北 京 邮 电 大 学

本科毕业设计（论文）中期进展情况检查表

学院	信息与通信工程学院	专业	通信工程	班级	2014211112
学生姓名	林文鼎	学号	2014210328	班内序号	7
指导教师姓名	别红霞	所在单位	北京邮电大学	职称	教授
设计（论文）题目	（中文）基于腾讯定位数据的异常事件检测算法				
	（英文）Anomalous Event Detection Based on Tencent Positioning Data				

目前已完成任务

主要内容：(毕业设计（论文）进展情况，字数一般不少于 1000 字)

异常检测是找出其行为严重不同于预期对象的一个检测过程。这些对象被称为异常点或者离群点。而地图的定位数据中，终端的数量变化可以反映出当地的一些事件变化。在通常情况下，终端的分布应该会服从一个基于时间的规律变化。而在异常事件（如异常气象，交通管制等）出现时，定位数据较以往同时段的数值必然会有异常的波动，我们可以从这些波动中获取到异常事件发生时上述定位数据的异常特征，从而在相似事件发生时可以及时做好有关准备。目前我主要在以下几方面取得了一定进展。

(1) 调研异常检测的研究概况与发展状况，并对主要算法进行复现[1]

异常是指数据特征不符合该特征一般存在的区间的现象。寻找异常挑战来源于两个方面：首先，“异常”这个概念较为模糊，偏离正常数据中心多少可以被界定为异常没有一个定量的数值，甚至完全可以认为在划定的边界线附近的数据是正常的数据；再者，用于划定正常区间的数据中有时也会存在异常，导致划定边界线偏差，同时考虑到正常的数据往往远大于异常数据，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。

异常也有诸多分类。通常情况下我们所说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是异常的，那么其就是点异常。对于本课题，异常应被认为是环境异常，其是一个数据实体在特定环境中的异常；数据实体有两部分组成：环境属性&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。本课题中环境属性即是时间点与地理坐标，行为属性是在某时间点某地理坐标下的定位终端数量。

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，基于最近领的异常检测方法，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。

(2) 读入并分析定位数据，并对数据进行预处理与分析

给定的腾讯定位数据的异常事件为一次台风过境，会导致地图上的终端定位数量发生显著改变。对于该定位数据，我分为以下几个步骤进行处理：

1. 通过文件的分析得到其真实地理坐标并与实际地图进行比对，大致确定为广东省珠海市沿海一带；
2. 数据样本为 TIF 格式文件，横纵坐标分别为经纬度，组成的每一个点其上的值代表了当前时刻的该点存在的终端数量；
3. 通过 MATLAB 对数据进行处理，作出一天内各小时，整体数据内每天的某一小时和整体数据内每天的每一小时的图像，对该时空异常有一个大致的判断。
4. 由于该定位坐标沿海，位于海面上的坐标点存在大量接近零的数值点，对于异常判断是冗余的，采用最大值判断进行剔除。
5. 通过作图，我观察得到每一天内的数值变化大致符合一个曲线，所以通过采样所有日期某一小时的地图数据可以得到异常的日期。
6. 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，判断异常日期是哪一天或是全部为正常数据。如果最后一张图上的大部分点都指向某一天存在异常，即可以认为该天是异常天

(3) 使用异常点检测算法对处理后的数据进行异常检测，并对效果分析

通过上述算法的调研，并根据数据的特征，我采用了以下方法检测异常日期：

1. 差分分析法（Laplace 算子）：通过分析某点前后日期定位数值的变化比例，找到变化波动最大的那一天并且如果那一天的变化确实超过某个阈值，那么该点在那个时间点是异常的；对于数据量小且异常值较为明显的异常样本，该方法十分高效。
2. 小波分析法（离散小波变化）：通过对某点随时间变化的曲线进行一层二进离散小波分解，可以得到信号的高频分量（剧烈的变化）以及低频分量（信号的大致波形），将高频分量减少。使用低频分量以及模糊过的高频分量进行小波重建，将重建后的信号和原始信号相减，并取出峰值（即最大的噪声），认为峰值日期即为异常日期；较差分分析法计算复杂，但对于数据量大时拥有更好的效果。
3. 局部异常因子算法：通过计算某点的局部密度，使用局部离群因子来表示某点的领域点的局部可达密度与某点的局部可达密度之比的平均数来表征某点是否与其邻域内的点为同一簇的可能性。如果这个比值远大于 1，则认为该点代表的日期为异常日期；无论数据量大小，只要异常样本相对明显，则十分有效，缺点是计算复杂。
4. 最大似然法（3σ准则）：假设正常的数据符合高斯分布，采用最大似然法去拟合数据的平均值与方差，找到数据当中超过 $\mu+3\sigma$的点并采用最大值，那么这个点代表的日期即是异常天。由于数据不完全符合高斯分布，且异常值可能存在于 3σ内，所以该方法计算量大且效果不好。

参考文献：

[1] Chandola, Varun, A. Banerjee, and V. Kumar. Anomaly detection: A survey. ACM, 2009.

是否符合任务书要求进度：符合任务书要求进度

<table>
<tr><td rowspan="2">尚需完成的任务</td><td colspan="3">采用更大量的数据分析，并对曲线异常检测算法做进一步更新；
尝试采用图像的方法直接对异常天进行检测；
整合现有功能，使用 Matlab 生成可执行文件，达到输入地图数据→判断异常天的功能。</td></tr>
<tr><td colspan="3">能否按期完成设计（论文）：可以按期完成设计</td></tr>
<tr><td rowspan="2">存在问题和解决办法</td><td>存在问题</td><td colspan="2">(1) 如何判断多异常天数
(2) 地图数据中存在大量的冗余数据如果，采用图像的方法会显的低效。</td></tr>
<tr><td>拟采取的办法</td><td colspan="2">(1) 更适合采用邻域或统计的方法而不是阈值的方法去判断。
(2) 采用聚类的方法将图中划分为几个子图像区域并对每个子图像区域进行分析。</td></tr>
<tr><td colspan="2">指导教师签字</td><td></td><td>日期</td><td>年 月 日</td></tr>
<tr><td colspan="2">检查小组意见</td><td colspan="3">负责人签字： 年 月 日</td></tr>
</table>

注：可根据长度加页。