空间密度聚类模式挖掘方法 DBSCAN 研究回顾与进展

伏家云[1]，靖常峰[1,2]，杜明义[1]
（1. 北京建筑大学 测绘与城市空间信息学院，北京 100044；2. 城市空间信息工程北京市重点实验室，北京 100038）

摘 要：针对 DBSCAN 方法存在的参数 Eps 和 MinPts 需要事先人为输入及对密度分布层次大的数据集聚类效果较差的局限性，该文对其进行了文献回顾，总结了国内外学者们的研究现状与发展，并比较分析了引用量较高方法的优点和不足，最后得出结论。对于参数确定的问题，现有学者提出了大致两种解决方法：①利用启发式方法；②与其他智能算法相结合。对于具有较大密度差数据集的适用问题，现有学者也提出了大致两种解决方法：①利用曲线斜率将数据集分层；②利用一定的规则将数据集网格化。

关键词：数据挖掘；空间聚类；密度聚类；聚类模式挖掘
【中图分类号】 P208 **【文献标志码】** A **【文章编号】** 1009-2307(2018)12-0050-08
DOI：10.16251/j.cnki.1009-2307.2018.12.010

Review and progress of DBSCAN research on spatial density clustering pattern mining method

Abstract: Aiming at some limitations that the parameters, Eps and MinPts, need to be input by human previously, and its performance on clustering effect is barely satisfactory for high density distribution levels in the density-based spatial clustering of applications with noise(DBSCAN)method, this paper reviewed literatures on the limitations of DBSCAN and summarized the research status and development of domestic and foreign scholars. In addition, this paper compared and analyzed the advantages and disadvantages of the higher citation methods. Finally we can make some conclusions. For the issue that how to determine the parameters, there has been roughly two solutions proposed by the current scholars: ① using heuristic methods; ② combining with other intelligent algorithms. For the application of large density difference data sets, the current scholars have put forward roughly two solutions: ① using the curve slope to divide the data set; ② utilizing certain rules to mesh the dataset.

Keywords: data mining; spatial clustering; density clustering; clustering pattern mining

FU Jiayun[1], *JING Changfeng*[1,2], *DU Mingyi*[1] (1. School of Geomatics and Urban Information, Beijing University of Civil Engineering and Architecture, Beijing 100044, China; 2. Beijing Key Laboratory of Urban Spatial Information Engineering, Beijing 100038, China)

0 引言

地理学中的地理现象可以被抽象为具有地理位置和地理属性的空间数据，如基于微博、Twitter 等的社交媒体事件[1-3]、移动轨迹数据[4]、城市公共设施[5-6]、城市经济文化活动区域及移动模式等。地理现象根据其表达目的不同可以被抽象为点、线、面等地理实体，根据地理学第一定律，地理实体和事件在空间上都存在着联系，并且与空间位置紧密相关[7]。空间聚类方法能发现空间实体及地理现象之间潜在的聚集模式，并进一步挖掘地理现象的成因及影响因素，为行政决策、地理模式等提供科学依据，随着地理信息技术的发展和进步，空间聚类逐渐成为地理学及测绘学的一个重要研究分支[8]。

空间聚类是传统聚类为了满足数据挖掘需求而在空间上的延伸，在继承了传统聚类的特点之上又融入了新的元素，该方法结合各个领域的特点挖掘数据中潜在的知识，为人们提供分析现实问题的需求信息，在一定程度上对人类社会和经济的发展有促进作用。空间聚类算法根据空间度

作者简介：伏家云（1993—），女，甘肃白银人，硕士研究生，主要研究方向为 GIS 空间分析、城市管理与规划、城市物联网。
E-mail：1143896070@qq.com

收稿日期：2018-06-04
基金项目：北京市自然科学基金项目(41771412)；地理国情监测国家测绘地理信息局重点实验室项目(2016NGCM10)；城市空间信息工程北京市重点实验室经费资助项目(2016203)；北京市高精尖中心科研项目(X18058)
通信作者：靖常峰 副教授 E-mail：jingcf@bucea.edu.cn

量尺度将数据集分为若干个聚类簇，其中，簇与簇之间的差异性最大，簇内数据间的相似性最大，从而形成与全局或局部分布存在显著差异的异常聚类簇。目前，空间聚类方法已广泛应用到地震源分析[9]、鸟类迁徙分析[10]、物流就业群分析[11]、土壤 CO_2 浓度影响因子分析[12]、乳腺癌空间分布模式分析[13]等众多领域。

综上所述，空间聚类算法为人们挖掘数据中的潜在知识提供了很好的技术支撑，但是随着社会经济和网络技术的不断进步，各行各业产出的多样化数据几乎都具有了大数据的 5V[14] 特点：容量大、速度快、种类多、不确定性大、价值高，这就要求空间聚类算法对数据集的可伸缩性要强，噪声点敏感度要低，可以识别任意形状的聚类簇，并且对算法输入参数的领域知识要求较低[15]。针对这种需求，空间密度聚类方法 DBSCAN(density-based spatial clustering of applications with noise)引起了学者越来越多的关注。DBSCAN 算法由文献［16］于 1996 年提出，目前已在多个领域有相关应用[17-20]，但是该算法主要存在两方面问题：①参数的确定。该算法在运行时需主观确定全局参数 Eps 和 MinPts，并且参数对聚类结果敏感性较大，导致该算法在对数据集无任何先验知识情况下不适合直接应用于大数据集。②对密度差异显著数据聚类效果较差[21]。

因此，本文针对 DBSCAN 存在的问题进行了文献回顾，总结了近年来学者提出的多种解决方法，并比较分析了其优缺点。其中，对于参数确定的问题，现有学者提出了大致两种解决方法：一是利用启发式方法；二是结合其他主流算法。对于较大密度差数据集的适用问题，现有学者也提出了大致两种解决方法：一是利用曲线斜率将数据集分层；二是利用一定的规则将数据集网格化。

1　DBSCAN 原理

1.1　基本定义

1)邻域半径(Eps)：算法运行时的搜索半径，一般用 $N\mathrm{Eps}(p)$ 表示点 p 的 Eps 半径内的点集合。

2)密度(MinPts)：搜索半径 Eps 范围内的点密度。

3)核心点：Eps 邻域内的点数超过(包括) MinPts 时的点。

4)边界点：不是核心点，但是属于其他核心点的 Eps 邻域内。

5)噪音点：既不是核心点，也不是边界点的点对象。

6)直接密度可达：$(p, q)\in$ 数据集 D，p 是核心点，$q\in N\mathrm{Eps}(p)$，则 q 称为 p 的直接密度可达点。

7)密度可达：$(p, q)\in$ 数据集 D，$q_i\in$ 数据集 $D(1\leqslant i\leqslant n)$，$q_1, q_2, \cdots, q_n$，$q_1=p$，$q_n=q$，$q_{i+1}$ 是 q_i 的直接密度可达点，则称 q 是 p 的密度可达点。

8)密度相连：$(p, q, O)\in$ 数据集 D，p、q 是 O 的密度可达点，则称 p 和 q 密度相连。

如图 1 所示：设 MinPts=5，点 a、b、c、j、h、k 的 Eps 邻域内的点密度均大于等于 MinPts，则称 a、b、c、j、h、k 为核心点；h 是 j 的直接密度可达点，k 是 h 的直接密度可达点，k 是 j 的密度可达点；a 和 f 密度相连。

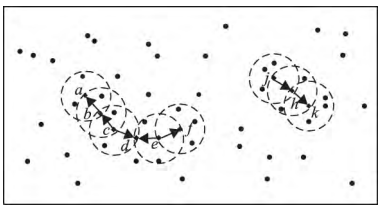


图 1　DBSCAN 算法名词图解

Fig. 1　The Illustration of DBSCAN Algorithm Nouns

1.2　DBSCAN 基本流程

1)初始化。①在数据集 O 中建立字段；②建立搜索数据集 S，临时存放符合条件的对象；③初始化参数 Eps、MinPts。

2)寻找簇。①遍历数据集 O，寻找种子核心点，核心点与其 Eps 邻域内的种子点形成一个聚类簇，将该点集合存入 S；②遍历数据集 S，寻找新的簇，并与该簇合并，将核心点移出 S，若 S 非空，重复此步骤；③执行步骤①，直至遍历完数据集 O。

3)删除空数据集 S，簇外点为噪声点。

DBSCAN 实现的伪代码如表 1 所示。

表 1　DBSCAN 核心代码

Tab. 1　DBSCAN Core Code

空间异常聚类模式挖掘算法 DBSCAN 核心代码	
目的：寻找全局聚类簇，判别其分布模式	
输入：参数 Eps、MinPts	
输出：聚类簇编号	
1	＃寻找簇
2	ExpandCluster(p, N, C, Eps, MinPts)
3	add p to cluster C
4	for each point p' in N
5	mark p'as visited
6	N'=getNeighbours(p',Eps)
7	if sizeOf(N')>=MinPts then
8	$N=N+N'$
9	end if
10	if p' is not member of any cluster
11	add p' to cluster C
12	end if
13	end for
14	End ExpandCluster

DBSCAN 原理流程图如图 2 所示。

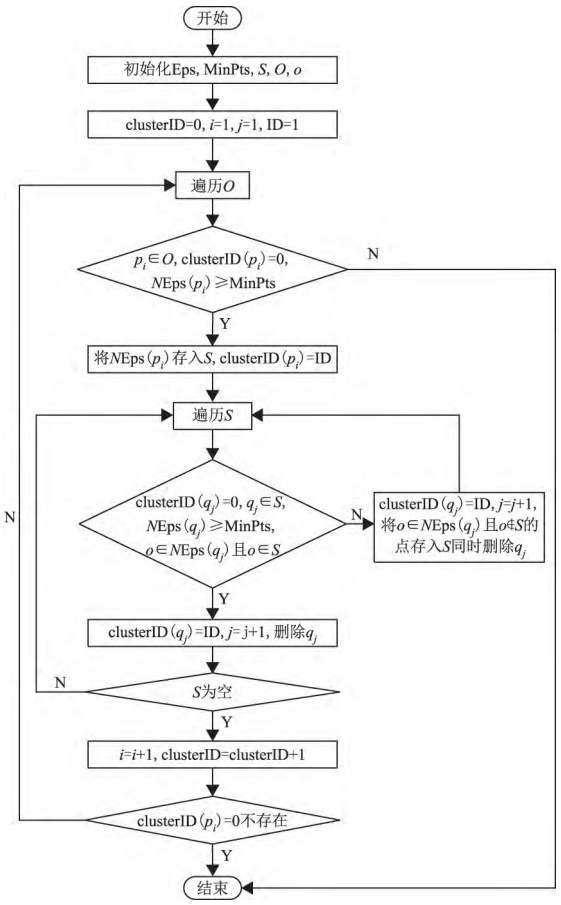


图 2 DBSCAN 原理流程图

Fig. 2 Flow Chart of DBSCAN Principle

1.3 DBSCAN 优缺点

DBSCAN 算法在考虑空间实体邻域内的局部分布情况下，可以发现：一方面，异于全部或局部的任意形状聚类簇，当算法遍历数据集时，噪声点的识别取决于核心点邻域内的点密度，而聚类簇的识别则不会受噪声点的影响，因而该算法常常被用作空间异常聚类模式的挖掘[22]；另一方面，由于 DBSCAN 对数据库中数据的输入顺序不敏感，遍历数据集时常采用 R 树或 K-D 树的数据检索方式，因此，对于低维数据其时间复杂度一般是 $O(n)$[23]，但是对于高维数据其时间复杂度则会随着参数的设置而增加。

DBSCAN 运行前需要人为输入全局参数 Eps 和 MinPts，并且由于空间数据的分布存在空间异质性，对于密度分布结构较复杂的数据集，当其参数无法实现全局动态适应时，聚类结果则会受到影响，甚至较差。

2 聚类参数确定方法研究

针对 DBSCAN 算法参数确定问题，国内外学者提出了较多的参数优化方法，并主要着重于参数自适应确定方向，对于学者提出的参数确定方法大致可以总结为以下两个方向：一是基于启发式的方向；二是结合其他主流算法，其中，大部分学者对 DBSCAN 参数的优化是基于启发式的方向。

经典 DBSCAN 算法中，参数 Eps 和 MinPts 的选取存在较大的主观性。文献[16]利用多个 2D 样本数据集，基于欧几里得距离和 k-dsit 思想分析数据集的统计特性，通过观察分析，k-dist 曲线与 4-dist 曲线没有显著性差异，鉴于分析每条 k-dist曲线特征其工作量和时间复杂度都太大，因此选取能够表征其他 k-dist 曲线特征的 4-dist(图 3)曲线为特征曲线。分析不同距离尺度下点对象的分布特征，选取噪音点与聚类簇的临界值为 Eps 值，该噪音的百分比根据用户经验输入。最后令 k 值为 MinPts，运行 DBSCAN 算法并与 Clarans 聚类算法比较分析，实验结果表明：DBSCAN 算法在发现任意形状聚类簇的同时，效率较高，其运行速度几乎达到了 Clarans 的 100 倍。这种基于 k-dist 曲线分析数据集统计特征确定参数的方法被称为启发式方法，该方法中 k 值的选取具有较大的人为主观性，而 k 值的合理选取会很大程度地提升 Eps 的功效。

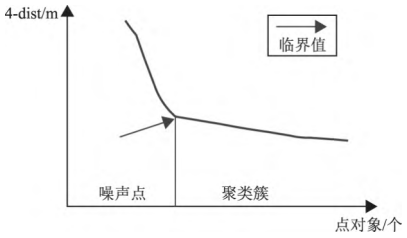


图 3 4-dist 曲线图

Fig. 3 4-dist Plot

针对 DBSCAN 算法参数确定问题，国内外很多学者进行了深入研究。文献[24]基于 Ester Martin 教授的研究提出了 GDBSCAN(generalized DBSCAN)算法，该算法由一个权重函数 wCard 和参数 Eps、MinPts 组成。通过实验分析，发现 k 值的选取对 DBSCAN 的影响较小，实验结果表明该算法的运行效率较高，但是即使如此，参数 Eps 和 MinPts 的值也是根据先验知识来判断。

在上述研究基础之上，部分学者遵循 DBSCAN 参数选取的启发式方法进行了后续的研究，如文献[25-28]基于欧几里得距离，在假设 k(或者 MinPts)为 4 的前提下，分别提出了 SA-DB-

SCAN(self-adaptive DBSCAN)、I-DBSCAN、AF-DBSCAN(adaptive and fast DBSCAN)以及AE-DBSCAN算法。SA-DBSCAN算法对4-dist曲线做逆高斯拟合，令Eps为对得到的概率函数求二阶导微分方程时的解，然后以固定的Eps值，绘制MinPts=k(k=1，2，…，n-1)时的Noise曲线，对Noise曲线函数求导，得到精确的MinPts值；当$\boldsymbol{DIST}_{n\times n}$矩阵中的每一列数据服从泊松分布时，I-DBSCAN算法计算每一列的最大似然估计值ε作为Eps_i值，在MinPts为4的前提下，绘制噪声点数量与k的关系图，通过对图形的分析解读，将噪声点和聚类簇不再急剧上升或者下降，即趋于稳定变化时的临界k值作为Eps_k，令$\mathrm{MinPts}=\frac{1}{n}\sum_{i=1}^{n}P_i$；AF-DBSCAN算法在SA-DBSCAN和I-DBSCAN算法的思想基础上，通过曲线拟合求其函数的拐点作为Eps，令$\mathrm{MinPts}=\frac{1}{n}\sum_{i=1}^{n}P_i$，实验结果表明，该算法的运行速度超于其他两种算法；AE-DBSCAN算法先假定k和MinPts为4，绘制4-dist曲线，计算曲线斜率的平均值mean(slopes)和标准差standard deviation(slopes)，然后选取曲线斜率介于mean(slopes)－standard deviation(slopes)与mean(slopes)＋standard deviation(slopes)之间的第一斜率值，将该第一斜率值对应的k-dist值作为Eps，将AE-DBSCAN算法应用到不同密度簇的2D数据中，最后评估k和MinPts值对聚类准确性和聚类簇数量的影响，实验结果表明，该算法对Eps的自适应确定结果较好，算法优于经典的DBSCAN算法。文献［29］基于启发式方法，将二进制差分进化(binary differential evolution，BDE)与DBSCAN算法相结合提出BDE-DBSCAN算法，实验结果表明，该算法可以发现多密度聚类簇，其优化纯度高于二进制遗传算法和二进制粒子群，但是该算法需要人为输入参数。

对上述引用量较高且基于启发式方法优化DBSCAN参数的算法性能进行比较分析，如表2，可以得到以下几点结论：①优化的DBSCAN算法参数通过分析数据集的统计特征自适应确定时，时间复杂度为DBSCAN算法的1.7倍，且不能适用于大规模数据集和密度分布层次较大的数据集；②优化的DBSCAN算法参数部分自适应确定时，其时间复杂度大约为DBSCAN算法的1.5倍，其中部分算法可以用于密度分布层次较大的数据集和大规模数据集，而部分算法则不适用；③优化的DBSCAN算法参数由用户指定时，其时间复杂度和DBSCAN算法相同，并且可以用于大规模数据集和密度分布层次较大的数据集。总结上述结论可以发现，根据启发式方法优化的DBSCAN算法很少且几乎没有在参数完全实现自适应确定的情况下还能应用到大规模数据集。

表2　相关密度聚类算法性能比较

Tab. 2　Performance Comparison of Relative Density Clustering Algorithm

性能	DBSCAN算法	SA-DBSCAN算法	AF-DBSCAN算法	I-DBSCAN算法	V-DBSCAN算法	AE-DBSCAN算法
实验数据大小/个	1 525～6 256	150～520	160～220	150～700	50～1 500	399～3 031
参数获取方式：用户指定(G)、自动生成(A)	G	A	A(部分)	A(部分)	G	A(部分)
参数敏感度：敏感(S)、不敏感(UNS)	S	S	S	S	S	S
时间复杂度	T	1.7T	1.46T	1.49T	T	无
能否处理大规模数据	能	否	否	否	能	能
能否处理异常点	能	能	能	能	能	能
能否表达数据集统计分布特征	能	能	能	否	能	能
能否用于密度层次较大的数据	否	否	否	否	能	能

注：A(部分)表示参数MinPts需要用户指定，Eps为自适应确定。

还有部分学者将DBSCAN算法与其他主流算法相结合，代替了DBSCAN算法参数选取的启发式方法，为DBSCAN算法参数确定开辟了全新的方向，如文献［30］首先利用K-means算法对未知协议的

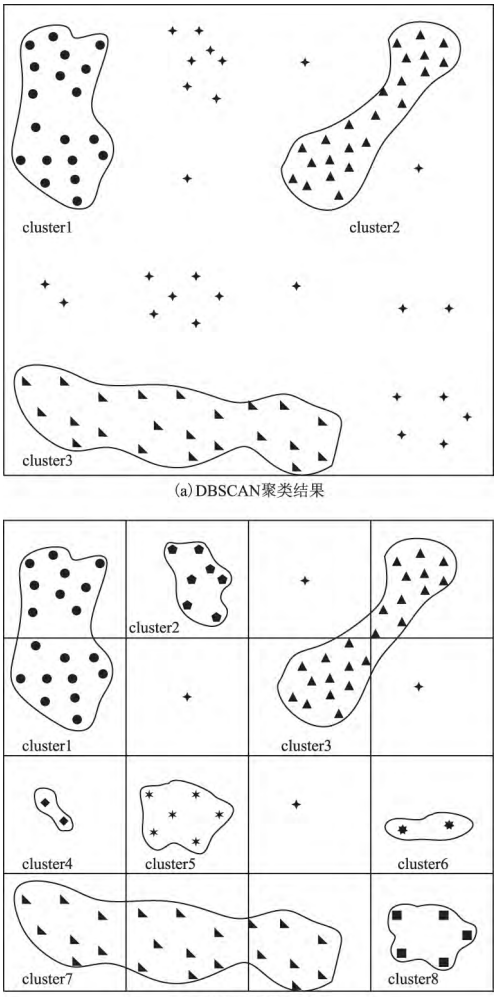
数据集进行聚类，利用最大最小距离(maximum and minimum distance，MMD)[31]选取初始聚类中心，得到 K 个类别簇，然后对第 C_i 个簇统计分析其样本距离 D_i，计算 $\mathrm{Max}D_i-\mathrm{Min}\,D_i$ 并将其划分为多个区间。令 Eps_i 为样本数最多区间的中心值，MinPts_i 为 C_i 簇内每个样本点在 Eps_i 范围内的最小点数，该方法虽然避免了 Eps 和 MinPts 的人为确定。但是因为 K-means 算法需要事先为数据集人为确定聚类簇数目 K 而引入了新的参数。文献[32]在假定 MinPts 的前提下，利用遗传算法自适应确定 Eps，但是时间复杂度仍然很大。文献[33-34]将遗传算法、并行编程模型 MapReduce 和 DBSCAN 算法相结合，对于低维数据在一定程度上减少了遗传算法和 DBSCAN 结合耗时大的问题，但是对于高维数据仍然比较耗时。因此，DBSCAN算法与其他主流算法相结合确定参数时，还存在新参数引入、计算复杂度高以及缺乏数据集的统计特性分析等问题。

3 针对非均匀数据集的应用研究

DBSCAN 算法对密度分布层次较大数据集的聚类效果不佳[13]，即不能用一个全局密度参数值去表征数据集的内部结构。实际生活中产出的数据集内部的密度结构较复杂，因此，利用 DBSCAN算法挖掘数据集中潜在的知识时，不仅需要解决 DBSCAN 参数敏感性问题，还要解决如何将优化的 DBSCAN 算法应用到密度分布差较大的复杂数据集。

针对如何将优化的 DBSCAN 算法应用于密度结构较复杂的数据集的问题，大致可以总结为以下两种解决方向：一是利用一定的规则将数据集网格化，计算每个单元格网的参数，然后根据密度相似性将参数对合并，形成一个密度层聚类簇；二是根据数据集的密度分布特性将其绘制为密度分布曲线，曲线斜率相对平缓的线段代表一个密度层次，最后在密度分布相对均匀的密度层运行优化的算法。

国内外学者在相关研究中提出了解决方法，其中一部分学者如文献[35]提出将密度分布层次较大的数据集按照一定的规则划分成 m 个网格，计算每个单元格的 Eps 和 MinPts，然后根据地理学第一定律和密度分布特征，将相似密度单元格的 Eps 和 MinPts 分别合并，利用具有密度分布差的实验数据对 DBSCAN 和其提出的E-DBSCAN(efficient DBSCAN)算法进行了聚类分析。如图 4 所示，其实验结果表明，DBSCAN 算法因密度较大的簇而忽略了密度较小的簇，并将其错误的识别为噪声点(未标有簇编号的点)，这一不足被 E-DBSCAN 算法很好地克服，与 DBSCAN 相比，E-DBSCAN 可以更有效地识别不同密度层次的簇。



(a) DBSCAN聚类结果

(b) E-DBSCAN聚类结果

图 4 DBSCAN、E-DBSCAN 运行结果[35]

Fig. 4 DBSCAN、E-DBSCAN Operation Results[35]

另一部分学者提出将数据集的密度层次按不同的曲线斜率划分，如图 5 所示，相较于斜率较大的曲线，其斜率相对平缓的曲线段代表一个密度层次。利用曲线斜率划分密度层次的有文献[36-37]提出的 VDBSCAN(varied DBSCAN)算法，其基于最近邻思想根据数据集的密度分布特性，计算相邻点之间的最近邻差得到 DK 曲线，利用 DK 曲线图将不同密度层次的数据进行分层，并进一步分析曲线的特征确定 Eps 值，最后利用曲线斜率验证分层数据的精确性。实验结果表明，VDBSCAN 算法可以在密度分布不均匀的数据集中有效地发现聚类簇，并且时间复杂度为 $O(n)$。但是 VDBSCAN 算法需要人为确定 DK 曲线。文献[38]也是基于最

近邻思想根据数据集的密度分布特性，利用最近邻曲线将数据集根据密度分层，然后利用启发式方法确定参数 Eps 值。文献［39］在 VDBSCAN 和 I-DBSCAN 算法的基础上，进一步优化了 DK 曲线和最近邻曲线的选取，他们将每个点到其最邻近 k 个点的距离求最大似然估计值，然后对该系列值升序，绘制成平均近邻曲线图，从而避免了 DK 曲线选取的人为主观性，该平均 k 距离图将数据集内部的密度分布结构进行可视化，其中不同的密度簇基于不同的曲线斜率进行划分，参数 Eps 则通过不同密度簇的 Eps_i 决定，MinPts 为每个点对象在 Eps 范围内的点对象的期望值。还有部分学者提出了易于上述两种方向的解决方法，如文献［40-41］提出的 MSDBSCAN(multi-density scale-independent DBSCAN)算法、GCMDDBSCAN(grid and contribution multi density DBSCAN)算法，实验结果表明，这两种算法都可以有效处理密度结构复杂的数据集。

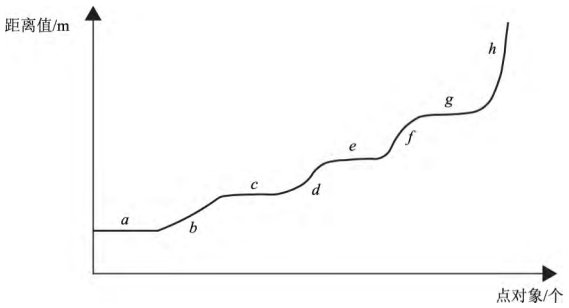


图 5　多密度层次曲线图

Fig. 5　Multi-density Level Graph

4　结束语

DBSCAN 作为空间密度聚类的典型方法，因其可以补充空间聚类在对数据集的适应性、噪声点的敏感性以及任意形状聚类簇识别方面的不足而得到了越来越多的关注，并且在实际案例分析中也得到了较多应用，如在学区房的选址中，利用 DBSCAN 将学校、城市基础设施等兴趣点的可视化与聚类可视化相结合，在保证兴趣点位置精确度的条件下还可以发现其空间集聚模式，从而为学区房的选址提供直观分析[42]。同时，由于 DBSCAN 在理论和应用上的突出优势，其在时空维度上也有了扩展，如 ST-DBSCAN 在城市路网时空轨迹上的二次聚类分析，通过在时空维度上精细化识别轨迹簇，确定道路拥堵的时空范围，进一步反映城市道路的拥堵状态及时空变化趋势[43]。随着大数据的产生，DBSCAN 除了在空间数据聚类模式识别方面的应用外，也与计算机技术及其他智能方法相结合，应用在了大数据处理的效率提升上，如在云计算环境中复杂任务的调度中，利用 DBSCAN 对云任务进行聚类，使其与已经分类的资源进行匹配，以此提升云任务与资源匹配的执行效率[44]。在未来，DBSCAN 应更多地与其他主流算法及智能技术相集成，取长补短，充分发挥其优势，并高效应用于各行业的数据聚类模式挖掘以及其他的需求问题中。

参考文献

［1］ 腾巧爽，孙尚宇，秘金钟. 众源地理空间数据的城市热点区域探测[J]. 测绘科学，2018，43(5)：74-80.（TENG Qiaoshuang，SUN Shangyu，BEI Jingzhong. Urbanhot pots detection based on crowdsourcing geospatial data[J]. Science of Surveying and Mapping，2018，43(5)：74-80.）

［2］ CHEN Huiling，ZHAO Guoqing，XU Ningyi. The analysis of research hotspots and fronts of knowledge visualization based on CiteSpace II[C]//Proceedings of the 5th international conference on Hybrid Learning. Berlin：Springer-Verlag，2012：57-68.

［3］ ADANA M. A geocomputational analysis of Twitter activity around different world cities[J]. Geo-spatial Information Science，2014，17(3)：145-152.

［4］ QIN K，ZHOU Q，WU T，et al. Hotspots detection from trajectory data based on spatiotemporal data field clustering[J]. The International Archives of the Photogrammetry，Remote Sensing and Spatial Information Sciences(ISPRS)，2017，XLII-2/W7：1319-1325.

［5］ FARKAS A. Route/Site selection of urban transportation facilities：an integrated GIS/MCDM aproach[J]. Plant Molecular Biology，2010，40(6)：969-976.

［6］ ZHAO Miaoxi，XU Gaofeng，LI Yun. Evaluating urban public facilities of Shenzhen by application of open source data[J]. Geo-spatial Information Science，2016，19(2)：1-11.

［7］ TOBLER W. On the first law of geography：a reply[J]. Annals of the Association of American Geographers，2004，94(2)：304-310.

［8］ 杨学习，徐枫，石岩，等. 一种基于场论的空间异常探测方法[J]. 武汉大学学报(信息科学版)，2018，43(3)：364-371.（YANG Xuexi，XU Feng，SHI Yan，et al. Feld-theory based spatial outlier detecting method[J]. Geomatics and Information Science of Wuhan University，2018，43(3)：364-371.）

［9］ JAGLA E A，KOLTON A B. A mechanism for spatial

and temporal earthquake clustering[J]. Journal of Geophysical Research (Solid Earth), 2010, 115 (B5): B05312.

[10] TANG Mingjie, ZHOU Yuanchun, CUI Peng, et al. Exploring the spatial distribution of bird habitat with cluster analysis[C]// 8th IEEE/ACIS International Conference on Computer and Information Science. Washington, DC, USA: IEEE Computer Society, 2009: 130-135.

[11] CHHETRI P, BUTCHER T, CORBITT B. Characterising spatial logistics employment clusters[J]. International Journal of Physical Distribution & Logistics Management, 2014, 44(3): 221-241.

[12] GIAMMANCO S, BONFANTI P. Cluster analysis of soil CO_2, data from Mt. Etna (Italy) reveals volcanic influences on temporal and spatial patterns of degassing [J]. Bulletin of Volcanology, 2009, 71 (2): 201-218.

[13] MELIKER J R, JACQUEZ G M, GOOVAERTS P, et al. Spatial cluster analysis of early stage breast cancer: a method for public health practice using cancer registry data [J]. Cancer Causes & Control, 2009, 20(7): 1061-1069.

[14] **李德仁. 展望大数据时代的地球空间信息学**[J]. **测绘学报**, 2016, 45(4): 379-384. (LI Deren. Towards Geospatial information science in big data era[J]. Acta Geodaetica et Cartographica Sinica, 2016, 45 (4): 379-384.)

[15] **陆锋, 张恒才. 大数据与广义** GIS[J]. **武汉大学学报(信息科学版)**, 2014, 39 (6): 645-654. (LU Feng, ZHANG Hengcai. Big data and generalized GIS[J]. Geomatics and Information Science of Wuhan University, 2014, 39(6): 645-654.)

[16] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//International Conference on Knowledge Discovery and Data Mining. Palo Alto, California, USA: AAAI Press, 1996: 226-231.

[17] JOSHI D. Polygonal spatial clustering[D]. Lincoln, Nebraska: University of Nebraska, 2011.

[18] KISILEVICH S, MANSMANN F, KEIM D. P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geotagged photos[C]//Proceedings of International Conference and Exhibition on Computing for Geospatial Research & Application. New York, NY, USA: ACM, 2010: 38.

[19] **李新延, 李德仁.** DBSCAN **空间聚类算法及其在城市规划中的应用**[J]. **测绘科学**, 2005, 30(3): 51-53. (LI Xinyan, LI Deren. DBSCAN spatial clustering algorithm and its application in urban planning[J]. Science of Surveying and Mapping, 2005, 30 (3): 51-53.)

[20] SUN Dayang, LI Binbin, QIAN Zhihong. Research of vehicle counting based on DBSCAN in video analysis [C]//IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing. Washington, DC, USA: IEEE Computer Society, 2013: 1523-1527.

[21] KUMAR P A, ROSHNI D. A survey on study of enhanced partition based DBSCAN algorithm[J]. International Journal of Management, IT and Engineering, 2014, 4(3): 115-123.

[22] **李军, 秦其明, 游林, 等. 利用浮动车数据提取停车场位置**[J]. **武汉大学学报(信息科学版)**, 2013, 38(5): 599-603. (LI Jun, QIN Qiming, YOU Lin, et al. Parking lot extraction method based on floating car data[J]. Geomatics and Information Science of Wuhan University, 2013, 38(5): 599-603.)

[23] SCHUBERT E, SANDER J, ESTER M, et al. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN[J]. ACM Transactions on Database Systems, 2017, 42(3): 19.

[24] SANDER J, ESTER M, KRIEGEL H P, et al. Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications[J]. Data Mining & Knowledge Discovery, 1998, 2(2): 169-194.

[25] **夏鲁宁, 荆继武.** 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, JING Jiwu. 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.)

[26] 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.

[27] **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. (ZHOU Zhiping, WANG Jiefeng, ZHU Shuwei, et al. An improved adaptive fast AF-DBSCAN clustering algorithm[J]. CAAI Transactions on Intelligent Systems, 2016, 11(1): 93-98.)

[28] OAKOK F O, CELIK M. A new approach to determine Eps parameter of DBSCAN algorithm[J]. International Journal of Intelligent Systems and Applications in Engineering, 2017, 5(4): 247-251.

[29] KARAMI A, JOHANSSON R. Choosing DBSCAN parameters automatically using differential evolution[J].

International Journal of Computer Applications，2014，91(7)：1-11.

［30］王兆丰，单甘霖. 一种基于 *K*-均值的 DBSCAN 算法参数动态选择方法［J］. 计算机工程与应用，2017，53(3)：80-86.（WANG Zhaofeng，SHAN Ganlin. *K*-means based method for dynamically selecting DBSCAN algorithm parameters［J］. Computer Engineering and Applications，2017，53(3)：80-86.）

［31］AZAR G，ALAJAJI F. On the equivalence between maximum likelihood and minimum distance decoding for binary contagion and queue-based channels with memory［J］. IEEE Transactions on communications，2015，63(1)：1-10.

［32］LIN C Y，CHANG C C，LIN C C. A new density-based scheme for clustering based on genetic algorithm［J］. Fundamenta Informaticae，2005，68(4)：315-331.

［33］HU Xiaojuan，LIU Lei，QIU Ningjia，et al. A MapReduce-based improvement algorithm for DBSCAN［J］. Journal of Algorithms & Computational Technology，2017，12(1)：53-61.

［34］SHARMA P，RATHI Y. Efficient density-based clustering using automatic parameter detection［C］//Proceedings of the International Congress on Information and Communication Technology. Singapore：Springer，2016：433-441.

［35］邱宁佳，李宾，王鹏，等. 基于 MapReduce 的密度聚类改进算法［J］. 计算机应用，2017，37(S1)：63-67.（QIU Ningjia，LI Bin，WANG Peng，et al. New density clustering algorithm based on MapReduce［J］. Journal of Computer Applications，2017，37(S1)：63-67.）

［36］CHOWDHURY A K M R，MOLLAH M E，RAHMAN M A. An efficient method for subjectively choosing parameter '*k*' automatically in VDBSCAN(varied density based spatial clustering of applications with noise) algorithm［C］//The 2nd International Conference on Computer and Automation Engineering. ［S. l.］：IEEE，2010：38-41.

［37］周董，刘鹏. VDBSCAN：变密度聚类算法［J］. 计算机工程与应用，2009，45(11)：137-141.（ZHOU Dong，LIU Peng. VDBSCAN：varied density based clustering algorithm［J］. Computer Engineering and Applications，2009，45(11)：137-141.）

［38］GAONKAR M N，SAWANT K. AutoEpsDBSCAN：DBSCAN with Eps automatic for large dataset［J］. International Journal on Advanced Computer Theory and Engineering(IJACTE)，2013，2(2)：11-16.

［39］SAWANT K. Adaptive methods for determining DBSCAN parameters［J/OL］. International Journal of Innovative Science，Engineering & Technology(IJISET)，2014，1(4)［2018-06-04］. http://www.ijiset.com.

［40］ESFANDANI G，ABOLHASSANI H. MSDBSCAN：Multi-density scale-independent clustering algorithm based on DBSCAN［C］//Proceedings of the 6th international conference on Advanced data mining and applications：Part I. Berlin：Springer-Verlag，2010：202-213.

［41］ZHANG Linmeng，XU Zhigao，SI Fengqi. GCMDDBSCAN：multi-density DBSCAN based on grid and contribution［C］//Proceedings of the 2013 IEEE 11th International Conference on Dependable，Autonomic and Secure Computing. Washington，DC，USA：IEEE Computer Society，2014：502-507.

［42］张铁映，李宏伟，许栋浩，等. 采用密度聚类算法的兴趣点数据可视化方法［J］. 测绘科学，2016，41(5)：157-162.（ZHANG Tieying，LI Hongwei，XU Donghao，et al. POI data visualization based on DBSCAN algorithm［J］. Science of Surveying and Mapping，2016，41(5)：157-162.）

［43］付子圣，李秋萍，柳林，等. 利用 GPS 轨迹二次聚类方法进行道路拥堵精细化识别［J］. 武汉大学学报（信息科学版），2017，42(9)：1264-1270.（FU Zisheng，LI Qiuping，LIU Lin，et al. Identification of urban network congested segments using GPS trajectories double-clustering method［J］. Geomatics and Information Science of Wuhan University，2017，42(9)：1264-1270.）

［44］王李彧，孙斌，秦童. 改进的 DBSCAN 聚类算法在云任务调度中的应用［J］. 北京邮电大学学报，2017，40(S1)：68-71.（WANG Liyu，SUN Bin，QIN Tong. Application of improved DBSCAN clustering algorithm in task scheduling of cloud computing［J］. Journal of Beijing University of Posts and Telecommunications，2017，40(S1)：68-71.）

（责任编辑：程锦）