**北京邮电大学**

**本科毕业设计论文**



**名称：基于腾讯定位数据的异常事件检测算法**

**学院： 信息与通信工程学院**

**班级： 2014211112**

**姓名： 林文鼎**

**学号： 2014210328**

**班内序号： 07**

**指导教师： 别红霞**

**验收日期： 2018年5月7日**

**摘 要**

摘要还没写

**关键词：**关键词1；2；3

目 录

1. 绪论 1

1.1 课题背景 1

1.2 异常检测研究现状 1

1.2.1 基本概念与挑战 1

1.2.2 异常检测算法分类（需完善） 1

1.3 论文主要工作 2

1.4 论文章节安排 2

2. 异常检测算法基础 2

2.1 时序曲线离群点检测 2

2.2 时序曲线预测 2

2.3 2

3. 定位数据 2

3.1 腾讯定位数据的形式 2

3.2 数据分析及应用 2

3.3 定位数据预处理及异常分析 3

4. 基于曲线分析的定位数据异常检测 3

4.1 基于小波的异常检测算法 3

4.1.1 离散小波变换 3

4.1.2 结果分析 3

4.2 基于最大似然法的异常检测算法 3

4.2.1 最大似然法与3西格玛 3

4.2.2 结果分析 3

4.3 基于差分的异常检测算法 3

4.3.1 差分算法 3

4.3.2 结果分析 3

4.4 局部异常因子检测算法 3

4.4.1 局部异常因子 3

4.4.2 结果分析 3

4.5 混合异常检测算法 4

4.5.1 传统方法的优化 4

4.5.2 结果分析 4

4.6 曲线异常检测算法总结 4

5. 定位数据的预测分析 4

5.1.1 生成模型与判别模型 4

5.1.2 生成模型的预测 4

5.2 基于动态神经网络的定位数据预测 4

6. 基于区域的定位数据异常检测 4

6.1 基于图像的异常区域检测 4

6.1.1 差分分析法 4

6.1.2 小波分析法 4

6.2 多异常天的检测 4

7. 总结与展望 4

7.1 内容总结 4

7.2 未来展望 4

1. 绪论

本章主要介绍了定位数据的异常事件检测的课题背景及其研究意义，其次介绍了异常检测这一领域的基本概念及常用算法分类，最后描述了论文的主要研究工作和章节安排。

1.1 课题背景

随着GPS定位，传感器网络和无线通信等应用的日益普及，越来越多的终端定位数据被收集和保存在应用服务器。除了定位数据本身所体现的空间特征，在相同的空间位置不同的时间点上进行记录还可以得到定位数据的时序特征。而通过分析某片区域上的时序定位数据，可以得到该区域上人口密度的变化特征，例如从北京市滴滴打车的定位数据可以明显看出以下特征：车辆在早高峰时将大量人群运送至各大工作区（例如中关村和西二旗），而在晚高峰时又将人群从工作区运送回家，这种形式的定位数据反应出了北京市的日人口密度变化特征。

在这种区域性的时间序列中，也会出现这样的观测点，它们较以往同时段的数据会有一个明显的波动，这些点被称为异常点或离群点。分析这些异常点也是一个很重要的课题，通常对于区域时序特征模型的建立，这些异常点是应当被剔除的噪声，它们会对模型的预测功能产生极大的阻碍。但同时异常点也可以作为一些突发事件（如异常气象，交通管制等）的判断因素，在异常事件发生时，区域的时序定位数据会在某一个时间间隔中出现较大的落差，即异常波动，通过异常检测算法将上述波动检出并分析，可以使有关部门及时做好准备。

1.2 异常检测研究现状

1.2.1 基本概念与挑战

异常是指数据特征不符合该特征一般所隶属区间的现象。寻找异常挑战主要来源于两个方面：首先，“异常”通常情况下只是一个定性的概念，偏离正常数据多少可以被界定为异常没有一个定量的比例数值，那么对于那些处于异常非异常边界线附近的异常数据来说，完全可以把边界线略微移动，使其能被归类为正常的数据；再者，用于划定数据特征正常区间的正常样本中有时也会存在异常数据，导致划定边界线偏差或是训练出的预测模型不准确。同时，考虑到正常的数据量远大于异常数据，使用机器学习的方法进行训练时很容易过拟合导致无法检测出异常。

异常也有多种分类。通常情况下直观的异常指的是点异常，其含义是多个数据实体中，如果存在一个实体对于其他实体来说有极大的偏差，那么这个实体就是一种点异常。另外一种异常被称为环境异常，它与点异常的异常概念是一样的，只是表征的是一个数据实体在特定环境中的异常，存在某种限定条件。这种异常类型的数据实体有两部分组成：环境属性与行为属性。环境属性表征了数据实体所处在的环境，例如时间序列数据的时间点，空间数据的地理坐标；行为属性表征了在上述特定环境属性下区分数据实体的属性，类似于地理数据的某地降雨量，行为属性即是固定了环境属性后的点异常中数据的特征。在本课题中，数据属于时序性的定位数据，其中的异常是一种环境异常。环境属性即是时空坐标与地理坐标，行为属性是某地理坐标下在某时间点上的定位终端数量。

1.2.2 异常检测算法分类（需完善）

异常检测是找出数据特征严重不同于预期对象的一个检测过程。传统检测异常的方法分为以下几类：基于分类的异常检测方法，基于最近邻的异常检测方法，基于聚类的异常检测方法，基于统计的异常检测方法。

 基于分类的异常检测方法：该方法分为两个步骤。第一阶段通过已有的标签数据训练分类器。第二阶段使用该分类器对未知数据进行分类。

 基于最近邻的异常检测方法：该方法基于“正常数据间的距离较近，异常数据与最近的数据点也较远”的假设展开，可以从密度的角度去区分正常点和异常点。

 基于聚类的异常检测方法：该方法基于“正常数据通常聚集在一起，同分类下存在大量数据，而异常数据不属于任何一个小组或是某分类下的数据样本极少”的假设展开，但聚类的思想更适合寻找聚类，即正常数据。

 基于统计的方法：对于一个统计模型，如果输入数据会处于统计概率中较低的位置，那么则认为其为异常数据。

1.3 论文主要工作

本文基于以上课题背景以及研究现状，针对腾讯地图所提供的定位数据，在已知某一天为异常天（台风过境）的前提下，研究并实现检出该天为异常天的算法，同时实现了根据现有时序定位数据预测后几日定位数据的生成模型，最后对区域的定位数据进行了讨论并将上述成果以MATLAB应用的形式输出。具体实现内容包含以下几部分：

1. 对腾讯定位数据进行解析及预处理。首先，由于研究的异常为台风过境某区域时的定位数异常，将定位数据的区域统一在该地域的经纬度；其次，对定位数据进一步作图分析，观察在相同位置处定位终端数量一天内的变化、每天同时段的变化，进而确定分析策略；最后，根据数据特点进行了预处理便于分析。
2. 对经过处理后的数据基于曲线分析进行异常检测。采用了诸如小波变换，最大似然法，差分分析法等传统方法以及一些混合改进算法，对这些算法的效果进行对比分析，总结在该定位数据下哪一种算法实现结果最佳。
3. 根据处理后的数据以及曲线异常检测的结论，设计并实现曲线预测模型。采用动态神经网络使用现有数据进行训练，使用后续补充的数据进行验证。
4. 区域+多异常天
5. 应用

1.4 论文章节安排

第一章干啥

1. 异常检测算法基础

2.1 时序曲线离群点检测

2.2 时序曲线预测

2.3

1. 定位数据

3.1 腾讯定位数据的形式

给定的腾讯定位数据是使用MATLAB的Mapping Toolbox生成的GeoTIFF格式图像文件，每张图像文件的分辨率为113\*150。GeoTIFF文件中包含图像所表示的地理位置信息，经过处理后确定该图像横坐标单位像素长度表示0.01经度，纵坐标单位像素长度表示0.01纬度，其它信息及图像热力图如**下图所示：**

将地理位置信息与实际世界地图进行比对，大致确定数据坐标为广东省珠海市沿海一带**（如图）**。同时，本数据集记录了8月14日至9月30日总计48天每一小时的区域定位终端数量。其中，已知的异常事件为8月23日的台风过境，由于台风袭来势必会导致地图上的终端定位数量发生显著改变，本课题通过分析该时段的终端定位数据来研究定位数据的异常。

3.2 数据分析及应用

对于本课题，首先需要对定位数据的规律进行分析，以便确定异常检测的算法思路。

3.2.1数据的小时变化规律

定位数据在时间上以小时的单位进行采样，可以通过观察某地一天24小时的终端数量值得出定位数据的小时变化规律。为便于观察，应选取终端数量较多的区域从而得出普适规律，而在3.1中我们通过比较已经确定该区域的实际地理位置，可以选择图中人口密度相对较高的珠海市进行研究。将珠海市的地理坐标范围确定后，取该区域的定位终端数量平均值并绘制出其在一天内的小时变化曲线图，如**下图所示：**

由图中可以看出：

为确定此规律符合每一个正常的自然天，而不是工作日或休息日的特殊情况，再取该区域的定位终端数量平均值并在一张图内绘制出其一周每一天内的小时变化曲线，**如图所示**

由图中可以分析得到数据的小时变化规律

3.2.2 数据的日变化规律

由于定位数据又在时空坐标上以自然天的单位进行采样，同时由3.2.1中数据的小时规律结论可知每一天的小时变化规律幅度大致相似，选取任意的小时时间节点分析不会影响结论，即可以通过观察某地每一天的同小时时刻终端数量值得出定位数据的日变化规律。为便于观察，同样选择图中人口密度相对较高的珠海市进行研究并选择一天当中定位终端数量较大的13:00时刻进行研究，取该区域每一天13:00的定位终端数量平均值并绘制出其在数据范围的48天内的日变化曲线图，如**下图所示：**

由图中可以分析得到数据的日变化规律

3.2.3 数据的总变化特征

将上述两章所分析的小时变化规律及日变化规律进行汇总，以X轴为数据范围内的自然日，Y轴为自然日内的每一小时，在Z轴绘出XY形成的<日-时>时间节点上的定位终端数量，**如下图所示：**

由图中可以分析得到和前述结论大致相似数据的总变化特征分析小时上的更有意义。

3.3 数据预处理及异常分析策略

3.3.1 数据预处理

经过**3.2**对数据的分析，我们得到了定位数据的基本形式同时分析了其变化规律。在3.2中，我们选取的大多是极具代表性的区域（人口密度较高的珠海市）进行分析，而由于该定位坐标沿海，从任意时间节点上的定位终端矩阵或是绘制的区域热力图中也能观察得出：位于海面上的坐标终端数值存在大量接近零的点**，如图所示**。这些点无论对于分析数据规律或是检测异常都是冗余的，比如海面上某点两时刻的值从1到2有100%的变化，会极大地影响基于变化率的检测方法，需要将这些点进行剔除。

平均数是一个衡量区域内定位终端数量量级的基本方法，但是考虑到海面上可能会在某时刻突然出现了高额终端数这种极端异常情况，不能轻易地忽略。采用平均数阈值去衡量有效点可能会因为天数过多而将这种异常点舍去，因此本课题更适合采用最大值阈值的方法对数据进行预处理。

创建一个与定位数据地图相同大小的0-1矩阵表征定位数据图中像素点是否有效（以下称为有效矩阵），读取定位数据中每一个像素点在**所有时刻**的值，如果这些值中没有一个超过10（1平方公里的区域中没有一个时刻超过10个定位终端），则将有效矩阵相同位置处置为0，否则置1。经过这样处理后定位数据地图中只有约**2000多个点**有效，极大地加快检测速度同时也避免了突变的错误舍去。

3.3.2 台风天异常分析策略

**由，**基于腾讯的时序定位数据，对于本课题所讨论的异常，应被归类为环境异常。环境属性即是时空坐标与地理坐标，行为属性是某地理坐标下在某时间点上的定位终端数量。本课题的检测目标即是输入所有的时空与地理坐标上的定位数据来检测出**某一天时空或是地理异常（哪一天的问题？）。**

**3.2**中我们讨论了数据的时空特征，对于某固定的地理坐标，分析其每天的同一小时时刻上的定位数据可以归纳出该地理坐标上的天时空特征，简化了时空分析量。而在**3.3.1**中我们又对数据进行了预处理，减少了地理分析量。经过上述讨论，对于本课题所研究的台风天异常，对经过数据预处理后的筛选点进行时空维度上的曲线异常检测，使用每天同一小时点上的数据来分析，判断异常日期是哪一天或是全部为正常数据；再从地理上统计地图上所有已筛点的异常日期，如果地图上的大部分点都指向某一天存在异常的，即可认为该天是异常天。

对上述的地图数据上所有已筛选过的点进行曲线异常检测，。如果最后一张图上的大部分点都指向某一天存在异常，即可认为该天是异常天

1. 基于曲线分析的定位数据异常检测

4.1 基于小波的异常检测算法

4.1.1 离散小波变换

小波变换的定义基础，

4.1.2 结果分析

用什么量级的小波，用小波怎么检测 结果 以及分析

4.2 基于最大似然法的异常检测算法

4.2.1 最大似然法与3西格玛

最大似然法 3西格玛准则

4.2.2 结果分析

通过最大似然法求得MIU和SIGMA，再通过3西格玛准则判断，结果的分析

4.3 基于差分的异常检测算法

4.3.1 差分算法

拉普拉斯算子的介绍

4.3.2 结果分析

怎么求，结果如何

4.4 局部异常因子检测算法

4.4.1 局部异常因子

LOF的概念以及算法介绍

4.4.2 结果分析

怎么求，结果如何

4.5 混合异常检测算法

4.5.1 传统方法的优化

小波+差分结合效果为什么好一些

4.5.2 结果分析

4.6 曲线异常检测算法总结

从算法复杂度、算法准确度来衡量，在这种短小数据量下什么算法比较好

1. 定位数据的预测分析

5.1.1 生成模型与判别模型

5.1.2 生成模型的预测

5.2 基于动态神经网络的定位数据预测

1. 基于区域的定位数据异常检测

6.1 基于图像的异常区域检测

6.1.1 差分分析法

6.1.2 小波分析法

6.2 多异常天的检测

1. 总结与展望

7.1 内容总结

7.2 未来展望