

共享单车入栏检测算法研究与实现^{*}朱江淼¹ 张 菁¹ 黄 艳² 金森林² 高春柳²

(1.北京工业大学 信息学部 北京 100124; 2.北京市计量检测科学研究院 北京 100029)

摘 要: 正确引导共享单车用户规范停车,是政府管理的重要举措。为此提出了利用聚类的原理、基于 K 均值的共享单车入栏检测算法。首先模拟产生服从高斯分布的随机单车定位数据;其次对定位数据进行聚类分组,不断迭代向均值移动确定每簇的中心点;最后将数据量最多的簇中心点作为单车的停放位置,对其进行入栏检测分析。用实际测得的单车定位数据验证入栏检测准确率,结果表明,在符合政府管理要求的条件下可准确检测单车是否入栏,正确率高达 80%~100%,具有较好的实用价值。

关键词: 共享单车;围栏;GNSS 高精度接收机;K 均值聚类;入栏检测算法

中图分类号: TN98 **文献标识码:** A **国家标准学科分类代码:** 510.40

Research and implementation on algorithm for entry
fence detection of shared bicyclesZhu Jiangmiao¹ Zhang Jing¹ Huang Yan² Jin Senlin² Gao Chunliu²

(1. Information Department, Beijing University of Technology, Beijing 100124, China;

2. Beijing Scientific Research Institute of Metrology and Testing, Beijing 100029, China)

Abstract: In order to standardize parking, city managers mark parking areas with rectangular solid white lines on the pavement to remind riders to park their bicycles within the fence. A new method uses the K-means clustering principle for fence of shared bicycles is proposed. First, the location data of Gauss distribution are simulated. Second, to get multiple clusters, the data of each location are clustered, and then by means of continuous iteration, the final center of each cluster is determined. Finally, the data of the largest cluster of central points for entry fence detection are analyzed. The accurate rate of entry fence detection is verified by the actual measurement of location data of shared bicycles. The results show that the algorithm can accurately detect whether the shared bicycles enter the fence under the prescribed conditions, accuracy is as high as 80%-100%, and has high practical value.

Keywords: shared bicycles; fence; high precision receiver of GNSS; K-means clustering; algorithm for entry fence detection

0 引 言

共享单车发展迅速,便捷群众,但是乱停乱放现象严重。为规范单车停放,合理分配停车区域,停放围栏的管理手段应运而生,而能否准确检测出单车的入栏状况,已成为政府引导用户规范停车的关键。从另一方面看,为了节约成本,共享单车内装的 GPS 接收机十分廉价,定位精度低、误差大,无法获知非常准确的共享单车的真实停放位置,也就无法准确判定单车是否入栏,因此,共享单车入栏检测的研究成为政府管理的一个重要内容。

考查快速而强大的机器学习(machine learning, ML)

算法,发现聚类算法作为其主要的方法之一,越来越受到人们的关注,近年来中外学者对其中的 K 均值算法多有研究。所谓聚类(clustering)是在寻找数据集的“自然结构”过程中,把数据对象区分为不同的集群(cluster),保证相类似的对象是在相同的集群中,不类似的对象则被分到不同的集群中^[1-2]。K-means 算法是一种以平均值作为聚类中心的分割聚类方法,简单而且快速,解决了许多实际问题^[3]。

针对共享单车入栏检测的问题,首先要确定围栏的位置,可以利用 GNSS 高精度接收机获取围栏的位置参数值;其次在等精度测量条件下,通过对单车某一型号接收机的

多次测量,采集单车定位数据。共享单车的定位数据具有相关性和一定的规律性,利用聚类的思想、基于K均值算法可以有效地估计出单车的真实位置;最后本文研制出了共享单车入栏检测算法。结果表明,应用此方法可较为准确地检测出单车是否入栏,在符合政府管理要求的条件下,正确率高达80%~100%,提高了入栏检测的准确率,具有很好的实用价值。

1 整体研究方案概述

整体的研究方案流程如图1所示。

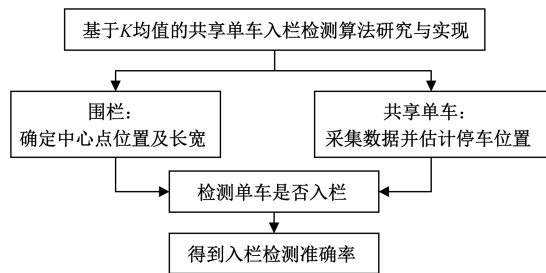


图1 研究方案流程

本文针对共享单车入栏检测问题进行研究,整体的研究方案分为3个部分。首先,确定围栏的位置信息,具体方法是确定围栏的中心点位置,在政府管理要求范围内选择不同的围栏的长和宽,从而明确围栏所在位置;其次,采集共享单车定位数据,通过算法有效地估计出单车停放的准确位置;最后,利用算法检测单车是否入栏,最终得到入栏检测的准确率。

2 共享单车入栏检测算法研究

2.1 围栏位置确定

所谓围栏,即政府为规范单车停放、合理分配停车区域,在地面用实线所划出的用户可见的地理边界。根据政府的管理要求和实际情况,多数围栏的尺寸长为6~10 m,宽为1.5~2 m,且目前政府需要管理的围栏的位置均面向如地铁口、商超门口等较为开阔的公共区域。这些公共区域受周边建筑的干扰较小,多为开阔场地,可以在一定精度下忽略多径干扰等因素的影响。

为研究针对不同长宽的围栏的单车入栏检测准确率,在算法仿真研究时,利用GNSS高精度接收机确定围栏中心点的位置,例如用接收机测得围栏中心点的经度值为116.632 317°,纬度值为39.892 849°,视为平面坐标(0,0)点。则长10 m、宽2 m的围栏位置随之确定,得到了栏框四个边界点的经纬度值分别为(116.632 255,39.892 858)、(116.632 255,39.892 841)、(116.632 383,39.892 859)、(116.632 382,39.892 842)(单位为°)。如此,再变换不同的围栏长宽可易得其边界点的对应经、纬度。

GNSS接收机是GNSS系统用户终端,主要由天线单

元、接收主机单元两部分组成。图2所示给出了接收机的原理基本结构。

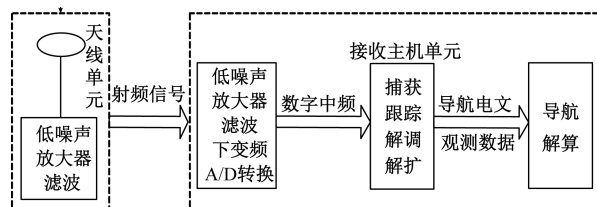


图2 接收机的原理结构^[4]

本文所采用的GNSS高精度接收机为司南M300C型接收机,是一种高度灵活的分体式接收机,采用北斗与GPS双星四频GNSS模块,支持GPS信号和L2C现代化改造后的GPS信号,可单北斗或单GPS定位,也可联合RTK定位,预留支持以太网、GSM/GPRS通信,且具备2个电源接口和2个串口,满足更多需求的使用^[5]。主机外观如图3所示。



图3 司南M300C型接收机

该接收机主要由射频单元、基带单元组成,具体来说,射频单元完成各频点导航信号的下变频处理;基带单元主要完成GPS-L1/L2P以及BDS-B1/B2/B3导航信号的捕获、跟踪、导航电文解调、伪距和载波相位测量以及定位解算等任务^[6]。内部储存为100 M,可设置自动记录原始数据,且内置2 000 V光电隔离,对接收机有效进行过流过压保护,使用AutoBas技术,开机即可快速实现参考站的配置,450 MHz超高频无线电台适用基准站和流动站作业,确保现场工作的电台信号覆盖范围以及数据安全性,通信协议支持透明传输、TT450S^[7]。时间同步支持1 PPS实时输出,支持PJK平面坐标输出,无需第三方软件做投影转换。该接收机支持自动差分,差分格式支持RTCM2.X、RTCM3.X以及CMR,可远程设置、下载、查看数据等^[8]。

这里要区分入栏率及入栏检测的准确率两个概念。入栏率系指单车存入栏内的比率,而所谓入栏检测准确率,是指正确检测出存入栏内单车的比率,也就是说若单车停在栏内,检测结果也在栏内,说明检测结果准确;若100辆单车都在栏内,则入栏率为100%,如果检测结果是99辆在栏内,则其检测准确率为99%。另外,由于单车体积比较大,可能出现部分入栏部分在栏外的情况,而单车的定位数据是由接收机给出的,因此,本文所讨论的入栏检测的准确

率,是针对单车所配备的接收机是否入栏而言的。

2.2 单车位置估计算法

检测单车是否入栏,其关键问题是获知单车准确的地理位置。实际操作时,单车每次落锁时能连续提供 28 个以经纬度数据表示的地理位置。理论上,这 28 个数据是服从正态分布的随机数据,应是一致的,但由于单车配备的接收机很廉价,导致这 28 个数据分散性较大,如何通过这 28 个数据推算出较为准确的单车实际存放位置,是本文要解决的一个问题。这实际上就是从测量结果和测量误差估计真值,是一个反问题。

1) K 均值聚类算法

所谓聚类分析,实际上就是针对某个数据集合,寻找一个数据,该数据对集合中的所有数据具关联性^[9];而 K 均值聚类算法的关联性系指集合中的任何一个数据到 K 数据(初始聚类中心)的几何距离,K 数据是随机选定的,它可以是一个或任意个,一般要远小于数据集合的总个数;以到聚类中心几何距离均值最小为准则,确定下一个聚类中心,如此进行直到聚类中心不再变化为止^[10]。K 均值是最普通的聚类算法,将数据聚类成不同的组,其中 K 代表聚类的类别数。K 均值算法是一个迭代过程,假设需要将数据聚类成 n 个组/簇,其方法为:

(1) 首先选择 K 个随机的点,称为聚类中心;

(2) 对于数据集中的每一个数据,按照距离 K 个中心点的距离,将其与距离最近的中心点关联起来,与同一个中心点关联的所有点聚成一类^[11];

(3) 计算每一个组的平均值,将该组所关联的中心点移动到平均值的位置^[12]。

重复步骤(2)~(3)直至中心点不再变化。

K-means 算法中包含几个循环步骤,赋值步骤是第 1 个循环,也就是针对各个数据点 i ,通过计算得知其应该属于哪一类;不断重复移动聚类中心是第 2 个循环,也就是针对各类 K 值,更新计算得到此类的质心^[13]。

K-means 算法具有最小化原则,要使得全部数据点与它们已关联好的聚类中心点间的距离和达到最小^[14]。所以,每一次迭代都是在减小上述距离之和,否则便是出现了错误。

2) 基于 K 均值的单车位置估计算法

由于共享单车每次落锁是一次独立事件,而二项分布(binomial distribution, BD)就是重复 N 次的独立事件的伯努利实验。因此为仿真得到最为随机的单车停放情况,取长 10 m,宽 2 m 的围栏,以其几何中心作为坐标原点(0, 0),围绕围栏周围(包括栏内与栏外),首先采用二项随机分布产生 1 个单车定位数据点,模拟共享单车的真实停放位置;其次在这个位置处产生包含 28 个数据的正态分布,从而得到一组 28 个服从正态分布位置点的模拟单车定位数据。

下面,用 K 均值聚类算法对这一组包含 28 个位置点的模拟单车定位数据进行入栏检测,取这 28 个定位数据,

以聚类数 $K=2$ 为例说明算法流程,即将单车定位数据分为两类,选择 2 个随机的点作为聚类中心,用实心点标出,如图 4 所示。

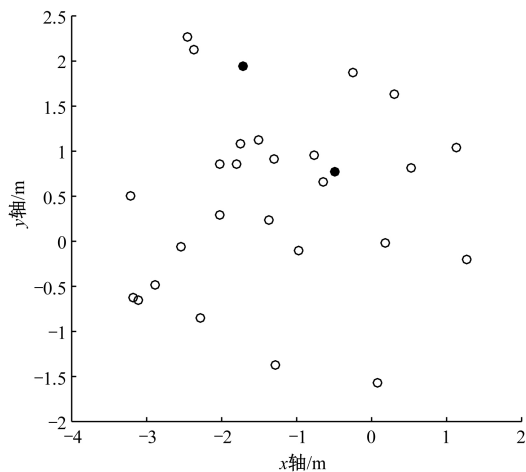


图 4 一组模拟单车定位数据

针对这组定位数据集中的各个数据点,计算出其与 K 个中心点的距离值,关联其与距离最为接近的一个中心点,将所有与相同中心点关联的数据点聚类分为一类^[15]。将聚好的两类(实心点一类,空心圆圈一类)及各自的聚类中心标出,如图 5 所示。

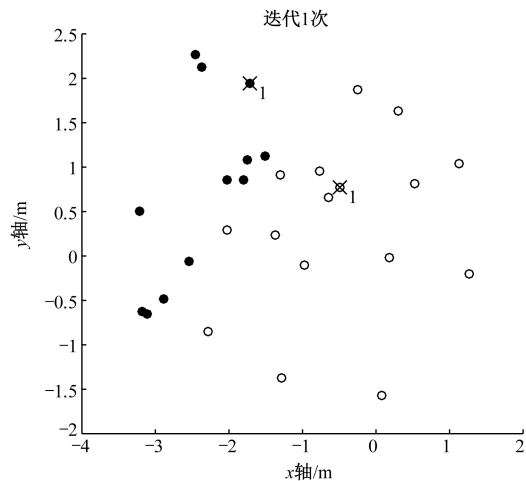


图 5 一组单车定位数据聚类情况($K=2$)

计算得到各类的均值,把此类已关联好的中心点移至均值的位置处。再重复地迭代进行计算并且不断移动,一直到中心点不再发生改变为止^[16]。迭代 3 次,显示结果如图 6 所示。

重复迭代至中心点不再变化为止,最终两类中心点如图 7 所示。

最后,对每一类所包含的数据量进行统计,取数据量最多的一类的聚类中心,作为最接近共享单车真实停放位置的定位数据,接下来对该中心点进行入栏检测。由图 7 可

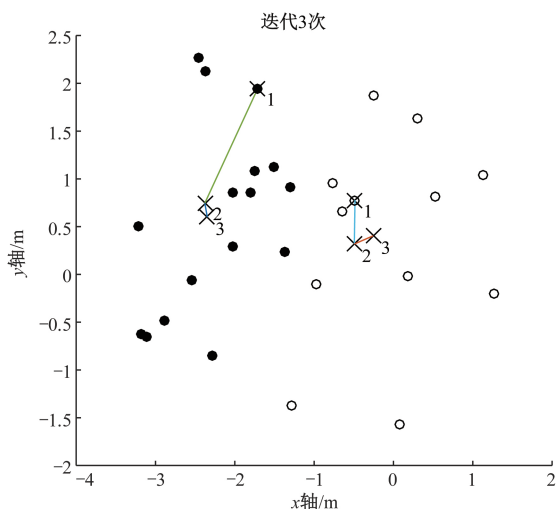


图6 迭代3次中心点移动情况

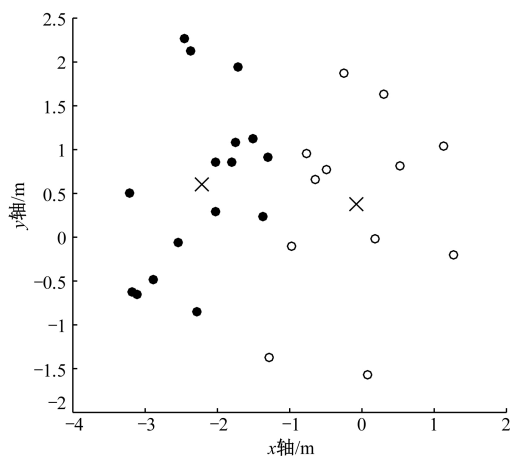


图7 最终聚类中心效果

知,实心类的数据个数大于空心类的数据个数,因此,实心类的聚类中心被确认为单车的停放位置,接下来检测该位置是否入栏。

综上,本文研制的基于K均值的入栏检测算法流程如图8所示。

2.3 入栏检测算法

取长10 m、宽2 m的围栏,判断单车位置是否在围栏内。分别计算单车所在位置与围栏4条边的距离,若该位置点同时满足大于左栏框、小于右栏框,且大于下栏框、小于上栏框的边界值的条件,则将输出值赋为1,说明该单车位于围栏内;若该位置点不同时满足上述条件,则将输出值赋为0,说明该单车不在围栏内。最后,将判断是否入栏的结果与之前模拟好的共享单车真实停放位置相比较,验证入栏检测结果是否正确。算法流程如图9所示。

为研究算法,本文采用二项随机分布产生50个单车定位数据点,模拟共享单车最为随意无序的真实停放位

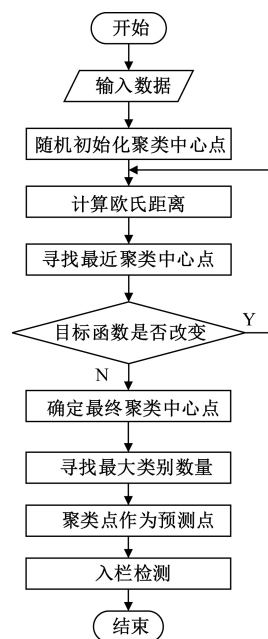


图8 基于K均值的单车位置估计算法流程

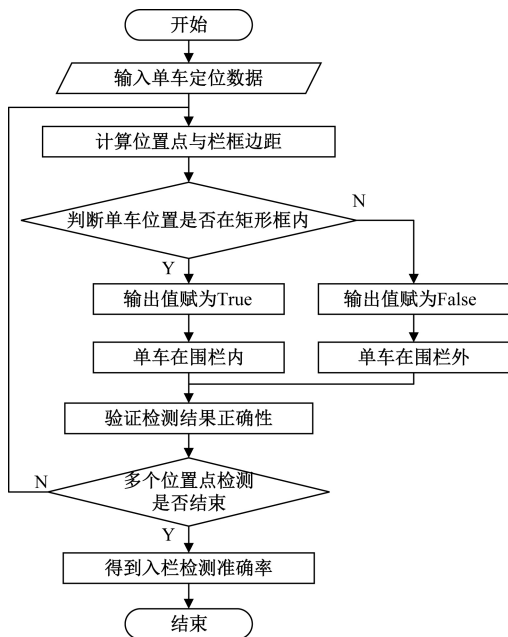


图9 入栏检测算法流程

置;其次使这50个模拟单车真实停放位置的定位数据点,分别呈每点包含28个数据的正态分布,从而得到50组且每组包含28个服从正态分布位置点的模拟单车定位数据。

应用上述基于K均值的入栏检测算法,取得输出值为1或0,依次与之前模拟好的共享单车真实停放位置相比较,验证入栏检测结果是否准确。统计全部50组单车定位数据的验证情况,从而得到入栏检测的准确率。

3 入栏检测准确率影响因素分析

3.1 算法参数的影响

K 均值聚类算法中,有两个参数会影响最后的结果,一个是 K 值,一个是迭代次数。本文分别取 $K=1,2,\cdots,10$,针对每一个 K 值分别迭代 10 次、100 次,对 50 组单车定位数据的入栏检测准确率进行分析。迭代次数超过 100 次会引起过拟合现象,入栏检测的准确率会降低;对于 $K=1$,迭代 10 次和 100 次均能使入栏检测的准确率达到 100%。因此针对 K 值只需迭代 10 次就可使聚类中心点不再变化,达到算法的最优效果。结果如表 1 所示,可见当取 $K=1$ 时,即聚类中心点接近均值点的位置被确定为单车的停放位置时,入栏检测的准确率高达 100%。

表 1 不同 K 值对应的入栏检测准确率

K 值	不同迭代次数准确率/%	
	10	100
1	100	100
2	92	90
3	82	78
4	74	76
5	84	78
6	78	82
7	86	78
8	76	74
9	80	80
10	78	70

3.2 围栏大小的影响

围栏的大小对入栏检测的准确率也有一定影响。本文分别取长度为 6~10 m,宽度固定为 2 m 的围栏,用同样的 50 组单车定位数据应用算法进行入栏检测,针对每一个的 K 值迭代 10 次,得到不同 K 值对应的入栏检测准确率如 2 表所示。

表 2 不同长度的围栏对应的入栏检测准确率

K 值	不同长度准确率/%				
	10 m	9 m	8 m	7 m	6 m
1	100	100	100	92	80
2	92	90	86	84	76
3	82	86	68	78	78
4	74	74	78	66	75
5	84	74	76	68	64
6	78	84	70	74	75
7	86	76	74	70	70
8	76	78	86	74	68
9	80	80	70	62	65
10	78	70	70	68	60

取不同宽度的围栏进行入栏情况检测,对入栏检测的准确率也有一定影响。分别取宽度为 2、1.5、1 m,长度固定为 10 m 的围栏,用同样的 50 组单车定位数据应用算法进行入栏检测,针对每一个的 K 值迭代 10 次,得到不同 K 值对应的入栏检测准确率如表 3 所示。

表 3 不同宽度的围栏对应的入栏检测准确率

K 值	不同宽度准确率/%		
	2 m	1.5 m	1 m
1	100	100	80
2	92	74	64
3	82	78	66
4	74	70	68
5	84	74	62
6	78	78	70
7	86	74	62
8	76	72	66
9	80	72	70
10	78	76	70

由表 1~3 可知,当取聚类中心数 K 值为 1,迭代次数为 10 时,按照政府管理规范,对于常规长 6~10 m 且宽 1.5~2 m 的围栏,本文提出的基于 K 均值的入栏检测算法,对共享单车的定位数据入栏检测的正确率达 80%~100%。如若在长 6~10 m 且宽 1.5~2 m 之间任意变换不同长宽的组合,入栏检测的正确率依然在 80%~100% 范围内,此处不再赘述。

4 算法验证与评估

为了验证基于 K 均值的入栏检测算法的准确性和可靠性,本文进行了实际测量,并用上述算法对测量数据进行了计算,得到了实际的入栏检测率。首先用 GNSS 高精度接收机确定了围栏的中心点位置,将该位置作为基准点。实验测得基准点的经纬度值为 $116.632\ 295^{\circ}$,纬度值为 $39.892\ 793^{\circ}$,变换围栏的长宽,进行入栏检测算法的验证。在等精度测量条件下,在围栏的不同位置对某一型号的单车接收机进行 100 次测量,从而采集到共享单车 100 组实际的位置信息。其中每次测量包含一次开锁、落锁操作,得到 28 个以经纬度数据表示的地理位置,视为 1 组定位数据。通过本文的基于 K 均值的入栏检测算法,估计出单车每次大概的真实停放位置,对这 100 个定位进行入栏检测,验证该算法能否对这些真实的单车定位数据得到较高的入栏检测的准确率。

基于这 100 组真实的共享单车定位数据,取 $K=1$,针对 K 值迭代 10 次。分别取长度为 6~10 m,宽度固定为 2 m 的围栏,得到的入栏检测准确率如表 4 所示。

分别取宽度为 2、1.5、1 m,长度固定为 10 m 的围栏,用同样的 100 组真实单车定位数据应用算法进行入栏检测的准确率如表 5 所示。

表4 不同长度的围栏验证入栏检测准确率

K 值	不同长度准确率/%				
	10 m	9 m	8 m	7 m	6 m
1	100	100	100	92	81
2	90	89	86	88	78
3	84	82	80	77	78
4	89	84	88	78	73
5	81	81	73	81	68
6	91	84	83	80	75
7	81	87	84	74	71
8	81	87	83	80	67
9	86	79	80	80	67
10	82	81	76	75	62

表5 不同宽度的围栏验证入栏检测准确率

K 值	不同宽度准确率/%		
	2 m	1.5 m	1 m
1	100	100	80
2	90	76	76
3	84	79	63
4	89	80	73
5	81	78	74
6	91	78	72
7	81	74	65
8	81	77	66
9	86	77	70
10	82	72	61

综上所述,当取聚类中心数 K 值为 1 时,对于常规长 6~10 m、宽 1.5~2 m 的围栏,本文提出的基于 K 均值的入栏检测算法,对共享单车的定位数据入栏检测的正确率达 80%~100%。这表明本文算法正确可靠,对判断共享单车是否停入围栏内的分析具有较好的实用价值,在符合政府管理的条件下,提高了入栏检测的准确率。

5 结 论

本文针对普通而廉价的共享单车内置接收机定位精度低,无法准确地检测出单车是否入栏的问题,研制出基于 K 均值的共享单车入栏检测算法,能较为准确地检测出共享单车是否停入栏内。经分析和实测数据验证表明,取 K 值为 1 时,对于常规长 6~10 m、宽 1.5~2 m 的围栏,入栏检测准确率达 80%~100%。其中,长度为 7 m、6 m 的围栏检测准确率为 80%~92%;长 8~10 m、宽 1.5~2 m 的围栏,入栏检测准确率高达 100%,具有较好的实用价值。

参考文献

- [1] 何明亮,陈泽茂,黄相静,等.基于改进 K 均值聚类的入侵检测算法研究[J].计算机与数字工程,2017,45(6):1145-1149.
- [2] 左进,陈泽茂.基于改进 K 均值聚类的异常检测算

法[J].计算机科学,2016,43(8):258-261.

- [3] 周爱武,于亚飞. K -means 聚类算法的研究[J].计算机技术与发展,2011,21(2):62-65.
- [4] 马文忠,李林欢,江丽丽.基于载波相位差分的北斗/GPS 双模定位系统研究[J].测绘工程,2015,24(9):25-30,33.
- [5] RAMON L L V, GARCIA J G, RONCAGLIOLO P A, et al. An experimental L1/L2 GNSS receiver for high precision applications[J]. IEEE Latin America Transactions, 2013, 11(1): 48-53.
- [6] 郑伦贵,尤政,张高飞,等.GNSS 接收机的一种快速捕获方法研究[J].仪器仪表学报,2014,35(4):807-813.
- [7] LI X, GE M, DAI X, et al. Accuracy and reliability of multi-GNSS real-time precise positioning: GPS, GLONASS, BeiDou, and Galileo [J]. Journal of Geodesy, 2015, 89(6):607-635.
- [8] 赵艳芝,张瑞敏.司南导航发布第一代高精度 GNSS 基带芯片[J].电子工业专用设备,2014,43(12):59-60.
- [9] 胡艳维,秦拯,张忠志,等.基于模拟退火与 K 均值聚类的入侵检测算法[J].计算机科学,2010,37(6):122-124.
- [10] 王林,吴海桥,郑友石.一种改进的 K 均值聚类算法[J].科技信息,2010(32):136-137.
- [11] CELEBI M E, KINGRAVI H A, VELA P A, et al. A comparative study of efficient initialization methods for the K -means clustering algorithm [M]. Pergamon Press, Inc.2013:6-121.
- [12] COATES A, NG A Y. Learning feature representations with K -means [J]. Lecture Notes in Computer Science, 2012, 7700: 561-580.
- [13] KANG S H, SANDBERG B, YIP A M, et al. A regularized K -means and multiphase scale segmentation[J]. Inverse Problems & Imaging, 2017, 5(2): 407-429.
- [14] 赵勇,李怀宇.基于通用距离测量的机器学习方法用于图像分类和聚类[J].电子测量技术,2017,40(9):136-140.
- [15] 孙振.基于机器学习的网络流量特征选择[J].电子测量技术,2017,40(7):131-136.
- [16] 李金涛,艾萍,岳兆新,等.基于 K -means 聚类算法的改进[J].国外电子测量技术,2017,36(6):9-13,21.

作者简介

朱江森,副教授、硕士生导师,主要研究方向为数字信号处理、现代测量、时域测量和数字去卷积应用。

E-mail:zhujiangmiao@bjut.edu.cn

张菁,硕士研究生,主要研究方向为现代信号处理及其在计量测试中的应用、高精度 GNSS 接收机的应用。

E-mail:zhangjinglilybei@163.com