

面向大规模交互数据空间划分的 Voronoi 图生成算法及应用*

熊鹏文¹, 周晓芸¹, 熊宏锦², 张婷婷³

(1. 南昌大学 信息工程学院, 江西 南昌 330031; 2. 海装驻武汉地区军事代表局, 湖北 武汉 333000;
3. 陆军工程大学 指挥控制工程学院, 江苏 南京 210007)

摘要:传统 Voronoi 图对大量点集进行 Voronoi 划分时会产生 Voronoi 单元格数过多的现象, 导致难以适用于地理信息系统、生物医学等诸多领域。为了解决这个问题, 提出一种自适应基于密度的聚类算法 (Density-Based Spatial Clustering of Applications with Noise, DBSCAN) 的 Voronoi 图。阐述了 Voronoi 单元合并的现象, 证明了其发生的充要条件, 提出该 Voronoi 图的生成算法并进行仿真。通过显微镜下嗜中性粒细胞、我国地表火点数据对算法进行验证, 结果表明, 该算法能够有效解决点集规模较大时, Voronoi 图划分过于细致的问题, 突破了传统 Voronoi 图单点对单点的划分形式。此外, 该算法拓宽了 Voronoi 图在图形图像处理、生物医学、地理信息系统等领域的应用。

关键词: Voronoi 图; 聚类; 自适应参数; 空间划分

中图分类号: TP3-0 **文献标志码:** A **文章编号:** 1001-2486(2022)01-129-08

Voronoi diagram generation algorithm and application for large-scale interactive data space partition

XIONG Pengwen¹, ZHOU Xiaoyun¹, XIONG Hongjin², ZHANG Tingting³

(1. School of Information Engineering, Nanchang University, Nanchang 330031, China;
2. Wuhan Military Representatives Bureau of Naval Equipment Department, Wuhan 333000, China;
3. Command and Control Engineering College, Army Engineering University, Nanjing 210007, China)

Abstract: When the traditional Voronoi diagram divides a large number of point sets into Voronoi, there are too many Voronoi cells, which makes it difficult to apply to such fields as geographic information systems and biomedicine. In order to solve this problem, a Voronoi diagram based on adaptive DBSCAN (density-based spatial clustering of applications with noise) was proposed. The phenomenon of Voronoi unit merge was explained. The necessary and sufficient conditions for its occurrence were proved. The algorithm for generating the Voronoi diagram was proposed and simulated. In order to verify its effectiveness, the algorithm was applied to neutrophils under the microscope and fire point data on the surface of China. The results show that the algorithm can effectively solve the problem that Voronoi diagram is too meticulous when the point set size is large, which breaks through the single point to single point division form of the traditional Voronoi diagram. In addition, the algorithm broadens the application of Voronoi diagrams in the fields of graphic image processing, biomedicine and geographic information systems.

Keywords: Voronoi diagram; clustering; adaptive parameters; space division

Shamos 和 Hoey 于 1975 年在“最近问题”中提出了 Voronoi 图这一概念, 从此拉开了计算几何的帷幕^[1]。传统的 Voronoi 图是由连接任意两邻点直线的中垂线连接构成的^[2]。平面中的点到最近的 Voronoi 单元都比其他的点到该 Voronoi 单元的距离要近。传统 Voronoi 图研究的是单个点对单个点的空间关系, 它将每个点都视为是相互独立的。然而, 在现实世界中, 一个物体通常都处于和其他物体相互关联的状态中, 我们不能仅

仅把目光投向局部某个点, 而应该从整体的角度去看待点与点之间的联系。

因此, 针对多核心 Voronoi 图的研究受到了广大研究人员的关注, 并取得了一些成果。但是由于 Voronoi 单元核心的数量限制, 或者点集数量的限制, 抑或是计算方式的限制, 多核心 Voronoi 图都不能快速地对大规模数据进行空间划分。Barequet 等^[3]提出了两点核心 Voronoi 图, 但是 Voronoi 单元核心的数量限制导致该算法的

* 收稿日期: 2020-08-07

基金项目: 国家自然科学基金资助项目(62163024, 61903175, 61663027); 江西省主要学科学术和技术带头人项目(20204BCJ23006)

作者简介: 熊鹏文(1987—), 男, 江西南昌人, 副教授, 博士, 硕士生导师, E-mail: steven.xpw@ncu.edu.cn;

张婷婷(通信作者), 女, 副教授, 博士, 硕士生导师, E-mail: 101101964@seu.edu.cn

应用具有局限性; Chen 等^[4]提出的聚类诱导 Voronoi 图 (Clustering Induced Voronoi Diagram, CIVD), 虽考虑了点与点之间的影响函数关系, 但是由于算法过于烦琐, 只能生成小规模数据的 Voronoi 图, 因此效率较低、普适性较差。陈学森等^[5]提出的基于集合影响力的 Voronoi 算法, 虽然对 Voronoi 单元核心数量受到限制的问题进行了处理, 但是需要人为输入参数来调整 Voronoi 图划分的密度, 实际处理起来较为复杂。卢嘉豪等^[6]提出了平面相交多边形的 Voronoi 图, 但是由于计算方式限制只能对少量目标进行 Voronoi 图生成。万静等^[7]提出了障碍空间中基于 Voronoi 图的不确定数据聚类算法, 但只考虑 2 级生成点以内的不确定数据和障碍情况, 适用范围窄。

对此, 为了更深入地研究点集在空间的分布关系, 本文将 Voronoi 图的概念推广到基于密度的聚类算法 (Density-Based Spatial Clustering of Applications with Noise, DBSCAN) 的 Voronoi 图 (Voronoi Diagram for DBSCAN, DBSCAN-VD), 不再将二维平面中的点集视作相互独立的, 而是通过聚类将点划分成一个一个的点集, 从而研究点集与点集之间的联系。本文还提出一种自适应确定 DBSCAN 参数的算法, 无须人为确定参数, 通过利用数据集自身分布特性生成候选和参数^[8], 自动寻找聚类结果的簇数变化稳定区间, 并将该区间中密度阈值最少时所对应的和参数作为最优参数。DBSCAN 参数的自动选取不仅可以节省计算参数的时间, 而且可以有效提高聚类的准确效果。最后以显微镜下嗜中性粒细胞的图像和我国地表火点分布图作为仿真实例, 验证 DBSCAN-VD 算法解决大规模数据下空间划分问题的有效性。

1 DBSCAN 聚类原理

DBSCAN 聚类算法是一种基于密度的聚类算法, 它可以在带有噪声的空间数据库中发现任意形状的聚类。DBSCAN 聚类算法具有良好的抗噪能力。关于 DBSCAN 聚类算法的定义^[9]如下:

定义 1 R_{eps} 邻域。假设存在空间中的点集 $D, \forall p \in D$, 以对象 p 为中心、以给定半径 R_{eps} 为半径的邻域称为对象 p 的 R_{eps} 邻域。

定义 2 密度阈值 T_{minpts} 。假设存在空间中的点集 $D, \forall p \in D$, 称满足对象 p 成为核心点的值为密度阈值 T_{minpts} 。

定义 3 核心点。假设存在空间中的点集 D ,

$\forall p \in D, \forall \{p_1, p_2, \dots, p_n\} \in D, p_1, p_2, \dots, p_n$ 在对象 p 的 R_{eps} 邻域内, 若 R_{eps} 内点数大于 T_{minpts} , 表示对象 p 的 R_{eps} 邻域内对象个数大于密度阈值, 则定义 p 为核心点。

定义 4 边界点。假设存在空间中的点集 $D, \forall p \in D, \forall \{p_1, p_2, \dots, p_n\} \in D$, 若 R_{eps} 邻域内点的数量小于 T_{minpts} , 且落在核心点的邻域内, 则称以上定义点为边界点。

定义 5 噪声点。假设存在空间中的点集 $D, \forall p \in D$, 既不是核心点也不是边界点, 则称 p 为噪声点。

DBSCAN 算法的流程首先是将所有点标记为核心点、边界点或噪声点^[10]。DBSCAN 聚类演示如图 1 所示, A 点为核心对象, 以 A 点为圆心的 R_{eps} 邻域内找到其他点。然后以各自的点为圆心、 R_{eps} 为半径作圆, 直到邻域内任意一个点在以 R_{eps} 为半径作圆时, R_{eps} 邻域不能囊括其他点, 则该点称之为边界点 B, C 。 N 点不能被任何以核心点为圆心、 R_{eps} 为半径的圆包裹进去, 所以称之为离群点。图中黑色箭头的指向表示密度可达, 以 A 点为圆心、 R_{eps} 为半径的圆构成的集合中包含了所有的密度直达样本, 集合外的均不能密度直达, 在这些密度可达的样本序列的 R_{eps} 邻域内, 所有的样本相互都是密度相连的^[11]。

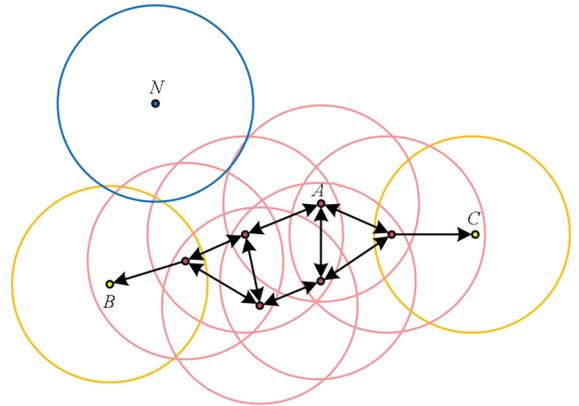


图 1 DBSCAN 聚类演示

Fig. 1 DBSCAN clustering demonstration

2 基于自适应 DBSCAN 聚类的 Voronoi 图的定义

2.1 传统 Voronoi 图的定义

Voronoi 图是对空间的最邻近划分^[12], 在给一定一些目标的情况下, 将空间划分成若干个区域, 所有划分区域内的任意一点都距离该区域内的目标最近。传统 Voronoi 图是根据平面点集绘制的分割图像, 是针对每一个点与其他点之间的位置

关系进行邻域划分。图 2 所示的是一个点集数量为 8 的 Voronoi 图,这 8 个点为 Voronoi 图的核心点,平面空间被任意两点连线的中垂线划分成若干个区域,这些区域为 Voronoi 单元,每个 Voronoi 单元中的任意一点都距离该 Voronoi 单元中的核心点最近。

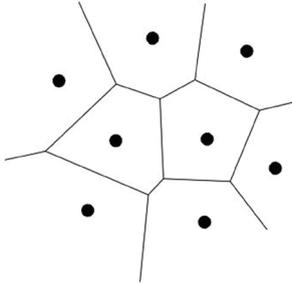


图 2 传统点集 Voronoi 图

Fig.2 Voronoi diagram of traditional point set

令 $P = \{p_1, p_2, \dots, p_n\}$ 为平面上 n 个点的集合,其中 p_i 表示集合中的任意一点, $d(p, q)$ 表示点 p 与点 q 之间的欧式距离,则定义:

$$V(p_i) = \{x \mid d(x, p_i) < d(x, p_j), p_i \in P, p_j \in P, p_i \neq p_j\}$$

集合 $V(p) = \{V(p_1), V(p_2), \dots, V(p_n)\}$ 便是 Voronoi 图^[13]。

Voronoi 图给出了平面的空间划分,集合中的元素为 Voronoi 核心点,如图 2 中的 8 个点。 $V(p_i)$ 是一个域,为对应于 p_i 的 Voronoi 域,也称作 Voronoi 单元。划分平面空间的线段,即 Voronoi 单元的边界,为 Voronoi 边。

2.2 DBSCAN-VD 与传统 Voronoi 图的关系

传统 Voronoi 图是针对每一个点进行 Voronoi 单元划分,Voronoi 单元的数量随着点集的数量增加而增加。在点集数量过大的情况下,Voronoi 图划分会产生图像分割过于密集、Voronoi 单元格过多的问题。如图 3 所示的点集为 500 的 Voronoi 图,由于目标数量较多,对每一个点进行 Voronoi 分割产生的划分结果过于细致,实际应用价值很小。

因此,由问题出发引出基于自适应 DBSCAN 聚类的 Voronoi 图(DBSCAN-VD)生成算法研究。DBSCAN-VD 的示意如图 4 所示,为了便于观察,同一种颜色的点表示通过 DBSCAN 聚类归纳为一类的点集。图中有五种颜色的小圆,分别代表五类点集,每个点集中的蓝色小点则是代表点集的核心,最后再以点集的核心生成传统 Voronoi 图。DBSCAN-VD 算法就是依据最近邻关系,将这五个点集进行划分。

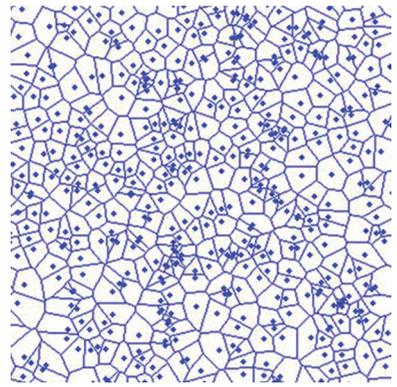


图 3 500 个点的 Voronoi 图

Fig.3 Voronoi diagram of 500 points

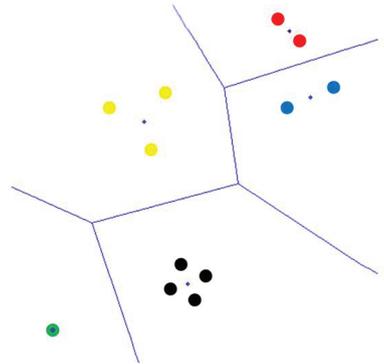


图 4 DBSCAN-VD 的简化形式

Fig.4 Simplified form of DBSCAN-VD

2.3 DBSCAN-VD 的定义

DBSCAN-VD 算法通过对生成的点集进行聚类,从而生成若干个与点之间有密切联系的点集的子集,再在此基础上生成点集的子集的 Voronoi 图。在通过对该 Voronoi 图的定义、性质进行研究之后,得出基于 DBSCAN 聚类的 Voronoi 图的生成算法。

定义 6 平面上存在一个点 p 和一个点集 s ,定义 p 到 s 的距离为 p 到点集 s 中最近一点的距离值。

$$d(p, s) = \inf(d(x, p) \mid x \in s) \quad (1)$$

定义 7 平面上存在一个以点集为元素的集合 $S, S = \{s_1, s_2, \dots, s_n\}$,其中 s_i 表示一个点集, Voronoi 图定义为平面上距离 s_i 最近的点的集合。 $V(s_i) = \{x \mid d(s_i, x) \leq d(s_j, x), \forall s_i \in S, s_j \in S, i \neq j\}$ S 的 Voronoi 图为:

$$V(S) = \{V(s_1), V(s_2), \dots, V(s_n)\} \quad (2)$$

定理 1 给定平面上的一个集合 $S = \{s_1, s_2, \dots, s_n\}$,其中

$$s_i = \{p_1^i, p_2^i, \dots, p_{m_i}^i\} \quad (3)$$

p_i^j 表示一个点,可以得到:

$$V(s_i) = V(p_1^i) \cup V(p_2^i) \cup \dots \cup V(p_{mi}^i) \quad (4)$$

证明:

$$\forall p_0 \in (V(p_1^i) \cup V(p_2^i) \cup \dots \cup V(p_{mi}^i))$$

其中, p_0 表示一个点, p_0 必然离集合 $\{p_1^i, p_2^i, \dots, p_{mi}^i\}$ 中的某个点最近, 假设这个点是 p_i^i , 则

$$p_0 \in V(p_j^i)$$

$$p_j^i \in s_i = \{p_1^i, p_2^i, \dots, p_{mi}^i\}$$

根据定义 6, p_0 离 s_i 最近。再根据定义 7, 可以得到

$$p_0 \in V(s_i)$$

所以

$$(V(p_1^i) \cup V(p_2^i) \cup \dots \cup V(p_{mi}^i)) \in V(s_i)$$

另外,

$$\forall \dot{p}_0 \in V(s_i)$$

其中, \dot{p}_0 表示一个点, \dot{p}_0 必定离 s_i 中的一个点最近, 假设这个点是 $p_k^i (p_k^i \in s_i)$, 可以得到

$$\dot{p}_0 \in V(p_k^i)$$

然后

$$V(p_k^i) = V(p_1^i) \cup V(p_2^i) \cup \dots \cup V(p_{mi}^i)$$

可以得到

$$\dot{p}_0 \in (V(p_1^i) \cup V(p_2^i) \cup \dots \cup V(p_{mi}^i))$$

所以

$$V(s_i) \in (V(p_1^i) \cup V(p_2^i) \cup \dots \cup V(p_{mi}^i))$$

因此

$$V(s_i) = V(p_1^i) \cup V(p_2^i) \cup \dots \cup V(p_{mi}^i) \quad \square$$

定理 1 实际上是依据 Voronoi 图的最近邻关系, 对点集与点集之间进行划分。上述推导在数学上证明了 DBSCAN-VD 生成算法的合理性, 接下来通过仿真实验来证明算法的正确性。

3 DBSCAN-VD 的基本生成算法

本文提出的 DBSCAN-VD 的基本生成算法首先采用自适应参数最优法^[14] 得出 DBSCAN 聚类所需的两个参数值 R_{eps} 和 T_{minpts} , 然后对生成的点集进行 DBSCAN 聚类, 找出点与点之间的空间关系, 从而将点集划分为若干个点与点之间有密切联系的点集的子集, 最后生成聚类后形成的子集的 Voronoi 图。

3.1 DBSCAN 聚类参数的确定

DBSCAN 聚类的第一步就是确定 R_{eps} 和 T_{minpts} 这两个参数的值^[15]。 R_{eps} 和 T_{minpts} 体现了聚类的合理性, 从而决定了 DBSCAN-VD 生成的效果。为了确保聚类后产生的点集的子集规模不过大或过小, 本文采用的自适应 DBSCAN-VD 算法

可以自动确定参数值, 提高了聚类的准确度。

基于 DBSCAN 算法, 定义密度阈值 T_{density} 为 R_{eps} 邻域内存在的 T_{minpts} 个数据点, 可以得到:

$$T_{\text{density}} = \frac{T_{\text{minpts}}}{\pi \cdot R_{\text{eps}}^2} \quad (5)$$

首先定义聚类的“距离”: 本文将欧式距离作为聚类距离。计算数据集 D 的距离分布矩阵, 即

$$D_{n \times n} = \{D_{\text{distance}}(i, j) \mid 1 \leq i \leq n, 1 \leq j \leq n\} \quad (6)$$

其中: $D_{n \times n}$ 为 $n \times n$ 的实对称矩阵; $D_{\text{distance}}(i, j)$ 为数据集 D 中第 i 个对象到第 j 个对象的距离。

R_{eps} 参数列表是通过 K -平均最近邻法和数学期望法产生的, 通过计算数据集 D 中每个数据点与其第 K 个最近邻数据点之间的 K -最近邻距离, 并对所有数据点的 K -最近邻距离求平均值, 得到数据集的 K -平均最近邻距离。第 K 列的元素构成所有数据点的 K -最近邻距离向量 D_K 。对向量 D_K 中的元素求平均, 可得到向量 D_K 的 K -平均最近邻距离, 并将其作为候选 R_{eps} 参数, R_{eps} 参数列表表示为:

$$D_{\text{eps}} = \overline{D_K} \quad 1 \leq K \leq n \quad (7)$$

然后采用数学期望法生成 T_{minpts} 参数列表, 再依次求出每个 R_{eps} 参数对应的 R_{eps} 邻域对象数量, 并计算所有对象的 R_{eps} 邻域对象数量的数学期望值, 作为数据集 D 的 T_{minpts} 参数, 表示为:

$$T_{\text{minpts}} = \frac{1}{n} \sum_{i=1}^n P_i \quad (8)$$

其中, P_i 为第 i 个对象的 R_{eps} 邻域对象数量, n 为数据集 D 中的对象总数, 最后求解出聚类类别数列表。

根据数据集提取出 R_{eps} 候选项 (按从小到大排列), 然后再提取出 T_{minpts} 候选项, 随后用这些候选项尝试使用 DBSCAN 算法进行聚类, 如果连续的候选项聚类的类别数目相同, 那么选择 R_{eps} 相对较大的那个作为最终参数输入 DBSCAN 算法中去。

3.2 点集划分处理

首先对需要进行空间划分的物体提取二维平面点集, 接着以每个点为圆心、 R_{eps} 为半径的圆来搜索簇, 当点 p 的 R_{eps} 邻域包含的点数大于 T_{minpts} 时, 创建一个以 p 为核心点的新簇。聚类算法依据 R_{eps} 和 T_{minpts} 找出所有核心点, 然后分别以核心点为起点, 找出由其密度可达的对象生成的新簇, 直到遍历完所有点便结束。如图 5 所示, 虚线表示 R_{eps} 邻域, p_1 是核心对象, p_2 由 p_1 密度直达, p_3 由 p_1 密度可达, p_3 与 p_1 密度相连。

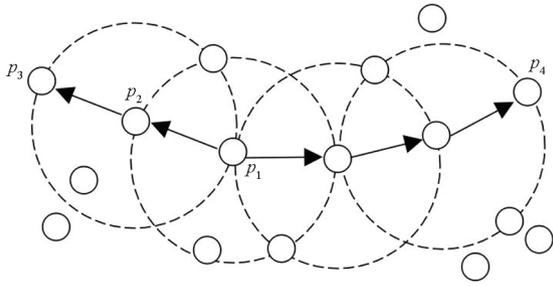


图 5 点集划分处理过程

Fig. 5 Point set division process

3.3 DBSCAN-VD 生成算法步骤

由定理 1 得到生成算法的具体步骤如算法 1 所示。

算法 1 DBSCAN-VD 的生成

Alg. 1 Generation of DBSCAN-VD

输入: 一个集合 $S, S = \{p_1, p_2, \dots, p_n\}$, 其中 p_i 表示一个点集

输出: Voronoi 图

Step 1: 输入所有点集中的所有点

Step 2: 确定 R_{eps} 和 T_{minpts} 的值

def $D_{\text{distance}}(a, b)$ 定义两个样本点之间的欧式距离

def R_{eps} 数据集

return: R_{eps} 候选集合

def T_{minpts} 距离矩阵

return: T_{minpts} 候选项列表

for i in range(距离矩阵横坐标)

for j in range(距离矩阵纵坐标)

if 矩阵中的点 $\leq R_{\text{eps}}$;

count = count + 1;

return: 聚类类别数列表

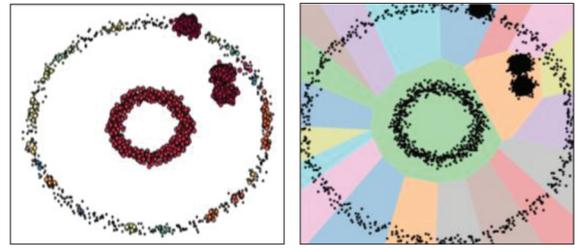
Step 3: 对点集中的所有点进行 DBSCAN 聚类, 得到若干个数据集 s_i ;

Step 4: 用传统 Voronoi 图算法构建数据集 s_i 基于 DBSCAN 聚类的 Voronoi 图

end

图 6 展现了算法的过程。由图 6 可以清晰地看出算法的核心思路: 首先随机生成点数为 5 000 的点集, 然后对点集进行 DBSCAN 聚类生成若干个子集, 最后用传统 Voronoi 图算法生成由点集的子集构建的 Voronoi 图。定理 1 可以证明该算法思路的正确性。

F 值是一种评价聚类结果的综合指标。设正确率 $C_{\text{precision}}$ 为判断正确的数据个数占识别出的数据个数的百分比; 召回率 C_{recall} 为判断正确的数据个数占实际的数据总数的百分比。则 F 值表



(a) 聚类图 (b) Voronoi 图
(a) Cluster diagram (b) Voronoi diagram

图 6 点集数量为 5 000 时的 Voronoi 图生成结果
Fig. 6 Result of Voronoi diagram generation when the number of point sets is 5 000

示为:

$$F = \frac{2 \cdot C_{\text{precision}} \cdot C_{\text{recall}}}{C_{\text{precision}} + C_{\text{recall}}} \quad (9)$$

在聚类结果簇数正确的前提下, 密度阈值越小, 则噪声率越低, 而且聚类结果的 F 值呈增大趋势。这是由于随着密度阈值的降低, 越多的低密度数据点被划分进簇, 因此噪声率变低, 同时 F 值的增大则表明准确率增加, 聚类结果趋于正确。聚类结果对比如表 1 所示。

表 1 聚类结果对比

Tab. 1 Comparison of clustering results

数据集	对象数	算法	F 值
D31	3 100	DBSCAN-VD	0.993 8
		AF-DBSCAN	0.815 3
		I-DBSCAN	0.896 7
		Kernel-DBSCAN	0.288 0
R15	600	DBSCAN-VD	0.995 0
		AF-DBSCAN	0.820 2
		I-DBSCAN	0.885 2
		Kernel-DBSCAN	0.677 1
Compound	399	DBSCAN-VD	0.914 6
		AF-DBSCAN	0.818 0
		I-DBSCAN	0.877 6
		Kernel-DBSCAN	0.730 3

从 F 值评价指标来看, 对不同类型的数据集进行聚类, DBSCAN-VD 算法均优于对比算法, 说明 DBSCAN-VD 算法具有较高的精确性。

4 仿真实例及分析

为了证明本文算法的有效性和可靠性, 采用 Python 和 OpenCV 2. 4. 11 对算法进行了实现。实验环境为 Windows7 操作系统, 8 GB 内存 Intel

i5 - 6500, 3.20 GHz 处理器。

4.1 嗜中性粒细胞的划分

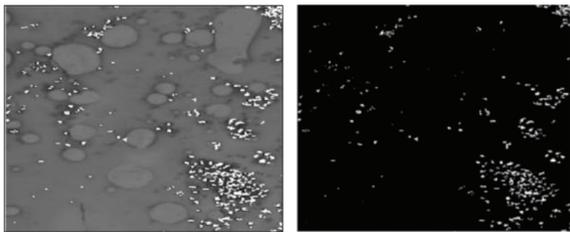
白细胞是人体内血液中的一种细胞,医学上往往检测白细胞的形态和数量作为疾病诊断的依据^[16]。正常情况下机体的白细胞数目为 $4.0 \times 10^9/L \sim 10.0 \times 10^9/L$ 。当机体发生炎症等疾病时均会造成白细胞数目的变化,因此检测白细胞的形态和数量对人体检测疾病来说具有重大意义。血常规检测中,医生根据各个参考值分析白细胞的各个指标,从而对机体进行初步分析^[17],具体白细胞的数值与人体机能的关系如下:

1) 白细胞计数增多。当人体内的白细胞总数每微升超过 10 000 个时,一般被诊断为病理性白细胞升高。病理性升高最常见的原因是感染,尤其是细菌感染,感染程度越高,白细胞的数量越大,呈正比关系。

2) 白细胞计数减少。当人体内的白细胞总数每微升少于 4 000 个时,一般会被诊断为病理性白细胞降低。病理性降低最常见的原因是病毒感染,如流行性感冒、病毒性肝炎、水痘等,均能引起病理性白细胞降低。

因此,检测白细胞数目可以为人体疾病的诊断提供依据,白细胞数目的大小也可以反映该疾病的严重程度。

验证过程为:首先从嗜中性粒细胞镜下图片中提取出二维平面点集,如图 7 所示。



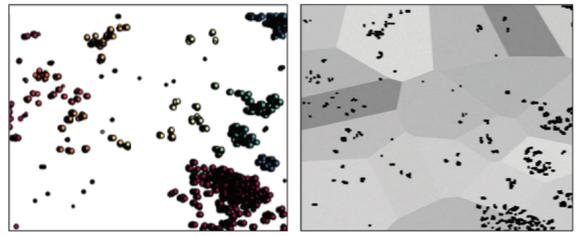
(a) 显微镜下图片 (b) 二维平面点集
(a) Picture under the microscope (b) Two-dimensional plane point set

图 7 嗜中性粒细胞显微镜下图片

Fig. 7 Microscope diagram of neutrophils

然后将采集到的点集调用 DBSCAN-VD 算法,得出结果如图 8 所示。由图 8 可知,实例产生了 19 个 Voronoi 单元格,在目标对象为大规模嗜中性粒细胞的情况下,DBSCAN-VD 算法大大减少了 Voronoi 单元的数量,更能清晰地观察嗜中性粒细胞群的分布情况,而不是像传统 Voronoi 图对每一个嗜中性粒细胞进行 Voronoi 划分。

本文采用该仿真实例对基于集合影响力的多



(a) 聚类后的图片 (b) DBSCAN-VD 的划分
(a) Picture after clustering (b) Division of DBSCAN-VD

图 8 嗜中性粒细胞显微镜下 DBSCAN-VD 图

Fig. 8 DBSCAN-VD diagram of neutrophil under microscope

点核心 Voronoi 图算法进行比较,结果如下:

MPUVD (multi-point and unique Voronoi division) 算法需要代入参数 $\epsilon = 1.2$ 进行计算 (ϵ 为调整点集划分的密度参数,需要人为调整),得出近似优结果如图 9 所示。实例共计算 1 038 个采样点,生成了 59 个多点核心单元格。

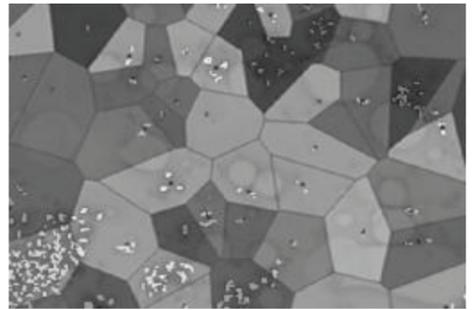


图 9 MPUVD 算法 Voronoi 划分图

Fig. 9 Voronoi partition diagram of MPUVD algorithm

本文 DBSCAN-VD 算法无须输入聚类参数,结果如图 10 所示。自动生成 19 个 Voronoi 单元格,在目标对象为大规模嗜中性粒细胞的情况下,DBSCAN-VD 算法大大减少了 Voronoi 单元的数量,更能清晰地观察嗜中性粒细胞群的分布情况。

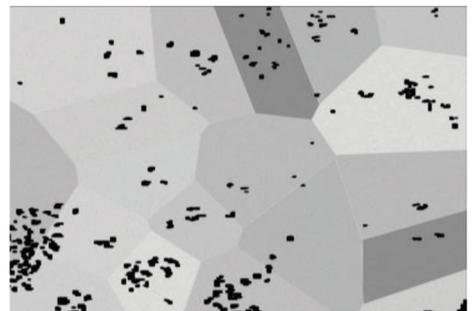


图 10 DBSCAN 算法 Voronoi 划分图

Fig. 10 Voronoi partition diagram of DBSCAN algorithm

DBSCAN-VD 算法的优势有以下几点:

1) 采用 DBSCAN-VD 算法可以减少 Voronoi 单元的数量,从而提高 Voronoi 图的生成效率。

2)采用 DBSCAN-VD 算法可以保证聚类效果的准确性和合理性。

3)DBSCAN-VD 算法可以根据离散点分布情况自动生成聚类参数,而 MPUVD 算法的参数需要多次调试才能获得。

4.2 地表火点的空间划分

火是引起地球表面异常的重要因素,可以驱使自然界森林、灌木、草原等生态系统植被的演替。植被燃烧过程中的火产生出大量的气溶胶粒子以及各种微量气体,对大气环境和全球碳循环有着重大的影响。如图 11 所示,本文采用地表火点数据来验证 DBSCAN-VD 的实际应用效果,地表火点数据来自中国科学院遥感与数字地球研究所研发的 SatSee-Fire 近实时地表高温异常点查询服务系统。

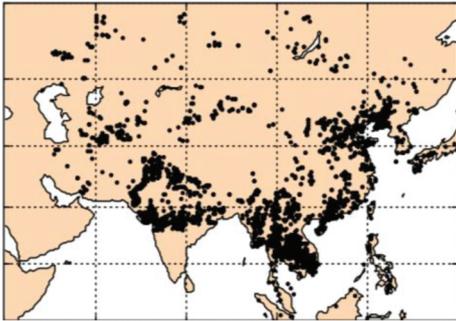


图 11 地表火点数据分布

Fig. 11 Distribution diagram of surface fire data

验证过程为:首先由地表火点数据分布图找出火点的地理位置,然后提取二维平面中点集,如图 12 所示。

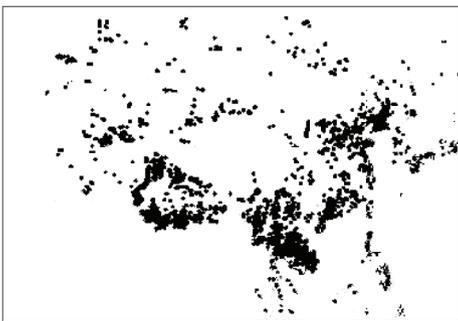
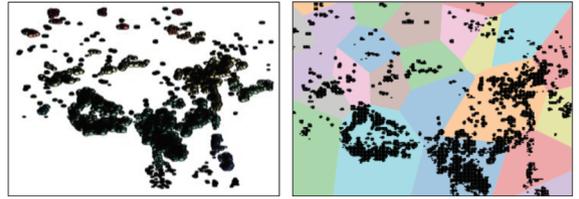


图 12 地表火点数据分布的二维平面点集

Fig. 12 Two-dimensional plane point set for the distribution of surface fire data

将采集到的点集调用 DBSCAN-VD 算法,得出结果如图 13 所示。由图 13 可知,实例产生了 23 个 Voronoi 单元格,在目标对象为大规模地表火点的情况下, DBSCAN-VD 算法大大减少了 Voronoi 单元的数量,可以更清晰地看出我国地表

火点在空间上的划分,由此得到精确的火点高发区域。如图 13 所示,我国南部、东南亚和西亚是地表火点的高发区域, DBSCAN-VD 算法可以提高地表火点的预防和监测效率、精度。



(a) 聚类后的图片 (b) DBSCAN-VD 的划分
(a) Picture after clustering (b) Division of DBSCAN-VD

图 13 地表火点的 DBSCAN-VD 图

Fig. 13 DBSCAN-VD diagram of surface fire

上述嗜中性粒细胞、地表火点的仿真实例充分证明了本文提出的 DBSCAN-VD 算法突破了传统 Voronoi 图在划分对象为大规模点集时,产生图像分割过度的问题。这不仅为 Voronoi 图发展做出了贡献,也让 Voronoi 图应用更加广泛。

5 算法复杂度分析

对于二维数据而言, DBSCAN 算法的基本时间复杂度为 $O(n^2)$, n 为数据集中数据点的数目。自适应 DBSCAN 算法在 DBSCAN 算法的基础上进行迭代运算,迭代次数由最优参数对应的 K 值大小决定,在算法实际运行过程中, K 远小于 n 时,聚类过程即发生收敛。因此, DBSCAN-VD 的时间复杂度为 $O(Kn \ln n)$ 。Voronoi 图的基本时间复杂度为 $O(n \ln n)$, 本文提出的该算法复杂度为 $O(Kn \ln n) + O(n \ln n)$ 。

综上,本文算法的算法复杂度虽然较传统 Voronoi 图算法略高,但是提升了聚类的精确性,适用于点集数量较大时生成 Voronoi 图的应用场景。

6 结论

目前,传统单点 Voronoi 图很难适用于大规模数据下的空间划分。本文从大规模点集进行 Voronoi 图划分时产生 Voronoi 单元个数过多而导致分割过度的问题出发,提出了一种基于自适应 DBSCAN 聚类的 Voronoi 图。通过对显微镜下嗜中性粒细胞、地表火点数据进行 DBSCAN-VD 仿真,结果可得 DBSCAN-VD 算法适用于大规模数据下的空间划分,突破了传统 Voronoi 图单点对单点的一种划分形式,为 Voronoi 图应用于地理信息系统、生物医学等领域拓宽了应用。在今后

的工作中,将从点集的 Voronoi 图推广到多边形集合的 Voronoi 图,并且针对 Voronoi 图尝试改良其生成算法,提高算法的运算效率。

参考文献 (References)

- [1] 刘金义, 刘爽. Voronoi 图应用综述[J]. 工程图学学报, 2004, 25(2): 125-132.
LIU J Y, LIU S. A survey on applications of Voronoi diagrams[J]. Journal of Engineering Graphics, 2004, 25(2): 125-132. (in Chinese)
- [2] AURENHAMMER F. Voronoi diagrams—a survey of a fundamental geometric data structure[J]. ACM Computing Surveys, 1991, 23(3): 345-405.
- [3] BAREQUET G, DICKERSON M, EPPSTEIN D, et al. On 2-site Voronoi diagrams under geometric distance functions[J]. Journal of Computer Science and Technology, 2013, 28(2): 267-277.
- [4] CHEN D Z, HUANG Z Y, LIU Y W, et al. On clustering induced Voronoi diagrams[J]. SIAM Journal on Computing, 2017, 46(6): 1679-1711.
- [5] 陈学森, 蔚承建, 王开, 等. 基于集合影响力的多点核心维诺图算法实现[J]. 计算机工程与设计, 2016, 37(8): 2110-2115.
CHEN X S, WEI C J, WANG K, et al. Realization of collection influence based multipoint as unit Voronoi diagram algorithm[J]. Computer Engineering and Design, 2016, 37(8): 2110-2115. (in Chinese)
- [6] 卢嘉豪, 熊鹏文, 闵卫东, 等. 平面上可相交凸多边形的 Voronoi 图[J]. 计算机辅助设计与图形学学报, 2019, 31(9): 1609-1616.
LU J H, XIONG P W, MIN W D, et al. Voronoi diagram for intersecting convex polygons in the plane[J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(9): 1609-1616. (in Chinese)
- [7] 万静, 崔美玉, 何云斌, 等. 障碍空间中基于 Voronoi 图的不确定数据聚类算法[J]. 计算机研究与发展, 2019, 56(5): 977-991.
WAN J, CUI M Y, HE Y B, et al. Uncertain data clustering algorithm based on Voronoi diagram in obstacle space[J]. Journal of Computer Research and Development, 2019, 56(5): 977-991. (in Chinese)
- [8] 刘青宝, 侯东风, 邓苏, 等. 基于相对密度的增量式聚类算法[J]. 国防科技大学学报, 2006, 28(5): 73-79.
LIU Q B, HOU D F, DENG S, et al. Relative density based incremental clustering algorithm[J]. Journal of National University of Defense Technology, 2006, 28(5): 73-79. (in Chinese)
- [9] 许丽利. 聚类分析的算法及应用[D]. 长春: 吉林大学, 2010.
XU L L. Cluster analysis algorithm and its application[D]. Changchun: Jilin University, 2010. (in Chinese)
- [10] SUN Q, DIAO M, LI Y B, et al. An improved binocular visual odometry algorithm based on the random sample consensus in visual navigation systems[J]. Industrial Robot, 2017, 44(4): 542-551.
- [11] TRAN T N, DRAB K, DASZYKOWSKI M. Revised DBSCAN algorithm to cluster data with dense adjacent clusters[J]. Chemometrics and Intelligent Laboratory Systems, 2013, 120: 92-96.
- [12] 张雷, 孙振江, 王道波, 等. 一种用于 SEAD 任务的改进型 Voronoi 图[J]. 国防科技大学学报, 2010, 32(3): 121-126.
ZHANG L, SUN Z J, WANG D B, et al. An improved Voronoi diagram for suppression of enemy air defense[J]. Journal of National University of Defense Technology, 2010, 32(3): 121-126. (in Chinese)
- [13] 杨钦, 张俊安, 李吉刚, 等. 二维限定 Voronoi 网格剖分细化算法[J]. 计算机辅助设计与图形学学报, 2006, 18(10): 1547-1552.
YANG Q, ZHANG J A, LI J G, et al. 2D conforming Voronoi mesh generation by refinement algorithm[J]. Journal of Computer-Aided Design & Computer Graphics, 2006, 18(10): 1547-1552. (in Chinese)
- [14] 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 L N, JING J W. 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. (in Chinese)
- [15] HE Y B, TAN H Y, LUO W M, et al. MR-DBSCAN: a scalable MapReduce-based DBSCAN algorithm for heavily skewed data[J]. Frontiers of Computer Science, 2014, 8(1): 83-99.
- [16] 麻若珊. 数字图像处理在白细胞分割中的应用研究[J]. 现代信息技术, 2019, 3(19): 111-112.
MA R S. Application of digital image processing in leukocyte segmentation[J]. Modern Information Technology, 2019, 3(19): 111-112. (in Chinese)
- [17] 李永霞, 吴蔚, 杨涛, 等. 新型冠状病毒肺炎患者早期外周血白细胞分类特点[J]. 中华内科杂志, 2020, 59(5): 372-374.
LI Y X, WU W, YANG T, et al. Characteristics of peripheral blood leukocyte differential counts in patients with COVID-19[J]. Chinese Journal of Internal Medicine, 2020, 59(5): 372-374. (in Chinese)