MinPts 的值。根据每个点邻域数据点的统计分布特性,依次计算出每个点的 Eps 邻域的对象数量;计算数据对象的数学期望,即 MinPts,如式(6)所示:

$$\text{MinPts}=\frac{1}{n}\sum_{i=1}^{n}P_i \tag{6}$$

式中:P_i 表示在点 i 的 Eps 邻域的点数。

本文将密度聚类算法与基于统计模型相结合，基于数理统计理论，假定数据集由统计过程产生，并通过找出最佳拟合模型来描述数据集，自适应计算出最优全局参数 Eps 和 Minpts。

2.2 种子代表对象的选择

本文提出一种改进的基于 DBSCAN 的快速聚类算法，在通过选用核心对象附近区域包含的所有对象的代表对象作为种子对象扩展类，减少了区域查询的次数，减低了聚类时间和 I/O 开销。

对于一个给出 Eps 和 MinPts 的核心对象 P，为了便于阐述，仅考虑二维对象，算法可用于其他大于二维的高维对象。代表对象选择过多则难以发挥算法效率，选择过少则容易造成对象丢失，影响算法聚类质量。FDBSCAN[15] 算法在区域查询后，在第 1 轮核心点区域查询时无丢失对象现象，而在以种子对象进行类扩展时，产生丢失对象，因此需要选择足够多的代表对象；而 I-DBSCAN[6] 在二维数据中采用至多 8 个代表对象，不存在对象丢失的情况。本文结合 FDBSCAN 与 I-DBSCAN，第 1 轮区域查询时采用 4 个代表对象进行类扩展，继续扩展类时，选择 8 个代表对象进行类扩展。本算法在提高查询效率的基础上，解决了类扩展时丢失对象的问题。

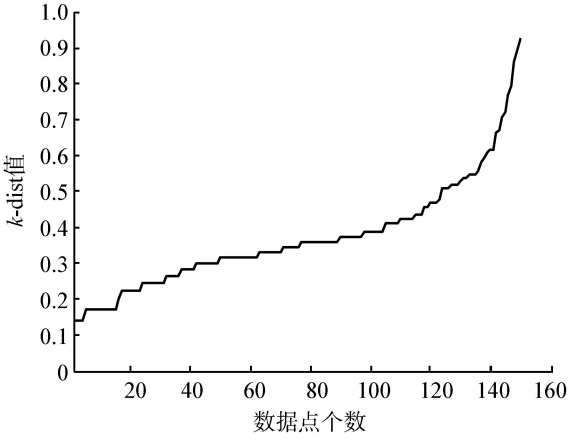
本文提出的代表对象选择方式如下：以核心对象 p 为中心，Eps 为半径画圆，以对象 p 为原点画坐标系交圆周于 A、C、E 和 G 4 点，再画 2 条分别与 x 轴成 45°和 135°角的直径交圆周于 B、D、F 和 H 4 点。第 1 轮选择代表对象时，以核心点边界的 A、C、E 和 G 点为参照，在 p 的 Eps 区域中分别选择离 A、C、E 和 G 点最近的点作为代表对象。当对于不同参照点存在离其距离最近的点为同一点时，此点只能被选择 1 次，且属于第 1 个参考点的代表对象。如果对象是 n 维数据，则至多可以选择 $2n$ 个代表对象。

在继续扩展类选择代表对象时，以核心点边界的 A、B、C、D、E、F、G 和 H 点为参照点选择代表对象，其原则为 p 的 Eps 区域中选择离参考点对象最近的点作为代表对象，即使 1 个代表对象到 2 个以上的参考点都是最近的，它也只被选 1 次，且归入第 1 个参考点的代表对象。因此，在二维空间范围内，对任一对象的被选代表对象数最多为 8 个。一般情况下，对 n 维空间，由于有 3^n-1 个参考点和 2^n 个象限，因此被选种子数最多为 3^n-1 个，按照以上方式实现区域查询，有效提高聚类效率以及解决对象丢失的问题。

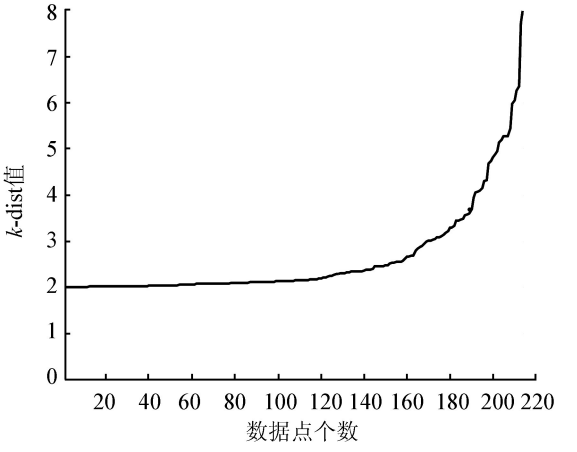
3 实验与分析

本文算法采用了 Java 语言，在 Windows XP 系统和 eclipse 环境下运行，PC 机硬件配置：Pentium (R) CPU，3 GB 内存，300 GB 硬盘。为了验证本文改进算法的有效性，根据数据集的维度、数据量和密度分布 3 种标准进行数据库的选择，选取 UCI 数据库中的 4 种典型数据集 Iris、Wine、Glass 和 cmc。根据聚类准确度和时间特性分析 2 项指标对 DBSCAN、I-DBSCAN[8] 和 AF-DBSCAN 算法性能进行比较分析，其中聚类准确度采用 F-Measure[13]。DBSCAN 中根据 k-dist 曲线，选取 dist_4 曲线图进行参数 Eps 值的确定，如图 5 所示。

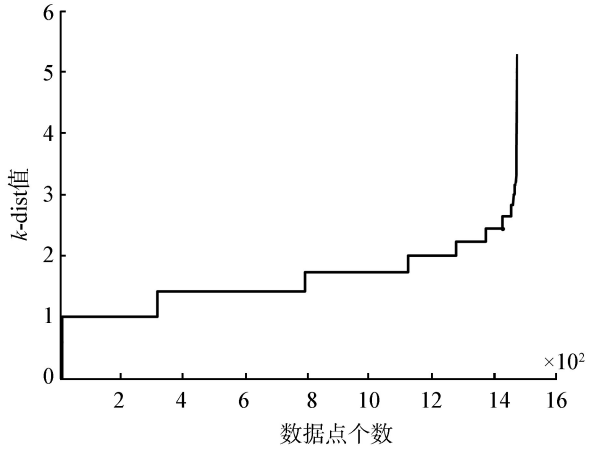
根据图 5 中平缓变化后急剧上升处对应的 k-dist 值作为全局参数 Eps 的值，且 Minpts 值设为 4。得到 4 种数据集 Iris、Wine、Glass 和 cmc 的 (Minpts, Eps) 分别为 (4, 0.436)、(4, 27.330)、(4, 3.700) 和 (4, 1.732)。



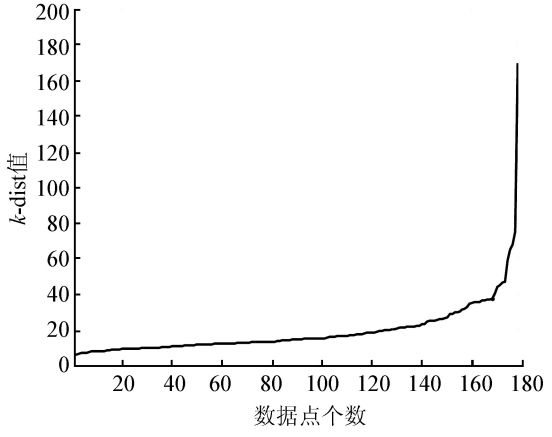
(a) Iris 数据集



(b) Glass 数据集



(c)cmc 数据集



(d)Wine 数据集

图 5　$dist_4$ 曲线

Fig.5　Curve of $dist_4$

本文提出的 AF-DBSCAN 算法的(Minpts,Eps)分别为(7,0.389)、(6,29.870)、(4,2.695)和(5,1.646)。4 种数据集聚类结果如表 1 所示。由表 1 可以看出,本文提出的 AF-DBSCAN 算法自适应计算出的全局参数减少了人为根据 *k*-dist 曲线确定全局参数 Eps 的误差及工作量,以及设定 MinPts 为固定值 4,而使聚类结果达不到全局最优的效果。通过比较分析 4 种数据集的聚类结果,AF-DBSCAN 的 *F*-Measure 值均优于其他 2 种典型算法,尤其在 Iris 和 Glass 数据集上,聚类精度比 DBSCAN 算法分别高 12.55%和 13.18%。而 I-DBSCAN 算法规定数据符合泊松分布,对于不同数据集 *F*-Measure 值不稳定,不能适应不同统计特性的数据集。由于密度衡量指标单一,AF-DBSCAN 算法适用于簇密度差异不明显的数据集。经过区域查询改进后的 AF-DBSCAN 算法,运行速度明显比 DBSCAN 和 I-DBSCAN 算法快,有效减少了密度聚类的时间。

表 1　实验比较

Table 1　Experiment comparison

数据集	算法	MinPts	Eps	时间/s	精度
Iris	DBSCAN	4	0.436	0.342	0.740 7
	I-DBSCAN	6	0.405	0.335	0.8803
	AF-DBSCAN	7	0.389	0.157	0.866 2
Wine	DBSACN	4	27.330	0.481	0.599 4
	I-DBSCAN	6	22.890	0.467	0.566 7
	AF-DBSCAN	6	29.870	0.172	0.609 1
Glass	DBSCAN	4	3.700	0.516	0.656 1
	I-DBSCAN	4	2.980	0.525	0.652 2
	AF-DBSCAN	4	2.695	0.188	0.787 9
cmc	DBSCAN	4	1.732	3.239	0.449 1
	I-DBSCAN	6	1.691	3.145	0.449 1
	AF-DBSCAN	5	1.646	1.266	0.449 1

4　结束语

本文针对 DBSCAN 算法的参数选取困难,计算效率低以及区域查询中代表对象选择后类扩展易丢失对象点等问题,提出一种改进的自适应快速 AF-DBSCAN 聚类算法,通过分析数据的 KNN 的数学统计规律,辅助用户自适应确定全局参数 Eps 与 MinPts。通过改进的区域查询方法,有效提高类扩展的效率,AF-DBSCAN 算法解决了 DBSCAN 算法人工干预,给定全局参数导致聚类质量恶化以及大数据集计算效率低的问题。

参考文献:

[1]吉根林, 姚瑶. 一种分布式隐私保护的密度聚类算法[J]. 智能系统学报, 2009, 4(2): 137-141.
JI Genlin, YAO Yao. Density-based privacy preserving distributed clustering algorithm[J]. CAAI transactions on intelligent systems, 2009, 4(2): 137-141.

[2]SMITI A, ELOUEDI Z. DBSCAN-GM: An improved clustering method based on Gaussian means and DBSCAN techniques[C]//2012 IEEE 16th International Conference on Intelligent Engineering Systems (INES). Lisbon, 2012: 573-578.

[3]ZHANG Jiashu, KEREKES J. An adaptive density-based model for extracting surface returns from photon-counting laser altimeter data[J]. Geoscience and remote sensing letters, 2015, 12(4): 726-730.

[4] MIMAROGLU S, AKSEHIRLI E. Improving DBSCAN's execution time by using a pruning technique on bit vectors[J]. Pattern Recognition Letters, 2011, 32(13): 1572-1580.

[5] JIANG Hua, LI Jing, YI Shenghe, et al. A new hybrid method based on partitioning-based DBSCAN and ant clustering[J]. Expert systems with applications, 2011, 38(8): 9373-9381.

[6] BORAH B, BHATTACHARYYA D K. An improved sampling-based DBSCAN for large spatial databases[C]//Proceedings of International Conference on Intelligent Sensing and Information Processing (ICISIP). Chennai, India, 2004: 92-96.

[7] KELLNER D, KLAPPSTEIN J, DIETMAYER K. Grid-based DBSCAN for clustering extended objects in radar data [C]//2012 IEEE Intelligent Vehicles Symposium. Alcal de Henares, Madrid, Spain, 2012: 365-370.

[8] ZHOU Hongfang, Wang Peng, LI Hongyan. Research on adaptive parameters determination in DBSCAN algorithm[J]. Journal of information & computational science, 2012, 9(7): 1967-1973.

[9] YUE Shihong, LI Ping, GUO Jidong, et al. A statistical information-based clustering approach in distance space[J]. Journal of Zhejiang university science, 2005, 6A(1): 71-78.

[10] MA Yu, GAO Yuling, SONG Shaoyun. The algorithm of DBSCAN based on probability distribution[C]//5th International Symposium on IT in Medicine and Education. Xining, China, 2014: 2785-2792.

[11] JAHIRABADKAR S, KULKARNI P. Algorithm to determine ε-distance parameter in density based clustering[J]. Expert systems with applications, 2014, 41(6): 2939-2946.

[12] XIONG Zhongyang, CHEN Ruotian, ZHANG Yufang, et al. Multi-density DBSCAN algorithm based on density levels partitioning[J]. Journal of information and computational science, 2012, 9(10): 2739-2749.

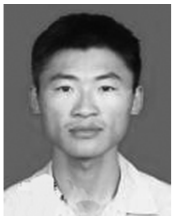
[13] LIU Bing. A fast density-based clustering algorithm for large databases[C]//2006 International Conference on Machine Learning and Cybernetics. Dalian, China, 2006: 996-1000.

[14] 夏鲁宁.SA-DBSCAN:一种自适应基于密度聚类算法[J]. 中国科学院研究生院学报, 2009, 26(4): 530-538.
XIA Luning. SA-DBSCAN: A self-adaptive density-based clustering algorithm[J]. Journal of the graduate school of the Chinese academy of sciences, 2009, 26(4): 530-538.

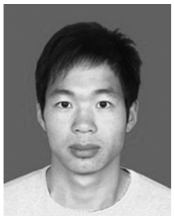
[15] 周水庚, 周傲英, 曹晶, 等. 一种基于密度的快速聚类算法[J]. 计算机研究与发展, 2000, 37(11): 1287-1292.
ZHOU Shuigeng, ZHOU Aoying, CAO Jing, et al. A fast density-based clustering algorithm[J]. Journal of computer research & development, 2000, 37(11): 1287-1292.

作者简介:

周治平,男,1962年生,教授,博士,主要研究方向为检测技术与自动化装置、信息安全等。



王杰锋,男,1989年生,硕士研究生,主要研究方向为智能信息处理。



朱书伟,男,1990年生,硕士研究生,主要研究方向为数据挖掘与人工智能。