**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

2.1.1 最大似然法与3σ准则 2

2.1.2 离散小波变换 2

2.1.3 差分拉普拉斯算子 3

2.1.4 局部异常因子 3

2.2 时序曲线预测 3

3. 定位数据 3

3.1 腾讯定位数据的形式 3

3.2 数据分析及应用 3

3.2.1数据的小时变化规律 3

3.2.2 数据的日变化规律 3

3.2.3 数据的总变化特征 4

3.3 数据预处理及异常分析策略 4

3.3.1 数据预处理 4

3.3.2 台风天异常分析策略 4

4. 基于曲线分析的定位数据异常检测 4

4.1 基于小波的异常检测算法 5

4.1.1 离散小波变换 5

4.1.2 结果分析 5

4.2 基于最大似然法的异常检测算法 5

4.2.1 最大似然法与3西格玛 5

4.2.2 结果分析 5

4.3 基于差分的异常检测算法 5

4.3.1 差分算法 5

4.3.2 结果分析 5

4.4 局部异常因子检测算法 5

4.4.1 局部异常因子 5

4.4.2 结果分析 5

4.5 混合异常检测算法 5

4.5.1 传统方法的优化 5

4.5.2 结果分析 6

4.6 曲线异常检测算法总结 6

5. 定位数据的预测分析 6

5.1.1 生成模型与判别模型 6

5.1.2 生成模型的预测 6

5.2 基于动态神经网络的定位数据预测 6

6. 基于区域的定位数据异常检测 6

6.1 基于图像的异常区域检测 6

6.1.1 差分分析法 6

6.1.2 小波分析法 6

6.2 多异常天的检测 6

7. 总结与展望 6

7.1 内容总结 6

7.2 未来展望 6

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

2.1.1 极大似然估计

https://blog.csdn.net/zengxiantao1994/article/details/72787849

极大似然估计使用来估计一个概率模型的参数的一种方法，即通过若干次试验，观察其结果，利用试验结果反推最有可能（最大概率）导致这样结果的参数值。

由于样本集中的样本都是独立同分布，可以只考虑一类样本集D，来估计参数向量θ。记已知的样本集为： 似然函数（linkehood function）：联合概率密度函数称为相对于的θ的似然函数。如果是参数空间中能使似然函数最大的θ值，则应该是“最可能”的参数值，那么就是θ的极大似然估计量。它是样本集的函数，记作：

ML估计：求使得出现该组样本的概率最大的θ值。

2.1.2 离散序列小波变换

https://blog.csdn.net/jbb0523/article/details/42554693

小波变换（wavelet transform，WT）是一种新的变换分析方法，它继承和发展了短时傅立叶变换局部化的思想，同时又克服了窗口大小不随频率变化等缺点，能够提供一个随频率改变的“时间-频率”窗口，是进行信号时频分析和处理的理想工具。

离散序列的小波变换就是基于著明的Mallat算法，实际上Mallat算法也是针对连续信号的，但在每一层的分解过程中（这里说的每一层实际上就是前面尺度的概念），各层分解系数之间有着某种关系，什么关系呢？那就看看Mallat算法框图吧，如下图所示：

初始系数x（暂且这么称呼）与其第一层分解后的高频系数D1（细节部分Detail）的关系是x经过高通滤波器g滤波后再下采样，与低频系数A1（近似部分Approximate）的关系是x经过低通滤波器h滤波后再下采样；然后继续对低频系数A1进行第二层分解，低频系数A1与其第二层分解后的高频系数D2（细节部分Detail）的关系是A1经过高通滤波器g滤波后再下采样，与低频系数A1（近似部分Approximate）的关系是A1经过低通滤波器h滤波后再下采样；后面依次类推即可。由于一直在下采样，所以虽然滤波器系数g和h不变，但其滤波带宽一直在减半。初始系数是怎么来的呢？肯定是根据信号得到的，最简单最粗糙的办法就是对信号直接抽样。这是对连续信号进行正交小波分解，有了这些系数，再利用正交小波基，就可以表示出信号了，这类似于连续周期信号的傅里叶级数分解吧。

2.1.3 拉普拉斯算子

2.1.4 局部异常因子

2.2 时序曲线预测

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。**补一张图**

1. 基于曲线分析的定位数据异常检测

经过在第3章中对数据的分析，我们将本身<时-空>的坐标分开分析，先对单个空间上的坐标点进行时空曲线异常检测，再统计空间上的规律，得出台风异常天的检测结果。在本章中，我们对每一种方法进行了设计与验证，查看其是否能够成功检出8月23日的台风天结果，并分析了各个算法的优势以及弊端。

4.1 基于小波的异常检测算法

4.1.1 离散序列小波变换

离散序列的小波变换就是基于著名的Mallat算法，离散序列值x与其第一层分解后的高频系数D1（细节部分Detail）的关系是x经过高通滤波器g滤波后再下采样，与低频系数A1（近似部分Approximate）的关系是x经过低通滤波器h滤波后再下采样；然后继续对低频系数A1进行第二层分解，依此类推，即离散信号x，经过多层分解后最后各分解系数合起来就是变换的结果。

对于本课题的数据，因为数据规模较小，直接采用一层分解即可从原始离散序列分离出有效高频部分D1，即原始信号的突变分量（异常分量）在D1中体现。对D1进行模糊处理并结合A1重建信号与原始信号差分，取最大值的横坐标（日期）即可得到该地理坐标下的异常日期。

4.1.2 结果分析

**图**

使用一层的离散小波变换成功检出了异常天（如图所示），但是有数量相对较大的误检测，并且小波分解计算量较大。

考虑到离散小波分解后是直接采用寻找最大值寻找异常天的横坐标的方法，在某些地理位置上其变化幅度较小，导致曲线本身就很平滑，采用小波变换后提取到的信号高频特征不明显，从而导致取最大值时发生错误造成了误检测。小波变换的确可以将离散的曲线信号中最为明显的部分突出，但后续处理的方法仍需改进。

4.1.3 改进算法

在4.1中我们讨论的小波变换能够将时序数据的高频分量提出，进而放大原始数据的噪声异常点。但由于在那一章中将异常放大后直接采取取最大值的策略导致很多本身没有异常的点也被错误的认为是异常点。所以，我们对该方法进行了优化，将原始时序数据经过小波分解后再进行差分运算，得到异常天的日期。同时，我们又对4.4中提到的原始差分运算进行改进，从相邻点的差分拓展到对相邻三点间的差分，减少异常点对周围点差分的影响。

4.2 基于极大似然估计的异常检测算法

4.2.1 极大似然估计与3σ准则

在第2章中我们讨论的极大似然估计的基本含义，其是用来估计一个概率模型的参数的一种方法，即通过若干次试验，观察其结果，利用试验结果反推最有可能（最大概率）导致这样结果的参数值。经过第3章的讨论，我们已知某地理坐标上的定位终端数在没有异常事件到来的情况下浮动规模应大致符合正态分布。基于此假设后，对于某地的时间序列求解最大似然估计，求解正态分布下似然方程得到唯一解（μ\*, σ\*2）。

在有了似然估计的解之后，我们得到了数据本身的一种拟合分布情况。在正态分布下，数值分布在（μ-3σ,μ+3σ)中的概率为0.9973。可以认为，正态分布中Y（在本课题中为定位终端数量）的取值几乎全部集中在（μ-3σ,μ+3σ)区间内，超出这个范围的可能性仅占不到0.3%。本课题中所涉及的异常数据如果较正常数据偏差大，很有可能会落在此小区间中被检出，即可得到异常天的日期，否则认为不存在异常天。

4.2.2 结果分析

使用极大似然估计拟合数据的分布，并基于3σ准则将异常值筛选出后，得到了如图所示的异常天分布情况。本算法仍然检出了异常天，但误检率仍然较高，且由于要求解每一个坐标点的似然估计方程，计算量相对于小波变换更加大。

考虑到极大似然估计是基于现有的数据对原始分布进行拟合，需要较为庞大的数据量支撑以便充分拟合，才能忽略个别噪声的影。而在本课题所涉及的数据中，数据量较小且这些用来估计的现有数据中也包括了异常的数据，如果异常值偏离很大，估计出的参数会极为不准确，从而导致误检。另外，为了便于计算，本算法认为原始数据基于正态分布，实际情况下需要进行长时间的统计，对区域的定位终端数量有一个充分的采样继而判断该地的数据符合哪一种分布，最后对这种分布的参数进行极大似然估计，才能得到较为准确的模型以及判断异常的条件。

4.3 局部异常因子检测算法

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 基于差分的异常检测算法

4.4.1 差分算法

异常的定义是某个数据严重偏离正常数据的范围之内，那么同样也可以认为在时序数据中，异常数据较其周围的数据有很大的波动。使用差分算法计算每一个点与其左右两个时间节点点上的数据浮动比例，当这个比例超过某种阈值后，即可认为该点是异常的。

4.4.2 结果分析

采用20%的变化率阈值的差分算法得到了如图所示的异常天分布情况。本算法成功检出了异常天误检率已经较低，并且由于本算法完全是线性运算，计算量较低。

差分算法是一种更加针对平稳的时序数据算法，因为它只考虑每一个数据点相邻时序上的点，所以对于时序上突然抖动的数据更加敏感。而且，使用变化率来衡量的差分算法也对小数据的微量浮动很敏感。在本课题中，因为已经事先对数据进行预处理，排除掉了小数据的干扰才使得差分算法准确度较高。在实际情况中，更应该使用变化率加上变化阈值的双条件来判定数据是否异常。

另外，本算法中所涉及的差分采取的是时序上左右两点间的差分值，真实异常点的附近几点也会受到异常点的影响导致可能被归类为异常点，应扩大差分的范围，减少异常点对周围点的影响。并且，对于实时的异常判断来说，不存在时序右侧的差分，可以将其替换为时序前一点与其前一点之间的差分，再求它们之间的平均值。

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 异常的检测与预测

时间序列是根据时间顺序得到跟时间相关的变量或者参数的观测数据[1]。对时间序列的研究主要是挖掘其中有价值的信息，找到其中变化的内在规律[2]。在第4章中我们所讨论的问题都是基于现有的<空-时>数据，从其中检出哪一天存在区域性的异常，检测是一个判别的过程，是从已知数据中检测出异常的模型，这种检测方法可以用于数据清洗。同时，对于本课题所研究的定位数据来说，如需实际应用那么检测方法更加强调实时性，即输入一组新的数据后基于现有的数据能够判别新数据是否存在异常。

时间序列预测是指根据现有的和历史的时间序列的数据，建立能反映时间序列中所包含的动态依存关系的数学模型[3]。我们可以通过建立模型来预测新的数据理应符合什么区间，并与实际数据进行比较，判断其是否异常。

5.2 基于动态神经网络的定位数据预测

在4.4中，我们曾用差分算法计算

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 相邻帧间差分法

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

[1]田铮.时间序列的理论与方法.北京:高等教育出版社，2001.

[2]E. KEOGH.Data mining and machine learning in time series database[C]//Proc of the 5th Industrial Conference on Data Mining(ICDM), Leipzig. 2005.

[3]刘大同.基于Online SVR的在线时间序列预测方法及其应用研究[D].哈尔滨工业大学.2010: 1~3.