

DOI:10.13296/j.1001-1609.hva.2019.04.024

基于改进DBScan算法的雷暴挖掘与研究

高攀^{1,2,3}, 田浩^{2,3}, 李健^{2,3}, 陶汉涛^{2,3}, 王钊^{2,3}, 姜志博^{2,3}

(1. 南瑞集团公司(国网电力科学研究院), 南京 211106; 2. 国网电力科学研究院武汉南瑞有限责任公司, 武汉 430074;
3. 电网雷击风险预防湖北省重点实验室, 武汉 430074)

摘要: 为了从已发生的雷电现象中提取雷暴并研究雷暴活动规律。文中根据雷电定位系统采集的雷电数据,使用大数据、数据挖掘等计算机技术,开发基于Spark框架的改进DBScan算法,进行雷电聚类分析,提取雷暴,在ArcGIS软件中动态分析雷暴产生、爆发、移动、消亡全过程,最后通过计算机对雷暴进行分析统计,寻找雷暴活动规律。根据历史雷电数据逆向挖掘雷暴,掌握雷暴的活动规律,对今后雷暴预警奠定基础。

关键词: 海量雷电数据; 雷暴聚类; 改进DBScan算法; 雷暴动态分析; 雷暴活动规律

Excavation and Research of Thunderstorm Based on Improved DBScan Algorithm

GAO Pan^{1,2,3}, TIAN Hao^{2,3}, LI Jian^{2,3}, TAO Hantao^{2,3}, WANG Zhao^{2,3}, JIANG Zhibo^{2,3}

(1. NARI Group Corporation(State Grid Electric Power Research Institute), Nanjing 211106, China;
2. Wuhan NARI Co., Ltd., State Grid Electric Power Research Institute, Wuhan 430074, China; 3. Hubei Key Laboratory of Power Grid Lightning Risk Prevention, Wuhan 430074, China)

Abstract: In order to extract thunderstorms from existing lightning phenomena and study thunderstorm activity law, based on the lightning data collected by the lightning location system, this paper develops an improved DBScan algorithm based on Spark framework, carries out lightning clustering analysis, extracts thunderstorms, dynamically analyzes the whole process of thunderstorms' occurrence, outbreak, movement and extinction in ArcGIS software, and finally checks them by computer. Thunderstorms are analyzed statistically to find thunderstorm activity rules. According to the historical thunderstorm data, the thunderstorm is excavated in reverse, and the thunderstorm activity law is mastered, which lay a foundation for thunderstorm early warning in the future.

Key words: massive lightning data; thunderstorm clustering; improved DBScan algorithm; thunderstorm dynamic analysis; thunderstorm activity law

0 引言

雷暴是指伴有雷鸣和闪电的强对流天气系统,它产生在发展旺盛的积雨云中,常伴有强烈的降雨或冰雹,在冬季甚至会随暴风雪而来。雷暴的发生通常需要层结不稳定、必要的水汽条件和抬升条件。雷暴与电力、通信、交通、建筑、核试验和导弹等现代高科技的发展密切相关,还与森林保护、国家基础建设相关^[1-2]。目前,国内雷暴的研究主要集中在雷暴的预报预警、雷暴的风场结构分析、利用多普勒天气雷达探测雷暴结构与演变、利用中尺度

模式模拟雷暴以及雷暴的电场结构观测与分析^[3]。

中国是雷击重灾区,最为严重的是广东省以南的地区,东莞、深圳、惠州一带的雷电自然灾害已经达到世界之最,其中东莞的雷击灾害最为严重,每年夏季5-8月份,东莞因为雷击所造成的GDP亏损达6%,上千万的经济损失^[4]。不同积雨云产生的雷电数量差异很大,并不是所有的积雨云都会形成强烈的雷电现象,当雷电数量少或不够密集时,对输电线路、通信、建筑等设施造成危害的可能性很小^[5],因此文中主要研究那种范围广、密度大(产生的雷电在时间和空间上密集)可能对输电线路造成危害的雷暴。雷暴在时间和范围上是有限的,积雨云电荷

收稿日期:2018-09-20; 修回日期:2018-11-10

基金项目:国家自然科学基金项目(51777037);国家电网公司科技指南项目(524625170011)。

Project Supported by National Natural Science Foundation(51777037), Science and Technology Guide Project of SGCC (524625170011).

积累到一定程度,正负能量足够大时雷电产生,积雨云内的能量释放,不再形成极性,伴随降雨,积雨云消亡,同一地点不同时间形成的积雨云也是完全不同的。

文中依托武汉某公司的电网雷暴活动定量分析项目,从近七年雷电监测定位装置采集的真实雷电数据入手,对雷暴进行从结果到过程的逆向研究,通过计算机数据挖掘技术对海量雷电数据预处理,提取雷暴,使用 ArcGIS 地理系统展示雷暴产生、爆发、移动、消亡的过程,统计雷暴活动规律。

1 DBScan 聚类算法及 Spark 简介

1.1 聚类算法简介

迄今为止,聚类还没有一个学术界公认的定义。简单来讲聚类是一种数据挖掘方法,属于无监督学习,能够自动发现数据源间的相似性,将有高相似度的数据聚集到一个簇中,将不相似的数据尽可能分开^[6]。根据雷电数据挖掘雷暴是研究雷暴活动规律的前提条件,雷暴是雷电在空间和时间上的聚集,使用合适的聚类算法就可以提取雷暴。

常见的聚类算法有基于密度聚类 DBScan^[7]、基于事先划分的 K-means^[8]、基于层次聚类的 BIRCH^[9]、基于网格聚类的 STING^[10] 算法。

DBScan 算法事先确定元素的邻域半径以及元素成为核心点的最小密度数,寻找核心点,逐步扩张簇类范围,确定簇类;再寻找核心点,以此循环,直到遍历整个数据集,DBScan 算法不受聚类顺序的影响,可以发现任意形状和大小的簇类,对噪声点不敏感。DBScan 算法需要事先确定聚类结果的簇类数,随机选取各簇类的初始点逐步向外扩展直至最终的聚类结果,聚类结果受选取的初始点和噪声点影响大,聚类结果不固定。BIRCH 算法需要事先根据待聚类数据在内存中生成聚类特征 CF-Tree^[11],利用 CF-Tree 以此处理整个数据集,得到聚类结果。雷电聚类不应受聚类顺序影响,聚类结果大小和形状复杂多样,雷暴(簇类)数量无法事先确定,雷暴是某区域内多个雷电聚集的现象,与区域内雷电密度直接相关。因此使用 DBScan 算法进行雷电聚类是最合适的。

1.2 DBScan 简介

DBScan(density-based spatial clustering of applications with noise)是由 M.Ester, H.P. Kriegel 等人提出的一种基于密度的空间聚类算法。DBScan 算法的本质是寻找类簇并不断扩展类簇的过程,直至不再扩大。它将簇定义为密度相连的点的最大集合,能够把具有足够高密度的区域划分为簇,并在有噪

声点的数据中发现任意形状的聚类,DBScan 算法聚类效果见图 1。

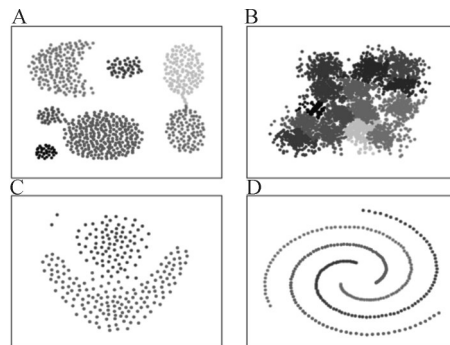


图1 DBScan 算法聚类效果图

Fig. 1 DBScan algorithm clustering effect diagram

DBScan 算法需要提前设置簇类的最小密度数 $MinPts$ 、节点邻域半径 D 。当节点邻域半径内的节点(邻域节点)数量大于 $MinPts$ 时,则该节点为核心点,从该节点向外循环扩展,直至到达簇类的边缘,不再扩展;然后再依次寻找其他簇类;直到遍历完所有节点^[12]。常规 DBScan 算法见图 2。

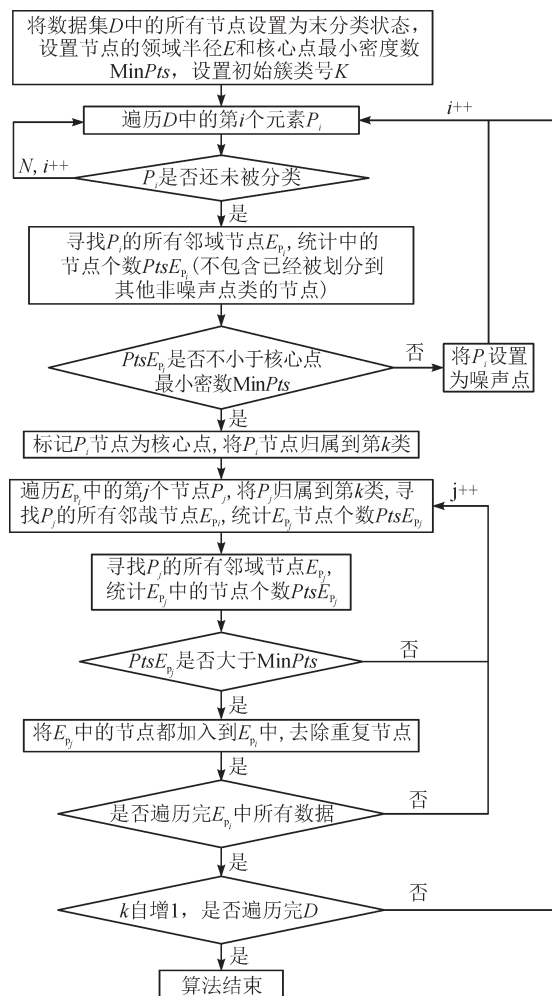


图2 常规 DBScan 流程图

Fig. 2 Regular DBScan flow chart

常规 DBScan 算法主要针对单核处理器,无法充

分发挥服务器的机器性能。将所有数据放在同一个内存中,而DBScan算法需要计算每个节点和其他所有节点的距离,寻找自己的所有邻域节点,算法的时间复杂度和空间复杂度为 $n \times n$ 。在面对海量数据时,常规DBScan算法显得束手无策。

1.3 Spark 简介

分布式处理是将不同地点,或具有不同功能的,或拥有不同数据的多台计算机(处理器)通过通信网络连接起来,在控制系统的统一管理控制下,协调地完成大规模信息处理任务的计算机系统。简单的说,就是将数据规模庞大、处理流程复杂的计算机任务在统一管理系统的控制下分拆到多个处理器计算,最后合并整理计算结果^[13-15]。

Spark^[16]就是针对于多核处理器、大容量内存的硬件环境所开发的处理框架,Spark主要基于弹性分布式数据集(RDD)、内存计算、多核并行计算的分布式处理模式,充分发挥机器的使用性能。虽然Spark技术已经相当成熟,但是目前Spark并没有支持DBScan算法的通用函数库,DBScan算法与其他众多聚类算法相比,有其独特的优点,具有其他算法无法替代的作用,如果能开发出基于Spark的改进DBScan程序,可以极大提高DBScan算法的处理能力,意义重大。

2 改进DBScan算法

2.1 雷电数据与参数设定

文中研究的雷电数据全部来源于分布在全国33个省市近1000个雷电探测站采集的真实雷电数据,通过雷电探测站—前置机—中心站—总站的数据传输模式,最终将全国的实时雷电数据存入数据库。雷电数据主要有产生时间、经度、维度等属性。雷暴不仅仅是雷电在空间上的聚集,同时也是时间上的聚集,即同一雷暴中雷电的发生时间应该呈现一定的连续性,前后两个雷电的发生时间间隔不应该超过一定时间范围 T_{span} 。DBScan算法采用欧式距离计算节点之间距离,时间尺度无法使用欧式距离进行计算,需要将时间尺度转换为对应的空间尺度。

冯万兴^[17]等在进行雷暴活动预测研究中,经过多次试验发现当DBScan的参数为以下值时接近真实的聚类现象:邻域半径 $D=0.05^\circ$ (经纬度上的 1° 约等于平面上的100 km),最小密度数 $MinPts=10$,不同雷暴间隔时间 $T_{span}=15$ min,雷暴中心点的预测轨迹与实际运动轨迹的误差范围约为 $0.1^\circ(\pm 0.1^\circ)$ 。

G. R. Huffines^[18]等在地闪雷电密度及雷暴持续时间的研究中统计探空站5 km范围内的15 min时

间内的雷电次数,当连续15 min内没有雷电发生,则雷暴结束。通过监测数据统计归纳,可以很好的反映全美国的整体雷暴状况。

孙凌^[19]等在雷暴持续时间与地闪活动的预报方法研究中将不同雷暴间隔时间取为15 min,基本可以反应出湖南省实际的地闪活动的变化规律。

空军司令部气象研究室^[20]在统计国外雷暴特性的研究中发现单个积雨云水平尺度通常在5~20 km,有一些超级单体雷暴甚至可以达到100 km以上,大部分都是多个积雨云倾斜堆积在一起,范围达数十公里。

文中在参考众多文献的基础上,结合积雨云特性(单个积雨云的水平尺度一般约为几公里至十几公里,单个积雨云形成的雷暴持续时间一般在十几分钟到几十分钟之间),将雷暴的时间间隔设置为15 min,雷暴核心点邻域半径为 0.05° ,核心点最小密度数为10。并在下文通过雷电聚集效应验证最小密度数取值的合理性,通过雷电最大生存周期验证不同雷暴时间间隔的合理性。

可以认为当两个雷电的发生时间间隔超过 T_{span} 时等同于它们在空间上不是对方的邻域节点,因此文中按照 T_{span} 的比例将雷电的时间属性转换为空间属性。

2.2 算法流程图

文中改进的DBScan算法流程图见图3。

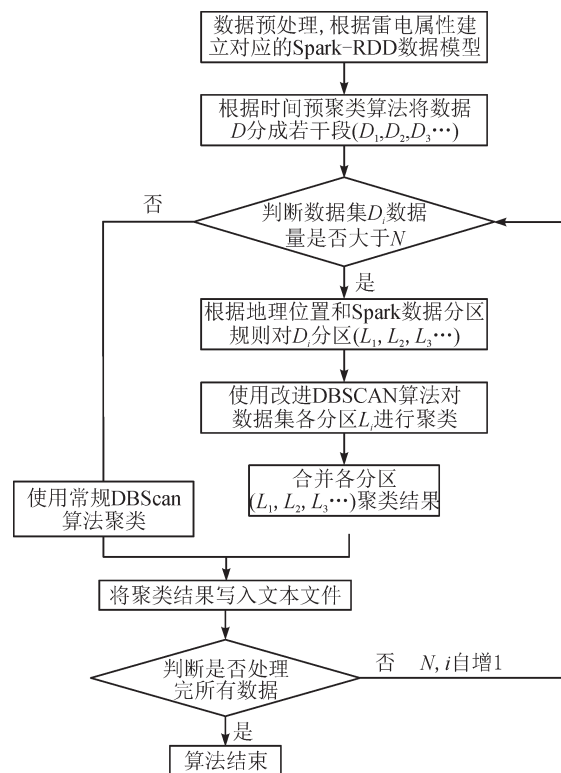


图3 改进DBScan算法流程图

Fig. 3 Improve DBScan algorithm flow chart

2.3 算法实现

2.3.1 数据预处理

建立雷电数据对应的 Thunder 模型类, Thunder 类包含雷电的序号、时间、经度、纬度、时间转换为空间距离的值、分区聚类簇类 ID(同时也为最终的簇类 ID, 0 代表未聚类、-1 代表噪声点、1~ n 代表簇类号)、地理分区 ID、Spark 分区 ID、边界聚类簇类 ID、雷电数据在分区聚类中的类别(-1: 噪声点, 1: 边界点, 2: 核心点)等属性。

2.3.2 时间分段预聚类

雷电活动在时间尺度上有很大的随机性, 并不是每一时刻都有雷电产生, 雷暴是多个雷电活动在短时间内聚集在一起产生的, 因此雷暴在时间尺度上有一定的连续性^[21], 一个雷暴中的雷电活动不应该出现时间上的断层, 否则认为是不同的雷暴。因此可以从雷电产生的时间上将雷电数据预聚类分为多个小的数据集, 然后依次对每个小数据集进行分析处理, 提高算法的处理能力和效率。根据以上分析, 将数据集进行如下处理

1) 将数据集进行时间上的排序。

2) 从数据集开始取第一个数据子集, 如果两个连续数据时间间隔在 $T_{span}(\text{min})$ 以内, 认为其属于同一个数据子集。

3) 如果两个连续数据时间间隔在 TS_{pan} 分钟以上, 则从后一个数据开始取第 2 个数据子集。依次类推, 直到整个数据集被划分为多个小数据集。

2.3.3 DBScan 算法边界问题

DBScan 算法中一直都存在边界点归属问题^[22]。如果两个簇类 C_1 、 C_2 在边缘部分有交集, 点 A 同时属于簇类 C_1 和 C_2 (边界点), 如果规定边界点归属于后一个簇类, 则 A 点在簇类 C_1 聚类时已经划分为 C_1 的元素, 在 C_2 聚类时 A 点又将被划分为 C_2 的元素, 如果簇类 C_1 除去类似于 A 点的这些点之后元素数量小于簇类阈值(既簇类的最小密度数), 就会出现簇类 C_1 中元素不足, 但是其也是一个簇类的结果, 这种结果与 DBScan 聚类期望的结果是相悖的; 如果规定边界点归属于前一个簇类, 则 A 点在计算 C_2 类的密度数时也会被计算在内, 但是 A 点却属于 C_1 类, 如果 C_2 类在除去类似于 A 点的这些点之后元素数量小于最小密度数, 则会出现 C_2 元素不足, 但也成为一个簇类的结果。见图 4。

文中从实际问题出发, 提出了一种解决这种问题的方法。

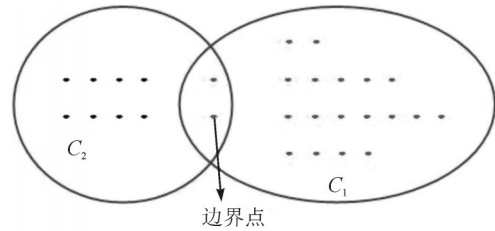


图4 DBScan 边界点归属示意图

Fig. 4 DBScan boundary point attribution diagram

1) 将边界点 A 归属于前一个簇类 C_1 。

2) 在 C_2 进行聚类时从核心点开始, 即在计算 C_2 聚类起始点的密度数时去除边界点。这样就可以保证 C_2 点在除去类似于 A 点的边界点之后还是满足最小密度数要求的。

2.3.4 改进 DBScan 算法

1) 数据分区。地理分区: 夏季是雷电活动高发期, 全国一天的雷电最多能达到上百万个, 雷电活动范围大而且持续时间长, 此时很难从时间尺度上将整个数据集划分为多个小的数据集, 而常规 DBScan 无法同时处理百万级的雷电数据。文中采用大而化小、并行处理、结果合并的思想, 将全国范围内的雷电按地理位置划分。根据雷电纬度进行地理划分, 每 N 个纬度的宽度划分为一个区域, 根据纬度位置为每个区域设置地理区号 geographyID, 区号为纬度对 N 整除的结果。Spark 分区: 经过时间分区和地理位置分区之后, DBScan 已经可以独立处理每个小的分区了, 但是由于中国幅员辽阔, 地理位置分区导致小分区较多, 而由于雷暴的聚集性, 有很多分区中的数据很少, 雷暴聚集的分区数据比较大。如果单纯的使用地理分区会导致分区数量过多、各个分区之间的数据量严重失衡。因此有必要在地理位置分区的基础上再进行 Spark 的 K 值分区, Key 值取每个区的地理区号 geographyID, 设定 Spark 的分区数为 $PartitionSize$, 则以地理分区 Key 值除以 $PartitionSize$ 的余数作为分区的 $HashCode$ 值, $HashCode$ 值相同的地理分区划分到 Spark 的第 $HashCode$ 个分区中。

2) 各分区聚类。经过前期的数据分区工作, 在 Spark 各分区中利用常规 DBScan 算法进行雷电聚类, 聚类结果划分出各雷电点所属的簇类以及各雷电点在簇类中的类别(核心点、边界点、噪声点), 便于在后续的分区合并中使用。

3) 分区合并。选取各分区的边界点: 选取各分区距离分区边界线 e (即为 DBScan 的邻域半径) 以内的元素点作为边界点。

对各分区相接的边界部分进行聚类: 对各个边

界区域进行 DBScan 聚类, 边界聚类与分区聚类算法基本一致, 边界聚类之后, 边界节点在聚类中更新边界节点的边界簇类号及其在簇类中的类别(核心点、边界点、噪声点)。

分区合并: 每个分区的簇类之间是存在一定关系的, 可能本来是一个簇类的元素, 由于分区的原因而被划分为两个簇类或者噪音点。经过分区聚类和边界聚类之后, 通过元素的分区簇类号以及边界簇类号可以将各分区的簇类联系起来, 形成最终的聚类结果。元素在边界聚类及分区聚类中的不同类别也决定了在簇类合并过程中的不同处理方式。处理方式见表 1。

表 1 分区聚类合并方式

Table 1 Partition clustering		
边界分区类别	分区聚类类别	类别合并过程中的处理方式
噪声点	噪声点	依旧噪声点
	非噪声点	继续保留之前的类号
边界点	噪声点	与同类中其他分区聚类类号相同①
	非噪声点	继续保留之前的类号
	噪声点	与同类中其他分区聚类类号相同②
核心点	非噪声点	找出边界聚类同一个类中的所有分区聚类类号, 合并分区簇类类号在同一个边界簇类中的簇类③

由表 1 可知, 分区合并过程中仅需要处理①、②、③这 3 种情况; 其他类型的元素在合并过程中无需处理。

1) 如果元素边界聚类时属于非噪声点, 而分区聚类时属于噪声点, 则将元素合并到同一类中其他元素的所属的分区簇类, 如果整个边界簇类中所有元素的分区聚类都为噪音点, 则设置整个边界簇类中的元素的类号为 100 000+边界聚类的类 ID(保证每一个簇类的类号唯一)。①、②两种情况按此合并处理。

2) 如果元素边界聚类时属于核心点, 分区聚类时不属于噪声点, 则应该合并。③情况按此合并处理。

基于数据分区—聚类—合并的 DBScan 算法示意图见图 5。

3 雷电聚类结果

3.1 最小密度数与雷电聚集效应

雷暴聚类过程中聚集在一起的节点形成雷暴,

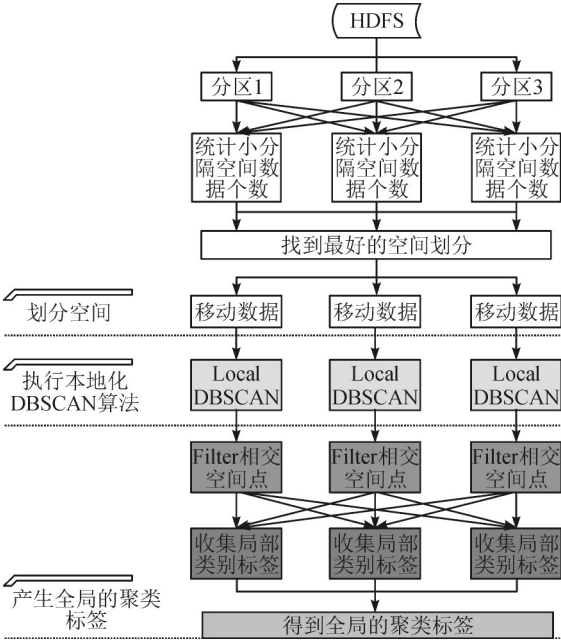


图 5 分区—聚类—合并算法示意图
Fig. 5 Schematic diagram of partition-cluster-merge algorithm

聚集成雷暴的点称为有效点, 其它的点就是噪声点, 而有效点的比例直接影响雷暴聚类的效果; 一次雷暴过程中, 雷电的空间位置应该是相对集中的, 雷暴应该数量少、密度大(包含的雷电个数多), 才符合雷暴的空间特性。文中采用有效点比例和平均雷暴聚集程度判定雷暴聚集效应, 见式(1)。

C = (Y/Z) × (Y/N) (1)

式(1)中: C 为 雷暴聚集效应; Y 为有效点个数; Z 为雷电总数; N 为雷暴数量。

当最小密度数减小时, 核心点增多, Y 增加, N 增加; 反之, 则相反。核心点的最小密度数 MinPts 直接影响最终的聚类效果。

文中以某段时间的 50 000 条雷电数据进行试验, 设定最小密度数为 5、10、20、40、80 进行对比, 结果见表 2。聚类结果显示当 MinPts=10 的时候雷暴的聚集效应最好。

表 2 5 组最小密度数对比试验

Table 2 Comparison test of 5 groups of minimum density

名称	最小密度数 MinPts				
	5	10	20	40	80
有效点数量 Y	38 643	35 463	27 075	19 942	8 644
雷点数量 Z	50 000	50 000	50 000	50 000	50 000
雷暴数量 N	1 190	723	450	244	70
雷暴聚集效应 C	25.1	34.8	32.6	29.4	21.5

3.2 聚类结果

根据文3.1节中的对比试验,设定核心点最小密度数 $\text{MinPts}=10$ 对全国雷电数据进行聚类,并将聚类结果放在 ArcGIS^[23]地理系统中进行展示。其中不同灰度代表不同的簇类(由于簇类类别太多,只能

使用若干种灰度循环使用,相邻簇类灰度不同,则代表不同簇类)。当天发生在湘鄂渝贵交界处地区的雷暴动态图见图6。该区域雷暴的完整分布图见图7,从图6、7中可以明显观测出雷暴产生、爆发、移动、消亡的全过程。

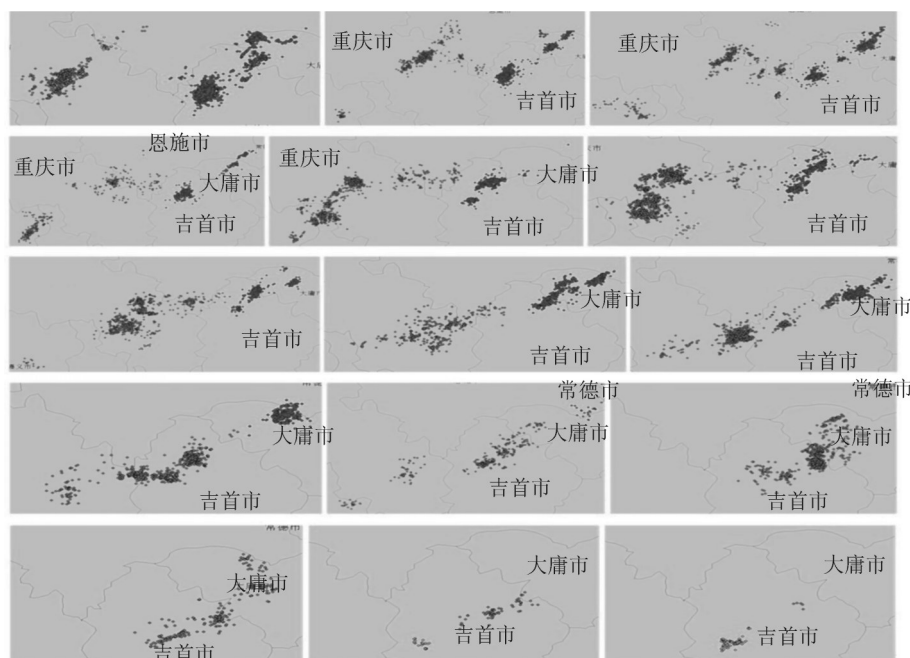


图6 一次完整雷暴的活动示意图

Fig. 6 Schematic diagram of a complete thunderstorm

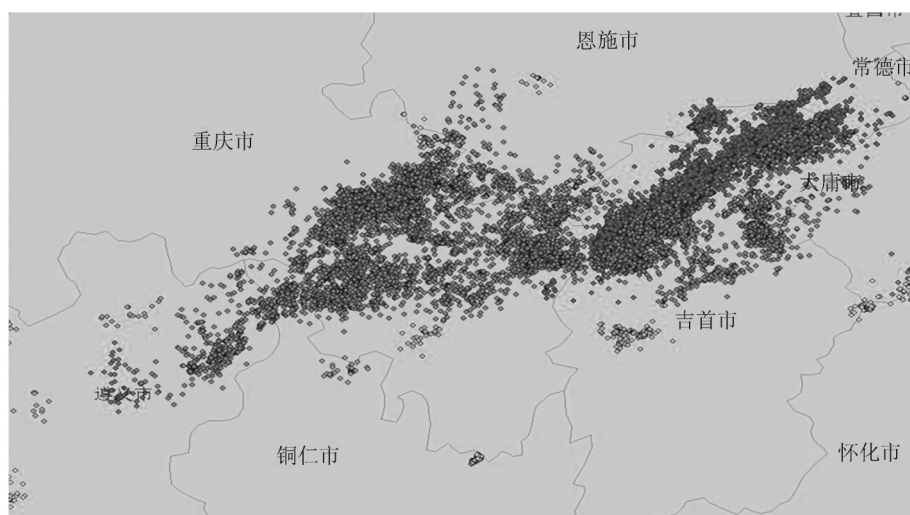


图7 雷电活动图

Fig. 7 Lightning activity map

4 雷暴活动规律统计

4.1 不同时间间隔对雷暴规律的影响

当其他参数不变时(邻域半径 0.05° , 最小密度数 10),取 10、15、20 min 这3组不同的时间间隔进行雷暴聚类对比试验,并进行雷暴生存时间统计,结果

见表3。由表3可以看出,当时间间隔为 10 min 时,雷暴的平均生存周期为 20.2 min,生存周期太短,不符合实际情况;当时间间隔为 20 min 时,雷暴的最长生存周期为 1 536.2 min,平均生存周期也较长,与雷暴的实际情况差距较大。因此文中取不同雷暴的时间间隔为 15 min,基本符合雷暴的实际生存时间,也为前文的参数选取提供佐证。

表3 不同时间间隔对比试验结果

Table 3 Comparison test results at different time intervals

min

时间间隔	雷暴平均生存周期	雷暴最长生存周期
10	20.2	304.7
15	59.4	812.1
20	83.3	1 536.2

4.2 雷暴统计分析

对湖北省 2011~2017 年的 10 303 个雷暴进行统计分析,结果见图 8~19。雷暴的平均生存时间约为 1 h,大部分雷暴生存时间在 30~80 min,雷暴面积的平均值约为 759 km²,大部分在 2 000 km² 以内。湖北省内恩施、宜昌、黄冈三地的雷暴数量较多,雷暴移动方向主要以南、西南、西、西北、北为主。夏季是雷暴的高发期,夏季雷暴总数占全年雷暴的 70% 以上。雷暴经过次数最多和持续时间最长的复奉线,其次是锦苏线,110 kV 的输电线路经过的雷暴数量最多。



图8 雷暴生命周期的最大值、最小值和平均值

Fig. 8 Maximum, minimum, and average values of the thunderstorm life cycle

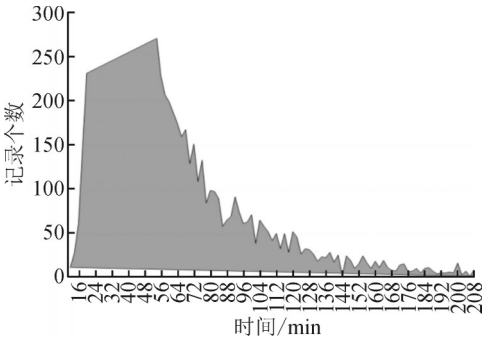


图9 雷暴的数量与雷暴的生存时间的关系图

Fig. 9 Relationship between the number of thunderstorms and the survival time of thunderstorms

5 结语

根据雷电探测站采集到的真实雷电数据,结合



图10 雷暴面积的最大值、最小值和平均值

Fig. 10 Maximum, minimum, and average of the area of the thunderstorm

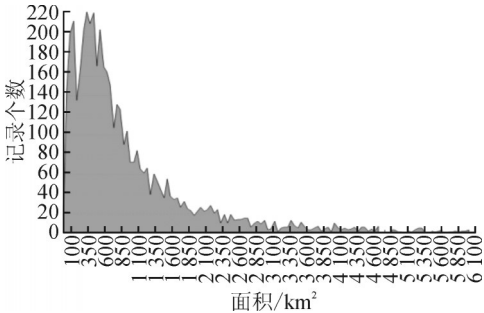


图11 雷暴的数量与雷暴的面积的关系图

Fig. 11 Diagram of the number of thunderstorms and the area of thunderstorms

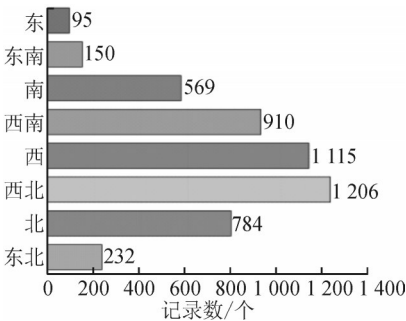


图12 每个方向上雷暴的数量的柱形表示

Fig. 12 Cylindrical representation of the number of thunderstorms in each direction

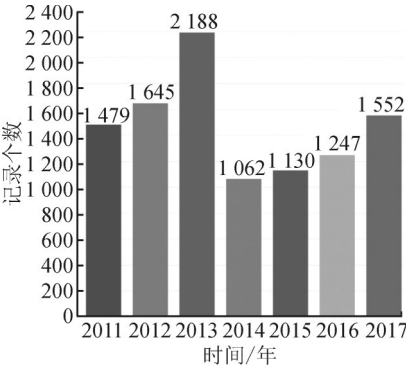


图13 所有雷暴在每个年度上的数量变化

Fig. 13 The number of all thunderstorms in each year

计算机聚类算法和雷暴的时空属性,编写基于 Spark 的改进 DBScan 算法,对全国范围内的海量雷电数据

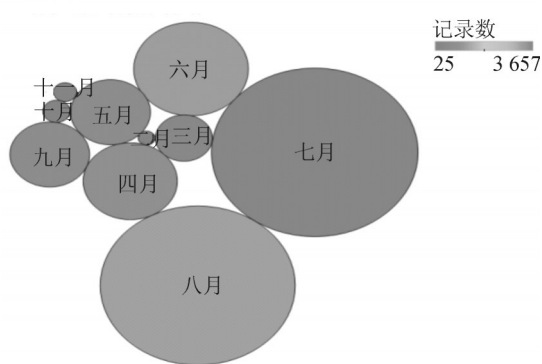


图14 所有雷暴数据在每个月度上的数量变化

Fig. 14 The number of all thunderstorm data changes on each month

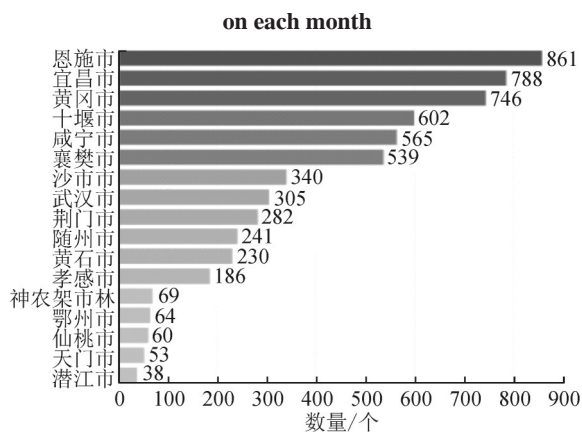


图15 雷暴在各个城市的数量

Fig. 15 The number of thunderstorms in each city

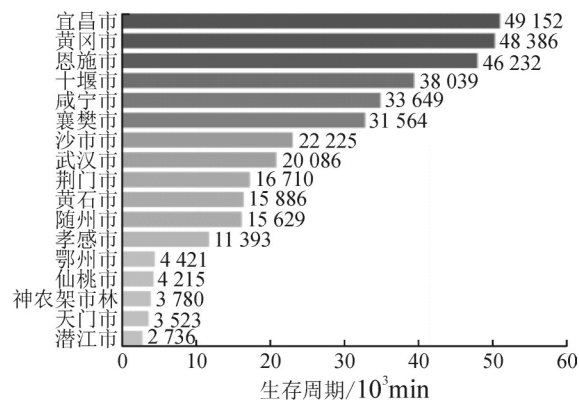


图16 雷暴在各个城市的总时间

Fig. 16 The number of thunderstorms in each city

聚类,挖掘所有雷暴,使用ArcGIS等软件绘制全国雷暴图,统计雷暴的活动规律,研究雷暴随时间产生、爆发、移动、消亡的全过程,为今后的雷暴预测工作奠定基础。

研究内容距离雷暴预测、雷暴防护等实际应用还很遥远,目前还不能产生明显的经济效益,雷暴分裂、合并、移动、预测等算法工作亟待完善。希望在后续的工作中可以进一步突破,向雷暴预测等实际应用更进一步。

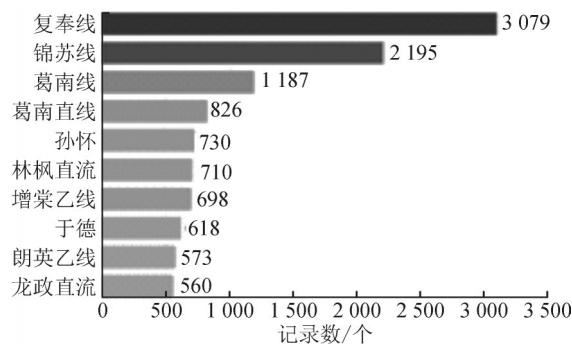


图17 线路上的雷暴数量排名前十

Fig. 17 Top ten thunderstorms on the line

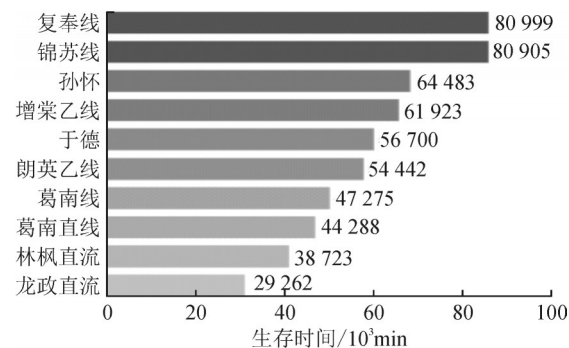


图18 线路上雷暴的总时间排名前十

Fig. 18 Top ten total time of thunderstorms on the line

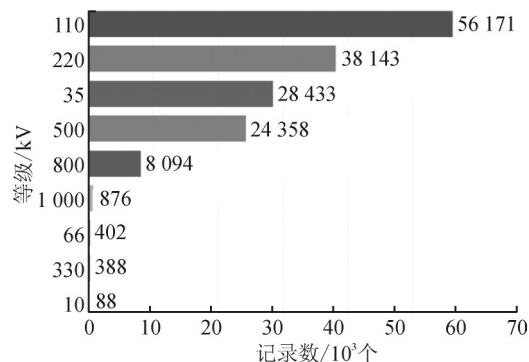


图19 电压等级的雷暴数量排名情况

Fig. 19 Ranking of thunderstorms at voltage levels

参考文献:

- [1] 王 蕾,钟思奕,曹 华,等.雷电灾害监测与防雷勘察技术研究[J].气象水文海洋仪器,2018,25(1):112-117.
WANG Lei, ZHONG Siyi, CAO Hua, et al. Research on lightning disaster monitoring and lightning protection exploration techniques[J]. Meteorological, Hydrological and Marine Instruments, 2018, 25(1):112-117.
- [2] 吴安坤.贵州省雷电灾害风险评价与区划研究[J].中国农业资源与区划,2018,39(2):88-93.
WU Ankun. Research on risk assessment and zoning of lightning disasters in Guizhou province[J]. Chinese Journal of Agricultural Resources and Regional Planning, 2018, 39(2):88-93.
- [3] 陈洪滨,朱彦良.雷暴探测研究的进展[J].大气科学,2012,36(2):411-422.
CHEN Hongbin, ZHU Yanliang. Review on the observation

- investigati-on of thunderstorms[J]. Chinese Journal of Atmospheric Sciences, 2012, 36(2):411-422.
- [4] 陈维江, 陈家宏, 谷山强, 等. 中国电网雷电监测与防护亟待研究的关键技术[J]. 高电压技术, 2008, 34(10): 2009-2015.
- CHEN Weijiang, CHEN Jiahong, GU Shanqiang, et al. Key Technologies of lightning detection and protection in china power grid[J]. High Voltage Engineering, 2008, 34(10):2009-2015.
- [5] 俞小鼎, 周小刚, 王秀明. 雷暴与强对流临近天气预报技术进展[J]. 气象学报, 2012, 70(3): 311-337.
- YU Xiaoding, ZHOU Xiaogang, WANG Xiuming. The advances in the nowcasting techniques on thunderstorms and severs convection[J]. Acta Meteorologica Sinica, 2012, 70(3): 311-337.
- [6] SUN J G, LIU J, ZHAO L Y. Clustering algorithm research [J]. Journal of Software, 2018, 19(1): 48-61.
- [7] ESTER M, KRIEGEL H P, SANDER J, et al. A density - based algorithm for discovering clusters inlarge spatial data-bases with noise[C]//Kdd. [S.l.]:[s.n.], 2017: 226-231.
- [8] JAIN A K, MURTY M N, FLYNN P J. Data clustering: A review [J]. ACM Computing Surveys(CSUR), 2017, 31(3): 264-323.
- [9] ZHANG T, RAMAKRISHNAN R, LIVNY M. BIRCH: An efficient data clustering method for very large databases[C]// ACM Sigmod Record. [S.l.]: ACM, 2017: 103-114.
- [10] WANG W, YANG J, MUNTZ R. STING: A statistical information-grid approach to spatial data mining[C]//VLDB. [S.l.]:[s. n.], 2016: 186-195.
- [11] 仰孝富, 齐建东, 吉鹏飞, 等. 一种CF树结合KNN图划分的文本聚类算法[J]. 计算机工程与应用, 2013, 23(1): 114-119.
- YANG Xiaofu, QI Jiangdong, JI Pengfei, et al. New text clustering algorithm based on CF tree and KNN graph partition[J]. Computer Engineering and Application, 2015, 51(6):114-119.
- [12] 包晓安, 鲍超, 滕赛娜, 等. 基于DBSCAN算法的测试用例优化方法[J]. 计算机系统应用, 2018, 27(3): 105-111.
- BAO Xiao'an, BAO Chao, TENG Saina, et al. Test case optimization method based on DBSCAN algorithm[J]. Computer Systems & Applications, 2018, 27(3):105-111.
- [13] 王佳玉, 张振宇, 褚征, 等. 一种基于轨迹数据密度分区的分布式并行聚类方法[J]. 中国科学技术大学学报, 2018, 48(1): 47-56.
- WANG Jiayu, ZHANG Zhenyu, CHU Zheng, et al. A trajectory data density partition based distributed parallel clustering method[J]. Journal of University of Science and Technology of China, 2018, 48(1):47-56.
- [14] 王童童, 荣垂田, 卢卫, 等. 分布式图处理系统技术综述[J]. 软件学报, 2018, 29(3): 569-586.
- WANG Tongtong, RONG Chuitian, LU Wei, et al. Survey on technologies of distributed graph processing systems[J]. Journal of Software, 2018, 29(3):569-586.
- [15] 周鹏飞, 郭乔进, 胡杰. 基于Hadoop平台的视频处理系统设计[J]. 信息化研究, 2016, 42(6):64-67.
- ZHOU Pengfei, GUO Qiaojin, HU Jie. Video processing system based on hadoop platform[J]. Informatization Research, 2016, 42(6):64-67.
- [16] ZAHARIA M, CHOWDHURY M, FRANKLIN M J, et al. Spark: Cluster computing with working sets[C]//Hot Cloud, 2010. Boston, US:[s.n.], 2017:10.
- [17] 冯万兴, 朱晔, 郭钧天, 等. 基于改进的DBSCAN方法和多项式拟合的雷电短时预测[J]. 计算机工程与科学, 2014, 36(10):2028-2033.
- FENG Wanxing, ZHU Ye, GUO Juntian, et al. Lightning forecast based on the improved DBSCAN and polynomial fitting[J]. Computer Engineering & Science, 2014, 36(10): 2028-2033.
- [18] HUFFINES G R, ORVILLE R E. Lightning ground flash density and thunderstorm duration in the continental United States:1989-1996[J]. Journal of Applied Meteorology, 1999, 38(7):1013-1019.
- [19] 孙凌, 周筠珺, 郭在华. 雷暴持续时间与地闪活动的预报方法研究[J]. 气象科学, 2012, 32(2):182-187.
- SUN Ling, ZHOU Yunjun, GUO Zaihua. Research on the forecasting method for thundrtstorm duration and lightning activities[J]. Journal of the Meteorological Sciences, 2012, 32(2):182-187
- [20] 空军司令部气象研究室. 国外关于积雨云特性的研究[J]. 气象科技, 1975(10):4.
- Air Force Command Department of Meteorological Research. Foreign research on the characteristics of cumulonimbus clouds[J]. Meteorological Science and Technology, 1975(10):4.
- [21] 刘希文, 李得勤, 韦惠红, 等. 2013年湖北地区两次雷暴大风过程环境及雷达回波特征[J]. 气象与环境学报, 2018, 34(1): 1-10.
- LIU Xiwen, LI Deqin, WEI Huihong, et al. Environmental and radar echoes characteristic analysis of two thunderstorm gale processes in Hubei province in 2013[J]. Journal of Meteorology and Environment, 2018, 34(1):1-10.
- [22] 冯玲, 刘克剑, 唐福喜, 等. 一种基于网格查询的改进DBSCAN算法[J]. 西华大学学报(自然科学版), 2016, 35(5): 25-29.
- FENG Ling, LIU Kejian, TANG Fuxi, et al. Improvements of DBSCAN algorithm based on grid[J]. Journal of Xihua University(Natural Science), 2016, 35(5): 25-29.
- [23] 张龙, 汪新庆, 刘岳, 等. 一种MapGIS向ArcGIS地图数据转换的方法[J]. 测绘科学, 2018, 39(7): 146-149.
- ZHANG Long, WANG Xinqing, LIU Yue, et al. A method of map data conversion from MapGIS into ArcGIS[J]. Science of Surveying and Mapping, 2018, 39(7):146-149.
- 高攀(1989—), 男, 硕士, 工程师, 主要从事电网雷电监测与预警技术、大数据技术研发工作。
- 田浩(1984—), 男, 硕士, 工程师, 主要从事电网雷电监测与预警技术、计算机系统、云计算、大数据研发工作。
- 李健(1986—), 男, 博士, 工程师, 主要从事输电线路外绝缘相关领域研究。