

北京郵電大學

本科 毕业 设计（论文）



题目：基于腾讯定位数据的异常事件检测算法

姓 名 林文鼎
学 院 信息与通信工程学院
专 业 通信工程
班 级 201421112
学 号 2014210328
班内序号 07
指导教师 别红霞

2018 年 5 月

北京邮电大学

本科毕业设计（论文）任务书

学院	信息与通信工程学院	专业	通信工程	班级	2014211199
学生姓名	猜猜	学号	2014210999	班内序号	99
指导教师姓名	猜猜	所在单位	信息与通信工程学院	职称	教授
设计(论文)题目	(中文) 猜猜看毕设题目是什么				
	(英文) Just Guess What On Earth My Title is				
题目分类	工程实践类 <input type="checkbox"/> 研究设计类 <input checked="" type="checkbox"/> 理论分析类 <input type="checkbox"/>				
题目来源	题目是否来源于科研项目 是 <input type="checkbox"/> 否 <input checked="" type="checkbox"/>				
	科研项目名称:				
	科研项目负责人:				
<p>主要任务及目标:</p> <ul style="list-style-type: none">给出一个 LaTeX 模板。减小大家花费在排版上的时间成本。但是代价是需要有一点基础学习成本，但研究生时可能会用到。					
<p>主要内容:</p> <p>同任务目标。</p>					
<p>主要参考文献:</p> <ul style="list-style-type: none">Zubiaga A, Aker A, Bontcheva K, et al. Detection and Resolution of Rumours in Social Media: A Survey[J]. 2017.Yan Z, Chen W, Chai K Y, et al. Detecting rumors on Online Social Networks using multi-layer autoencoder[C]// IEEE Technology & Engineering Management Conference. IEEE, 2017:437-441.					
<p>进度安排:</p> <ul style="list-style-type: none">2018.1.1 ~ 2018.2.10 完成领域内容调研，模板对应部分撰写。2018.2.28~2018.4.15 完成相关模板研究，设计模板。2018.4.16~2018.4.30 进行模板设计评估和比较分析。2018.5.1~2018.5.15 模板整体撰写。					
指导教师签字		日期	年 月 日		

编号: _____

北京邮电大学本科生毕业设计（论文）成绩评定表

学生姓名	猜 猜		所在学院	信息与通信工程学院					
学号	2014210999	专业	通信工程	班级	2014211199				
论文题目	(中文) 猜猜看毕设题目是什么								
	(英文) Just Guess What On Earth My Title Is								
指导教师姓名		指导教师职称		指导教师单位					
中期检查小组评分	(满分 10 分):			中期检查小组组长签字:		检查日期:			
指导教师评分	评价内容	具体要求			分值			评分	
	调研论证	能独立查阅文献和从事相关调研；能正确翻译外文资料；有收集、加工各种信息及获取新知识的能力和自学能力。			5	4	3.5	3	2
	方案设计	能独立提出符合选题的可行性研究方案、实验方案、设计方案，独立进行实验（如安装、调试、操作）和研究方案论证。			5	4	3.5	3	2
	能力水平	能综合运用所学知识和技能去分析与解决毕业设计（论文）过程中遇到的实际问题；能正确处理实验数据；能对课题进行理论分析，得出有价值的结论。			5	4	3.5	3	2
	学习态度	认真、勤奋、努力、诚实、严格遵守纪律，按期饱满完成规定的任务。			5	4	3.5	3	2
	设计（论文）水平	文题相符、综述简练完整，有见解；立论正确，论述充分，结论严谨合理；实验正确，分析处理科学；文字通顺，技术用语准确，设计（论文）有理论价值和应用价值。			5	4	3.5	3	2
	文本规范	装订顺序正确，字体字号等与基本规范相符，符号统一，编号齐全，图表完备、整洁、正确。			5	4	3.5	3	2
指导教师评分合计(满分 30 分): 评语:									
指导教师签字:					日期: 年 月 日				
复议	<input type="checkbox"/> 是 <input type="checkbox"/> 否 复议评分合计: 复议人签字: 复议日期: 复议有权限修改指导教师评分，选择复议后指导教师评分将由复议评分替换								

本科生毕业设计（论文）答辩成绩评定标准														
答辩小组成绩评定	评价内容	具体要求	分值											
	选题	符合专业培养目标，符合社会实际、结合工程实际，难易适度，体现新颖性、综合性。	5	4	3.5	3	2							
	设计（论文）质量水平	全面完成任务书中规定的各项要求，文题相符，工作量饱满，写作规范，达到综合训练的要求，有理论成果和应用价值。	20	16	14	12	8							
	答辩准备	准备充分；有简洁、清晰、美观的演示文稿；准时到场。	5	4	3.5	3	2							
	内容陈述	语言表达简洁、流利、清楚、准确，思路清晰，重点突出，逻辑性强，概念清楚，论点正确；实验方法科学，分析归纳合理；结论严谨；表现出对毕业设计（论文）内容掌握透彻。	20	18	14	12	8							
	回答问题	回答问题准确、有深度、有理论根据、基本概念清晰。	10	8	7	6	4							
	答辩小组评分合计（满分 60 分）													
意见：														
答辩小组组长签字：_____ 年 月 日														
答辩小组成员：														
学院意见	最终成绩：百分制_____； 五分制_____													
	院长签章：_____ 学院盖章：_____ 年 月 日													
备注														

注：1. 毕业设计（论文）成绩由中期检查评分（满分 10 分）、指导教师评分/复议评分（满分 30 分）和答辩小组评分（满分 60 分）相加，得出百分制成绩，再按 100-90 分为“优”、89-80 分为“良”、79-70 分为“中”、69-60 分为“及格”、60 分以下为“不及格”的标准折合成五级分制成绩；

2. 此表原件一式三份，一份存入学生档案，一份装订到毕业论文中，一份交教务处存入档案馆。

北 京 邮 电 大 学

本科毕业设计（论文）诚信声明

本人声明所呈交的毕业设计（论文），题目《社交网络多媒体信息可信度评估》是本人在指导教师的指导下，独立进行研究工作所取得的成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京邮电大学或其他教育机构的学位或证书而使用过的材料。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名：_____ 日期：_____

社猜猜看这个毕设题目是什么

摘要

这是中文摘要的部分。

它可以拥有多段。哈哈哈哈哈哈或

关键词 北京邮电大学 本科生 毕业设计 模板 示例

Have a try to guess what the title is

ABSTRACT

This is ABSTRACT.

You can write more than one paragraph here.

KEY WORDS BUPT undergraduate thesis template example

目 录

第一章 绪论	1
1.1 课题背景	1
1.2 异常检测研究现状	3
1.2.1 基本概念与挑战	3
1.2.2 异常检测算法分类	4
1.3 论文主要工作	5
1.4 论文章节安排	5
第二章 异常检测算法基础	7
2.1 时序曲线离群点检测	7
2.1.1 极大似然估计	7
2.1.2 离散序列小波变换	8
2.1.3 最近邻及邻域算法	9
2.2 时序曲线预测	10
第三章 定位数据	11
3.1 腾讯定位数据的形式	11
3.2 数据分析及应用	11
3.2.1 数据的小时变化规律	12
3.2.2 数据的日变化规律	13
3.2.3 数据的总时空特征	14
3.3 数据预处理及异常分析策略	15
3.3.1 数据预处理	15
3.3.2 曲线异常分析策略	16
第四章 基于曲线分析的定位数据异常检测	17
4.1 基于差分的异常检测算法	17
4.1.1 差分算法	17
4.1.2 结果分析	17
4.2 基于小波的异常检测算法	18
4.2.1 离散序列小波变换	18
4.2.2 结果分析	18
4.2.3 结果处理优化	19

4.3 基于极大似然估计的异常检测算法	19
4.3.1 极大似然估计与 3σ 准则	19
4.3.2 结果分析	20
4.4 局部异常因子检测算法	20
4.4.1 局部异常密度	20
4.4.2 结果分析	21
4.5 曲线异常检测算法总结	22
第五章 定位数据的预测分析	23
5.1 异常的检测与预测	23
5.2 基于动态神经网络的定位数据预测	24
5.2.1 动态神经网络	24
5.2.2 定位数据预测	26
5.2.3 分析结果	27
第六章 基于区域的定位数据异常检测	28
6.1 基于图像的异常区域检测	28
6.1.1 相邻帧间差分法	28
6.1.2 高浮动区域检测	28
第七章 总结与展望	32
7.1 内容总结	32
7.2 未来展望	33
7.3 特殊文本类型	33
7.3.1 脚注	33
7.3.2 定义、定理与引理等	34
7.3.3 中英文文献、学位论文引用	34
7.4 图表及其引用	35
7.5 公式与算法表示	36
7.5.1 例子：基于主成分分析	36
7.5.1.1 主成分分析算法	36
7.5.1.2 主成分分析可信度评估方法	37
7.6 代码表示	38
7.7 列表样式	38
参考文献	39
致 谢	39
附 录	40
附录 1 缩略语表	40

第一章 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，最后对论文的主要研究工作进行了总结并阐述了本文的行文章节安排。

1.1 课题背景

随着 GPS 定位，传感器网络和高速无线通信等技术的日益发展，越来越多的终端定位数据被收集和保存在应用服务器，它们是各地人口密度的一个衡量依据。除了定位数据本身所体现的人口密度空间特征，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人口密度时序特征。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，例如图??所示从北京市每日的滴滴打车的定位数据可以明显地总结出以下特征：滴滴车辆在早高峰时将大量住在郊区的人群运送至各大工作区（例如中关村和国贸区域），而在晚高峰时它们又将人群从工作区运送回家。这种区域性的时序定位数据反应出了北京市的日人口密度变化特征。

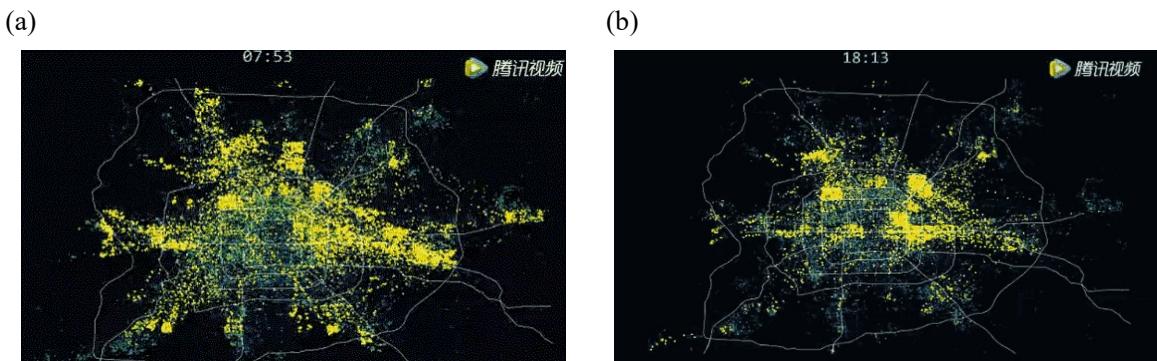


图 1-1 北京市滴滴打车定位数据图

而在这种区域性的时间序列中，也会在某些时间点上出现这样的观测点，它们较以往同时段的数据，例如某个星期天与前几周的星期天来说，会有一个明显的波动，这就是时序数据点中的异常值，如图附-1 所示。分析这些异常点也是一个很重要的课题，通常对于区域性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，可以使有关部门察觉到异常状况的发生并及时做好应急响应。

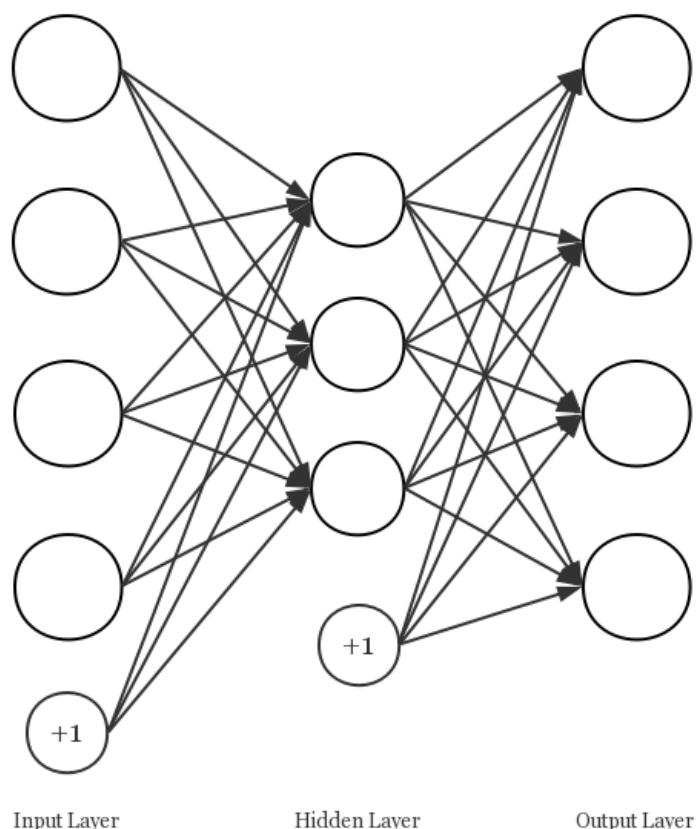


图 1-2 时序数据中的异常值

本课题基于腾讯定位数据，对已知的异常-某一天某区域的台风袭来时的定位数据进行研究，研究能够检测出该天整块区域的异常的算法，并对检测过程中的一些问题进行讨论。

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，如图 1-3 所示。寻找异常是一个非常困难的课题，其难点主要来源于以下两个角度：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使网络结构偏向于正常数据的分布，即过拟合导致无法检测出异常。

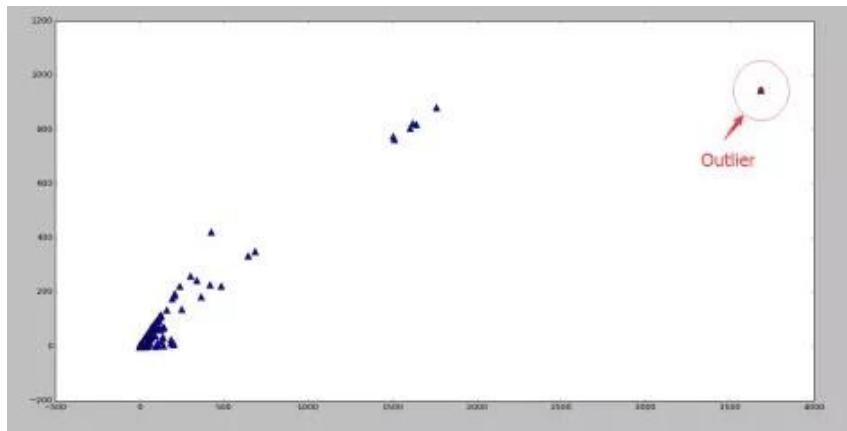


图 1-3 时序数据中的异常值

为了便于分析，异常也有多种分类：通常情况下直观理解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，那么这个实体数据所对应的特征就表明一种点异常。而另外一种异常被称为环境异常，它与点异常中的异常概念是一样的，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，行为属性即固定了环境属性后的数据特征，例如已知某地理坐标上的定位数据量。在本课题中，数据属于时序性的定位数据，而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，行为属性是在某时间点上某地理坐标下的定位终端数量。

1.2.2 异常检测算法分类

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。

1. **基于分类的异常检测方法：**分类是一种从一组已做好标注的数据实例（训练）中学习模型（分类器），然后使用学习模型（测试）将测试实例分类到其中一个类中的方法。基于分类的异常检测算法以类似的两阶段方式生成一个分类模型从而判断数据的特征是否异常：在训练阶段通过使用已进行标记的样本来训练分类器的模型和参数；测试阶段使用分类器将测试实例分类为正常或异常。基于分类的异常检测算法基于以下假设下实现：在给定的数据特征空间中学习可以区分正常类和异常类的分类器是可行的。
2. **基于最近邻的异常检测方法：**近邻分析的概念已用于多种异常检测算法，这些算法都基于以下关键假设：正常的数据实例发生在密集的邻域中，而异常发生在离它们最近的邻居很远的地方。最近邻的异常检测算法需要两个数据实例之间所定义的距离或相似度衡量，而这些距离又针对不同类别的属性有不同的衡量标准。对于连续的属性，欧几里得距离的效果优秀，可以表征出两个数据间的联系性，但在不同情况下也可以使用其他方法计算。对于分类属性，通常使用简单的匹配系数，但也可以使用更复杂的距离度量。对于多变量数据实例，通常为每个属性计算距离或相似度，然后进行合并。
3. **基于聚类的异常检测方法：**聚类用于将类似的数据实例分组到集群中。尽管聚类是无监督式学习，但聚类在半监督式学习中的应用也被最近探讨。尽管聚类和异常检测看起来彼此根本不同，但是已经有几种基于聚类方法被应用于异常检测：
 - 第一类基于聚类的算法依赖于以下假设：普通数据实例属于数据中的一个集群，而异常不属于任何集群。基于此假设的技术将已知的基于聚类的算法应用于数据集，并将任何不属于任何聚类的数据实例声明为异常。
 - 第二类基于聚类的技术依赖于以下假设：正常的数据实例靠近它们最接近的集群质心，而异常距离它们最近的集群质心很远。基于这种假设的技术由两个步骤组成。在第一步中，数据使用聚类算法进行聚类。在第二步中，对于每个数据实例，计算它到最近的集群质心的距离作为其异常分数。
 - 如果数据中的异常自身形成聚类，这些技术将无法检测到这种异常。为了解决这个问题，已经提出了第三类基于聚类的算法，它依赖于以下假设：普通数据实例属于大型且密集的集群，而异常集群属于小型集群或稀疏集群。基于此假设的技术将属于大小和/或密度低于阈值的集群的实例声明为异常。请注意，如果数据中的异常自身形成集群，则这些技术将无法检测到这种异常。

4. **基于统计的方法:** 统计异常检测技术基于以下关键假设: 正态数据实例出现在随机模型的高概率区域, 而异常发生在随机模型的低概率区域。统计技术将一个统计模型(通常用于正常行为)与给定数据相匹配, 然后应用统计推断测试来确定一个看不见的实例是否与该模型相符。基于应用的测试统计信息从学习模型中生成概率较低的实例被声明为异常。

1.3 论文主要工作

本文基于以上课题背景以及研究现状, 基于腾讯地图所提供的时序定位终端地图数据, 在已知某一天整块定位数据区域为异常天(台风过境)的前提下, 研究并实现检测出该天整块区域为异常天的算法, 实现了根据现有的前几日确定时刻时序定位数据预测当前相同时刻定位数据的生成模型来处理局部时刻异常, 最后使用图像处理的方法, 根据异常时间点检测结果识别出造成异常的主要区域。我们对这几块内容进行了讨论并将上述成果以简易 MATLAB 应用的形式输出。具体实现内容包含以下几部分:

1. **数据解析及预处理:** 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时某区域的定位数异常, 将定位数据的区域统一标定在该地域的经纬度; 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, 确定了分析已知异常的策略, 讨论了处理局部时刻异常的方法; 最后, 根据数据特点进行预处理便于分析。
2. **基于曲线的异常检测分析:** 对经过处理后的数据采取曲线分析的形式进行区域性整天异常检测。采用了诸如小波变换, 极大似然估计法, 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, 根据数据结果讨论在该定位数据下对于检测整天整块区域的异常各个算法的表现。
3. **基于曲线的时序数据预测:** 根据数据的分析, 设计并实现曲线预测局部时刻定位数据值模型。采用动态神经网络使用现有数据中的一部分进行训练, 并使用后续补充的数据进行验证, 采用预测模型的方法可以有效避免只分析整天异常而无法分析小时的弊端。
4. **基于图像的异常区域检测:** 在定位数据异常能够被成功检出的基础上, 通过使用图像分析中的帧间差分法以及不同大小的滑动窗计算区域定位数据变化程度来确定造成整体异常的区域。

1.4 论文章节安排

第一章绪论: 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, 最后对论文的主要研究工作进行了总结并阐述了本文的行文章节安排。

第二章异常检测算法基础: 本章主要介绍了异常检测的算法理论基础，涵盖后续几章所用的主要算法的一些理论背景，包括了极大似然估计，小波变换，神经网络等相关概念。

第三章定位数据: 本章主要介绍了本课题所研究的腾讯定位数据的形式，并在 MATLAB 中作图分析其数据特征，包含时序特征及地理特征，最后根据分析结果讨论如何进行检测。

第四章基于曲线分析的定位数据异常检测: 本章主要基于第三章数据分析的结果从曲线的角度对数据进行异常检测，研究了在数据中找出异常的方法。使用了例如小波变换、极大似然估计和邻域的方法进行了分析，并对这些方法的效果进行了讨论。

第五章定位数据的预测分析: 本章主要对异常检测的另一种思路进行了探讨，即实时判断数据是否异常的检测方法。采用了动态神经网络训练时序数据预测模型，并根据预测与实际值比对的结果对异常检测的效果进行了分析。

第六章基于区域的定位数据异常检测: 本章基于前述两章的异常检测结果，检测造成异常的重点区域。采用了图像分析的策略，例如帧间差分法，将异常的核心区域变化特征凸显并使用滑动窗标记。

第七章总结与展望: 本章主要对基于腾讯定位数据的异常检测算法进行分析总结，针对实验结果进行分析，获得本文方法存在的缺点并且提出解决存在问题的有效方法，提高异常检测的准确率和平台使用的有效性。

第二章 异常检测算法基础

本章主要介绍了本课题所研究的异常检测的算法基础，因为在后续几章中要根据本课题所研究的数据对象对算法进行改写，所以在本章中主要介绍理论基础，后续章节中再具体阐述实际应用策略。

2.1 时序曲线离群点检测

2.1.1 极大似然估计

极大似然估计是一种反推样本模型参数的统计方法。其通过对已知的样本信息建模，在数据符合某种特定分布下对分布中的参数进行似然估计。当给定了足够多的观测数据情况下，利用这些大量的试验结果去计算参数值为多少才最有可能导致这样的实验样本结果，即已经知道骰子的大量独立投掷结果去计算扔到骰子各个面的概率参数值。例如已知数据集的采样是独立且为正态分布时，我们可以使用极大似然估计法求出参数 μ 和 σ 的值。

设已知样本集为：

$$\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$$

似然函数(Link 函数)：联合概率密度函数 $P(\mathcal{D}|\theta)$ 称为相对于 $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N, \dots\}$ 的 θ 的似然函数。

$$l(\theta) = P(\mathcal{D}|\theta) = p(x_1, x_2, \dots, x_N | \theta) = \prod_{i=1}^N P(x_i | \theta)$$

$\hat{\theta}$ 值是未知的参数值，我们要对其进行估计，如果它的值可以在范围内使前述似然函数 $l(\theta)$ 的值最大，则 $\hat{\theta}$ 是造成这样的试验样本的最大可能值，即是 θ 参数的最大似然估计量。它是样本集的函数，记作：

$$\hat{\theta} = p(x_1, x_2, \dots, x_N) = d(\mathcal{D})$$

而 $\hat{\theta}(x_1, x_2, \dots, x_N)$ 被称作极大似然函数的估计值

已知试验样本服从独立的正态分布 $N(\mu, \sigma^2)$ ，则试验样本的似然函数如下式所示：

$$L(\mu, \sigma^2) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i-\mu)^2}$$

它的对数：

$$\ln L(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

求导，得方程组：

$$\begin{cases} \frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0 \\ \frac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0 \end{cases}$$

联合解得：

$$\begin{cases} \mu^* = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \\ \sigma^{*2} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \end{cases}$$

在上述条件下，似然方程存在唯一的解 (μ^*, σ^{*2}) ：而且它一定是最大值点，这是因为当 $|\mu| \rightarrow \infty$ 或 $\sigma^2 \rightarrow \infty$ 或 0 时，非负函数 $\ln L(\mu, \sigma^2) \rightarrow 0$ 。于是 μ 和 σ^2 的极大似然估计为 (μ^*, σ^{*2}) 。

2.1.2 离散序列小波变换

时序信号或序列存在大量信息，但是仅仅从时序的角度去分析信号会损失大量有效信息。Fourier 变换提供了一种变换域分析的方法，它利用大量的三角基去构造信号，从频域的角度对信号分析，得出更多信号的信息。

$$F(\omega) = F[f(t)] = \int_{-\infty}^{\infty} f(\omega) e^{-i\omega t} d\omega$$

美中不足的是，Fourier 变换只能反映信号的频域特征，即将整个时序信号的频域分量提出，而不能反映各个时间点上的频域特征。短时 Fourier 变换（STFT）对此进行了改进，使用了定长的窗口对信号的时间进行了限制从而可以将每个短时进行 Fourier 频谱分析，最终得到时间轴上的频域特征。

$$X(t, \omega) = \int_{-\infty}^{\infty} \omega(t - \tau) x(\tau) e^{-j\omega\tau} d\tau$$

小波变换是对短时 Fourier 变换的进一步改进，与后者不同的是，小波并非使用固定的窗口去限定时间，而是使用随时间变化的窗口-小波基去构造短时的 Fourier 分析。它可以有效地改善 STFT 中对于随时间信号幅度变化很不均匀的信号无法妥善处理的缺点。

$$W_f(a, b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} f(t) \overline{\Psi \frac{t-b}{a}} dt$$

小波变换和 Fourier 变换类似，都有应用于连续信号及离散信号的分析方法。本课

题所研究的时序数据应当使用离散序列小波变换，它基于 Mallat 算法将时序信号按照低频信号与高频信号进行逐层分解，每一层的低频信号被继续分解为低高频信号，以此类推，具体的分解关系如下图所示：时序离散序列的初始向量系数 x 在经过一层分解后得到高频部分 D_1 （系数的高频部分）与低频部分 A_1 （系数的低频部分），这两个分量是原始信号分别经过高通与低通滤波器后下采样得到的，代表着原始信号的细节以及近似部分。而后，因为 D_1 反应的是原始信号的细节部分，继续分解没有意义，所以我们分解近似部分 A_1 ，得到二层分解的 D_2 和 A_2 ，它们是 A_1 信号分别经过高通与低通滤波器后下采样得到的。我们仍可以继续对 A_2 进行分解，以此类推。由于在离散小波分解时滤波器系数 G 和 H 保持不变，所以带宽减半，但因为经过了下采样，所以每一部分仍然可以近似估计原始信号。将上述信号分解后的系数，经过正交小波基，可以还原原始信号的近似值。

2.1.3 最近邻及邻域算法

如图所示，对于 C_1 集合里的点，它们虽然互相之间相隔较远（相对于 C_2 来说），但它们的互相之间的间距以及分散情况是均匀的，可以认为是统一集合而不是异常的离散点。 C_2 集合显然是统一集合，而 o_2 虽然相对于 C_1 集合内的点的距离是不会被认为是孤立点，但是其距离最近的 C_2 集合相对于 C_2 集合是较远的，即 O_2 是异常点。LOF 算法从最近邻的思想展开，提供了一种检测异常的手段。首先，我们介绍 LOF 算法的前序概念：1) $d(p, o)$: P 点和 O 点之间的实际距离 2) k -distance: 第 k 距离对于点 p 的第 k 距离 $d_k(p)$ 定义如下： $d_k(p) = \min_{o \in N_k(p)} d(p, o)$ 和 satisfy: a) 至少 k 点 o , $c_x c_x p_o$, 它不包括集合中的 p 。 $c_x p$, $d(p, o_i) \leq d(p, o_j)$ $d(p, o_i) \leq d(p, o_j)$; b) 在集合中，最多有 1 个点 O , $c_x c_x p_o$, 不包括 p ，包括 k , $1k$ 。 $c_x p$, $d(p, o_i) < d(p, o_j)$ $d(p, o_i) < d(p, o_j)$; 从 k 到 p 的距离是点 p k 的距离，不包括 p ，如图 3 所示。3) k -distance neighborhood of p : 第 k 距离邻域点 P 的点 K 是邻域中的 $N_k(p)$ ，即 p 的 k 距离内的所有点，包括 k 距离。因此，在 K 上， k 距离的数量是 k $n_k(p)$ 上的 $n_k(p)$ 。4) reach-distance: 可达距离点 o 到点 p 的第 k 可达距离定义为： $\text{reach-distance}_k(p, o) = \max_{o' \in N_k(o)} d(o, o')$ $\text{reach-distance}_k(p, o) = \max_{o' \in N_k(o)} d(o, o')$ 也就是说，点 O 到 p 的可达距离，至少 O 的 k 距离，或者 O 和 P 之间的真实距离。这也意味着，从 O 的最近 k 点， O 到它们的可达距离被认为等于和等于 $d_k(o)$ $d_k(o)$ 。如图 4, $o_1 o_1$ 到 p 的第 5 可达距离为 $d(p, o_1)$ $d(p, o_1)$, $o_2 o_2$ 到 p 的第 5 可达距离为 $d(p, o_2)$ $d(p, o_2)$ 。5) local reachability density: 局部可达密度点 p 的局部可达密度表示为：表示点 P 中从点 K 到 P 的平均可达距离的倒数。注意 P 的邻域点 $N_k(p)$ $n_k(p)$ 到 p 的可达距离不是从 p 到 $n_k(p)$ $n_k(p)$ 的可达距离。我们必须澄清这种关系。此外，如果存在一个重复点，那么分母的可达距离的总和可以是 0，这将导致 LRD 变得无限大。这个值的含义可以用这种方式来理解。首先，这代表密度，密度越高，我们越有可能属于同一个簇，密度越低，离群点越有可能。如果 P 和周围邻域点是相同的簇，则可达距离越小，则 $D_k(O)$ $D_k(O)$ 越小，导致可达距离和

密度值越小，如果 P 和邻域越远，则距离越大。 $UE D(P, O) \propto D(P, O)$ ，导致密度较小，并且更可能是离群值。

2.2 时序曲线预测

在现代信息产业中，神经网络是人们根据人脑中神经细胞中的运作原理（虽然实际上复杂的多）所模拟出来的计算系统，它通过大量的近似模拟去解决普通计算机程序无法解决的复杂问题，例如模式识别这类对人来说相对轻松的任务。神经网络通常涉及大量处理器并行运行并按层排列的处理器。第一层接收与人类视觉处理中的视神经相似的原始输入信息，例如人眼接受的光信号在计算机中被表示为图像。每个连续的层次都接收来自其前一层的输出，而不是重新采样，越后层的神经元接受离它越近的前序神经元传来的信号。最后一层产生系统的输出。在这些一层层的神经元节点中，每一个神经元处理节点都在网络中扮演着自己的角色，即是我们无法去理解这些节点实际对信息做了什么处理，但所有的节点高度相连后，整个神经网络的输出便和输入存在某种对应关系（例如输入图像，输出判别结果）。神经网络在模型结构确立后，是根据训练样本来修改自身网络结构参数的，其最基本的学习集中在每一个神经元都根据前序输入进行加权，并不断权衡参数权重，使网络能够更加可能获得正确答案。通常情况下，一个网络需要经过大量已标注数据来进行训练，通过这些标注过的数据告诉网络在某种输入情况下理应输出什么，提供答案可让模型调整其内部权重，以了解如何更好地完成工作。

第三章 定位数据

本章主要介绍了本课题所研究的腾讯定位数据的基本形式，并在 MATLAB 中作图分析其数据特征，包含时序特征及地理特征，最后根据分析结果讨论如何进行异常检测。

3.1 腾讯定位数据的形式

本课题所给定的腾讯定位数据是使用 MATLAB 的 Mapping Toolbox 生成的 GeoTIFF 格式的图像文件，每张图像文件的分辨率为 113*150，文件中的 Reference 信息包含图像所表示的地理位置信息。本类图像数据经过处理后可以确定该图像单位像素上的值表示实际地图上的 0.01 经度与纬度覆盖面积上（约为平方一千米）定位终端数量，其中横坐标表示经度，纵坐标表示纬度。根据上述地理信息可以在 MATLAB 中画出该区域的定位终端热力图，为表现特征，采用归一化后终端定位数据作出该区域定位终端密度热力图如图 3-1 所示：

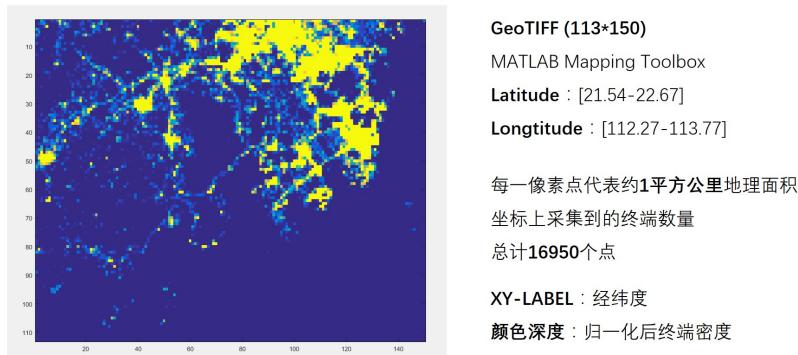


图 3-1 腾讯定位数据基本信息

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 3-2 所示。同时，本数据集记录了 8 月 14 日至 9 月 30 日总计 48 天每天的每一小时区域定位终端数量。其中，该数据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检测方式。

3.2 数据分析及应用

对于本课题，数据的维度涵盖时间与空间，直观上要分析单日的区域性台风影响带来的异常是困难的，我们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。

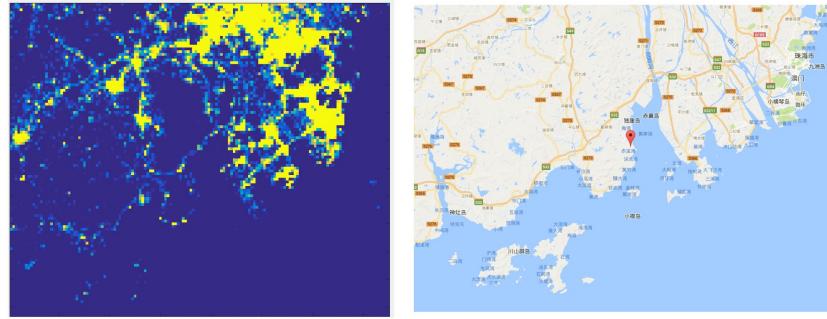


图 3-2 腾讯定位数据基于的实际地理坐标

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从午夜 0 点至次日午夜 0 点的小时变化曲线图，如图 3-3 所示：

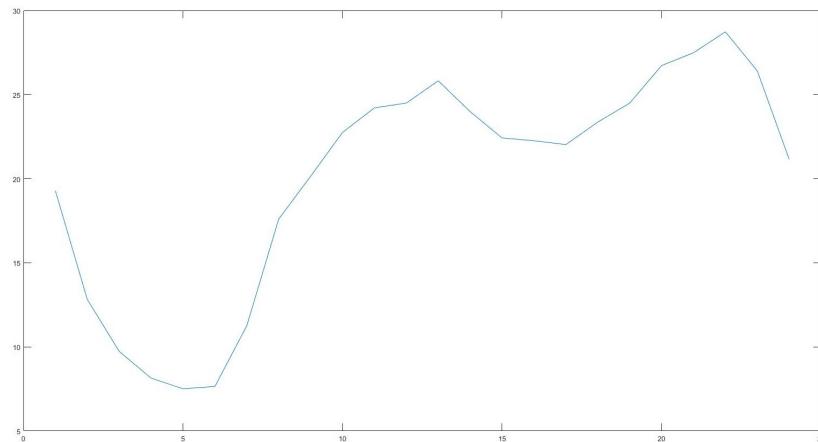


图 3-3 定位数据一天中随小时变化的曲线图

由图 3-3 中可以看出：午夜至清晨时间段为定位终端数量最少的时间段，大多数终端处于关闭状态，这是由于在这期间用户正处于睡眠状态造成的；而随着清晨至午间及晚间的推移，用户逐渐起床、工作、娱乐，定位终端数量也能观察到数量上的上升，而后从午夜开始再次下滑。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或是误采了某个节日的数据，再取该区域的定位终端数量平均值并在一张图内用不同颜色的曲线绘制出其一周每一天 24 小时内的小时变化曲线，如图 3-3 所示：

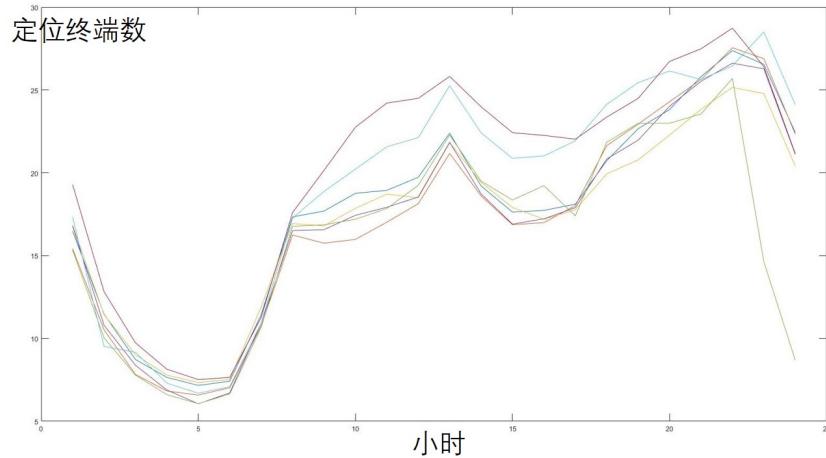


图 3-4 定位数据一周中各天随小时变化的曲线图

接下来我们再将一个正常天的 24 小时内终端数量变化的曲线与台风天的曲线进行比较, 如图 3-5 所示:

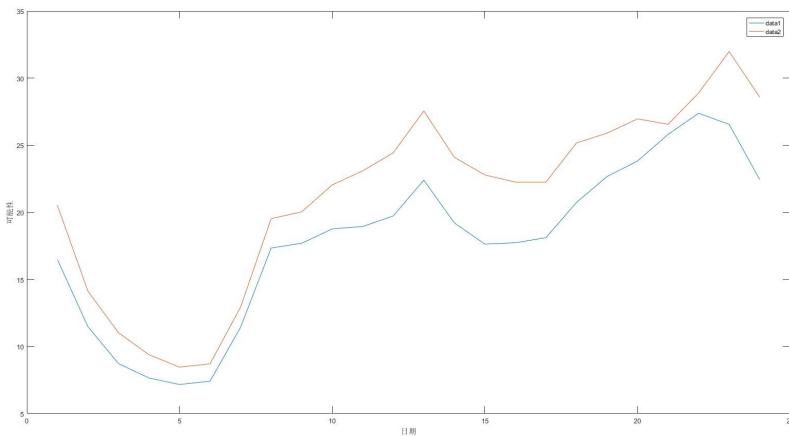


图 3-5 定位数据在正常日与台风日随小时变化的曲线图

由图 3-5 可以分析得到数据的小时变化规律: 定位终端数据在确定的地点内一周中每天的变化趋势大致相同, 而对于我们所已知的台风天异常, 所造成的影响将不仅限于某小时带来的影响, 而是会对整天各个小时的数据产生大的波动。

3.2.2 数据的日变化规律

在上一小节中, 我们绘制出了台风天与正常天一天内 24 小时定位终端数量的变化并且观察得到台风天在一天内的值较正常值有较大的偏差。由于定位数据又在时空坐标上以自然天的单位进行采样, 我们可以用同样的思路挖掘数据的日变化规律。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，如图 3-6 所示：

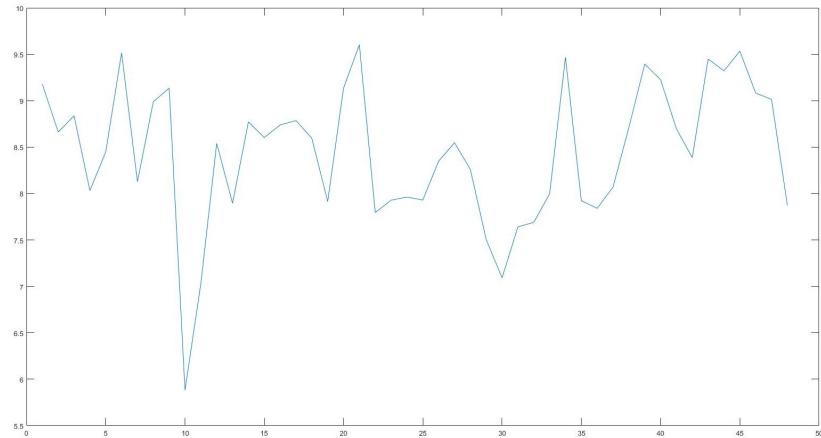


图 3-6 数据范围所有日期的 13:00 定位数据量变化曲线图

由图 3-6 中可以分析得到：在该地的每天 13:00 时刻，台风天相较于正常天的定位终端数有明显的偏差，是上一小节中 24 小时的定位终端变化曲线中的 13 点时刻偏差值的天数扩展。那么，如果能够通过选取任意的小时时间节点来代表整一天的定位终端数量值，台风天异常检测问题将会转化为每一天中的某确定小时的曲线异常检测问题。我们对此做进一步验证及讨论。

3.2.3 数据的总时空特征

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，如图 3-7 所示：

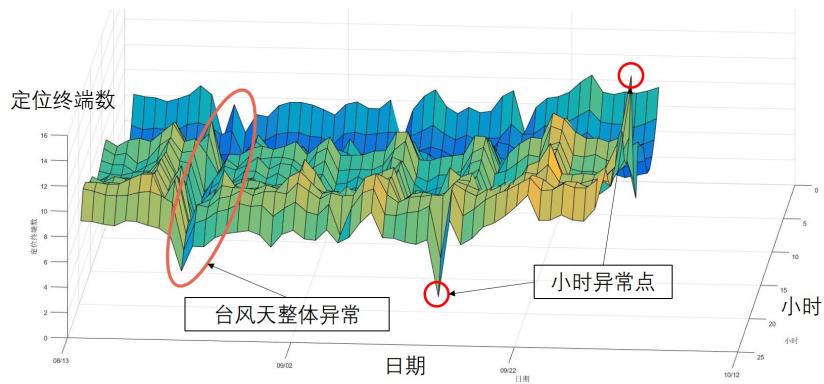


图 3-7 数据范围所有日期中定位数据量随小时变化曲线图

由该三维图的多角度观察可以分析得到：一天 24 小时内的定位数据变化确实大致相似，并且定位数据在 8 月 23 日中整体出现了一个明显的沟壑。但是，与前两小节的

小部分时间节点分析不同的是，在这张图上某些小时时间点上出现了数据的突变暴露出来。如果仅依据每天中单个小时的数据对整天进行分析，过大的单小时点可能会对算法的判断产生误差；另外，由于图中仅显示了某固定点的定位数据图，需要对每个地点的时空特征进行统计，得到总体的判断依据。

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。但实际由于该定位坐标沿海，或是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 3-8 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。

02	3	4	5	20	11	12	3	8	9
7	2	0	6	14	14	4	6	38	6
323	503	701	260	116	16	11	13	52	57
151	341	772	324	231	326	169	325	128	95
303	876	222	807	870	346	438	153	125	155
41	38	67	1343	1290	237	123	0	64	0
220	243	157	979	390	145	71	0	0	0
296	177	143	129	64	122	112	120	4	0
200	54	180	119	122	29	3	323	506	112
90	73	50	205	16	0	4	24	22	28
24	119	282	126	3	0	2	3	2	167
0	2	0	0	0	0	0	4	22	49
0	0	0	0	0	0	0	1	3	17
0	0	0	0	0	0	0	2	5	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	5	67
0	0	0	0	0	0	0	0	8	5
0	0	0	2	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
10	0	0	5	0	0	0	0	0	12
0	8	299	37	5	0	0	0	71	71
2	120	54	36	0	0	0	7	126	184

图 3-8 定位数据量中的高密度与低密度区域对比

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，不能轻易地将其忽略。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的同时也避免了突变的错误舍去。

3.3.2 曲线异常分析策略

在第一章中我们讨论过异常的分类，而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。台风天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的时空异常（出现台风的迹象）。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，其可能会在某个小时时间点上出现大的偏差，即使每天按照小时的变化定位数据的曲线大致相似，这些突然的抖动不能被忽略，所以应当使用一天中平均的定位数据来衡量。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，同时排除了一些会对算法造成影响的无效数据点。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，使用每天的平均数据来分析，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选点的异常日期，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。而对于其它类型的异常，例如某小时突然出现的剧烈抖动，将在后续章节进行讨论。

第四章 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整个区域8月23日的台风异常结果，并分析了各个算法的优势以及弊端。

4.1 基于差分的异常检测算法

4.1.1 差分算法

异常是指某个数据严重偏离正常数据的范围之内，在时序数据中，如果某时刻的定位数据较其周围时刻的数据有很大的波动，该时刻的定位数据也是异常的。使用差分算法计算每一个点与其左右两个时间节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。

4.1.2 结果分析

采用了20%的变化率阈值的差分算法得到了如图4-1所示的异常天分布情况。该图横坐标为日期，纵坐标为经过筛选后的定位点中有多少在某日期中被算法认为是异常的。由此可见本算法成功检出了异常天，并且误检率已经较低，并且由于本算法完全是线性运算，计算量较低。

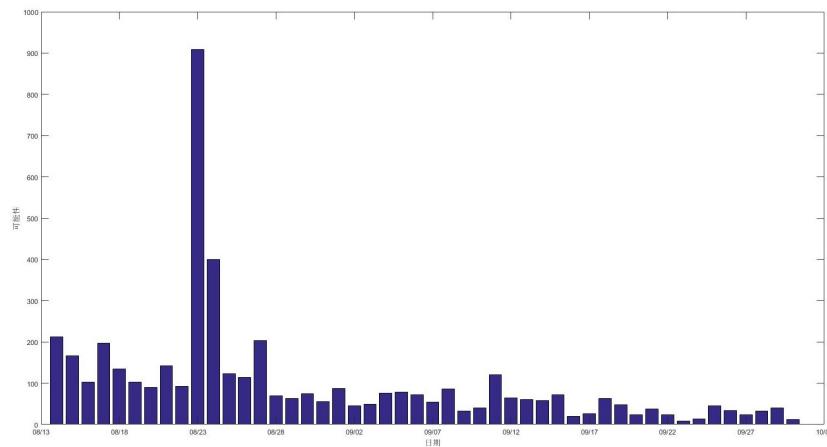


图 4-1 差分法定位数据异常天检出分布图

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。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，实际情况中可能会出现持续的异常，此时差分方法不再适用，它只会关注陡然的变化而忽视了变化之后连续的异常，甚至还会认为持续异常之后回到正常状态的变化是异常的。

4.2 基于小波的异常检测算法

4.2.1 离散序列小波变换

离散序列的小波变换基于著名的 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。

4.2.2 结果分析

使用一层的离散小波变换成功检出了异常天，如图 4-2 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。

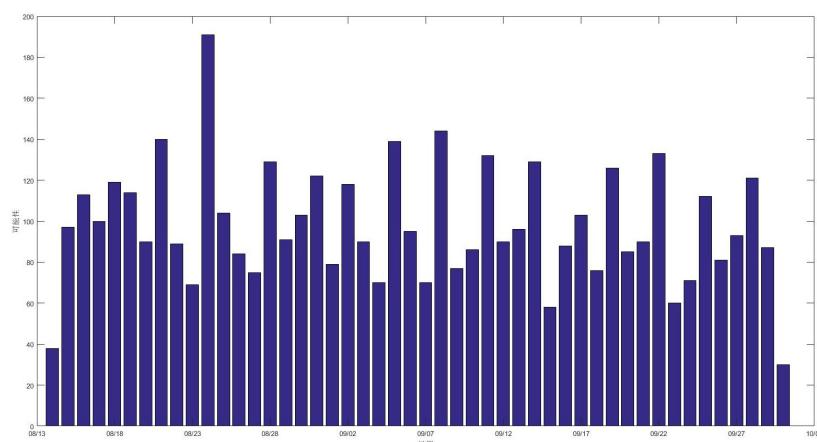


图 4-2 小波变换算法定位数据异常天检出分布图

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细微变化的部分突出，但由于台风异常的变化本身就很剧烈，导致数据的一层小波高频分量不明显。

4.2.3 结果处理优化

在 4.1.2 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。但由于在这一方法中将异常放大后直接使用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，得到异常天出现的日期。同时，我们又对 4.1 中提到的差分运算运用到小波运算之中，如图 4-3 所示。同样由于台风异常的变化本身就很剧烈，使用小波运算提取时序数据中的高频分量后再进行差分的效果不如直接对原始数据进行差分的效果好。在普通的差分算法那中，平滑的数据中出现了陡变分量则马上被检出。

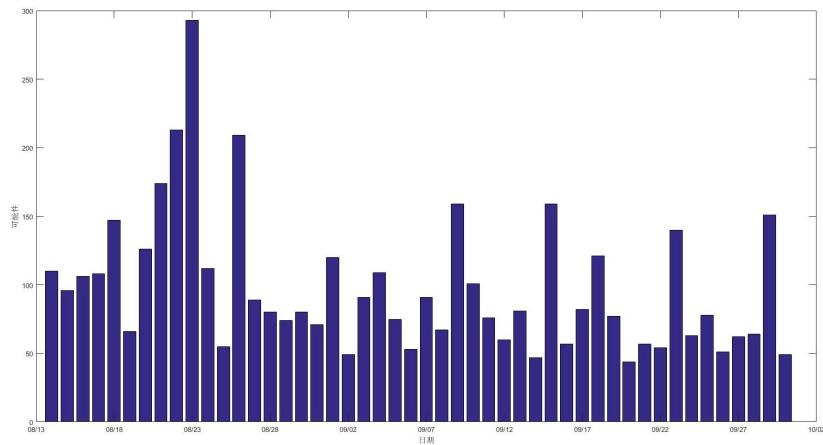


图 4-3 小波变换算法处理结果优化后的定位数据异常天检出分布图

4.3 基于极大似然估计的异常检测算法

4.3.1 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^*) 。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 $(\mu - 3\sigma, \mu + 3\sigma)$ 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 $(\mu-3\sigma, \mu+3\sigma)$ 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。

4.3.2 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。

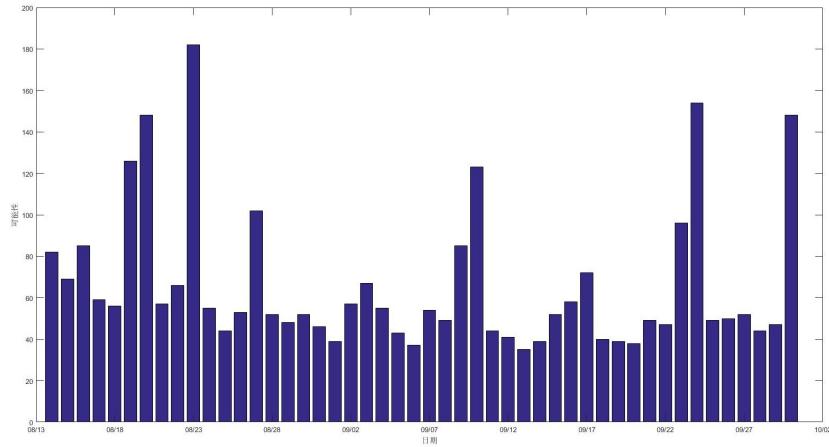


图 4-4 基于最大似然法的定位数据异常天检出分布图

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响。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进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。

4.4 局部异常因子检测算法

4.4.1 局部异常密度

点 p 的局部可达密度表示为:表示点 P 中从点 K 到 P 的平均可达距离的倒数。注意 P 的邻域点 $N_k(p)$ 到 p 的可达距离不是从 p 到 $N_k(p)$ 的可达距离。我们必须澄清这种关系。此外,如果存在一个重复点,那么分母的可达距离的总和可以是0,这将导致LRD变得无限大。这个值的含义可以用这种方式来理解。首先,这代表密度,密度越高,我们越有可能属于同一个簇,密度越低,离群点越有可能。如果 P 和

周围邻域点是相同的簇，则可达距离越小，则 $DK(O) / DK(O)$ 越小，导致可达距离和密度值越小，如果 P 和邻域越远，则距离越大。 $UE D(P, O) / D(P, O)$ ，导致密度较小，并且更可能是离群值。点 p 的局部离群因子表示为：点 $P(NK)(p) / NK(p)$ 的局部可达密度的均值与点 P 的局部可达密度成正比。如果比值接近 1，则 P 的邻域点密度相似， P 可能是邻域中的一个簇。如果比值小于 1， P 的密度高于邻域点密度， P 是稠密点。如果比值大于 1， P 的密度小于其邻域点密度， P 更可能是异常点。现在介绍了概念定义，现在我们回顾 LoF 的思想，主要是通过比较每个点 P 及其邻域点的密度来确定点是否是一个异常。如果点 P 的密度较低，则越有可能被识别为异常。至于密度，它是由点之间的距离来计算的。点越远，密度越低，距离越近，密度越高，这完全符合我们的理解。此外，由于 LF 通过点的 k 邻域而不是全局计算来计算密度，所以称为“局部”异常因子，因此，对于图 1 C1 和 C2 的两组数据集，LoF 可以被正确地处理，并且正常点不被判断为异常 POINT。TS 由于数据密度不同而分散。

4.4.2 结果分析

采用了局部离群因子来计算异常值的位置，本算法也成功检出了异常天，如图 4-5 误检率较低，但是由于局部异常因子算法需要对每个点局部异常因子进行计算，计算量复杂，效率较低。当 K 值（邻域范围）增加时，计算量进一步加大。

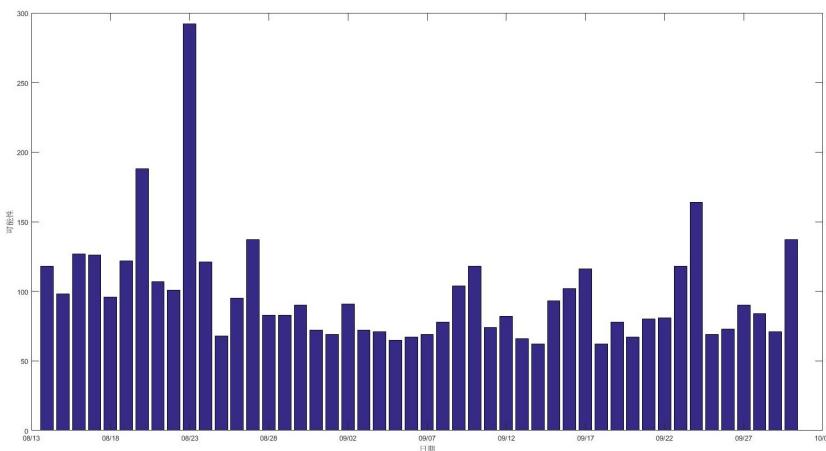


图 4-5 基于局部离群因子的定位数据异常天检出分布图

局部离群因子算法与差分算法等考虑时序数据的连贯性不同，它无论时序数据还是纯粹的行为异常点都可以处理，它仅仅将数据点放置在同一张平面，用数据之间的联系来计算某点是否异常。对于 K 值的选择问题，由于该值的选择标准没有一个普适的定量方法，本算法采取了 $K=2$ 的值（文献）。

	异常点数	运行速度 (秒)
差分分析	909	1.41
小波变换	191	7.38
小波变换-结果改进	293	8.75
局部异常因子算法	292	18.91
最大似然法 (3σ 准则)	182	14.32

图 4-6 曲线异常检测各算法对比

4.5 曲线异常检测算法总结

本章节对台风天的异常检测算法进行了讨论，将大范围的地理坐标下的整天异常问题细化为每一点的整天平均定位数据异常问题。在这个基础之上，我们采用 4 种算法对该问题进行了研究。对于本课题所研究的数据，时序关联性很强，每一天的平均定位数据量值相仿，而台风当天相对于邻近日整体的变化浮动相对较大，是一个极其明显的抖动。差分分析法对于本课题所研究的范围性异常检测问题快速有效，因为其能够捕捉到明显陡变的时序数据，如图 4-7。但是如果对于持续时间较长的异常，差分算法不再有效。由图 4-6 可见除了差分算法的检出量较高之外，其余方法的检出量大致处于同一水平。由于本课题的数据量较小，使用统计的方法来衡量异常的偏差较为困难；局部异常因子算法则存在和差分分析一样的问题，当异常点不再孤立时较难检出，但其对于数据的时序性没有要求，只是检出一个数据是否在一系列数据中是异常的。

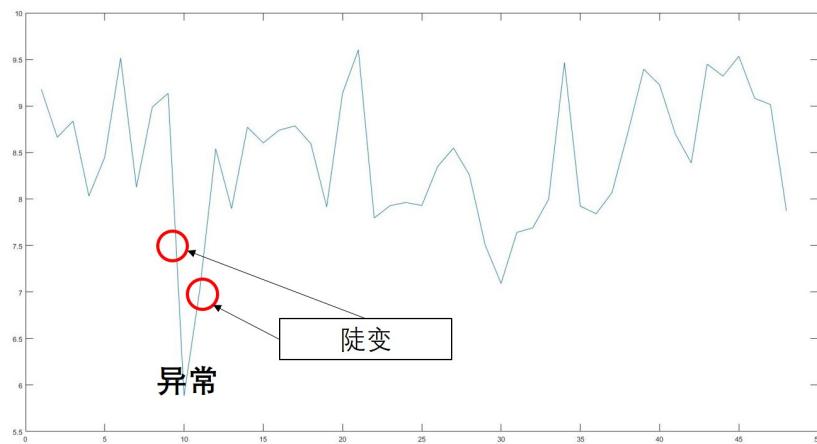


图 4-7 差分算法能够轻松检出异常时刻的数据陡变

我们在第 3 章最后讨论了某小时时刻的数据陡变情况，由于这些陡变的值被更大的台风天变化所忽略或是只出现在少部分地理位置，所以本章的算法没有将这些异常值检出，我们将会在下一章中对其进行进一步讨论。

第五章 定位数据的预测分析

本章主要基于第4章中的整体区域台风天异常检测基础，提出了时序数据的预测估计方法，并根据神经网络训练模型预测当前时间节点的值与实际值进行比对，进一步讨论了异常检测的另一种思路，探索一种方法去完善异常检测算法。

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，这种检测方法更倾向于数据清洗。同时，异常检测在生活中也多用于即时的判断，例如网络流量的异常检测可以对DDoS攻击进行相应，使网络运营商做好应急预案。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。

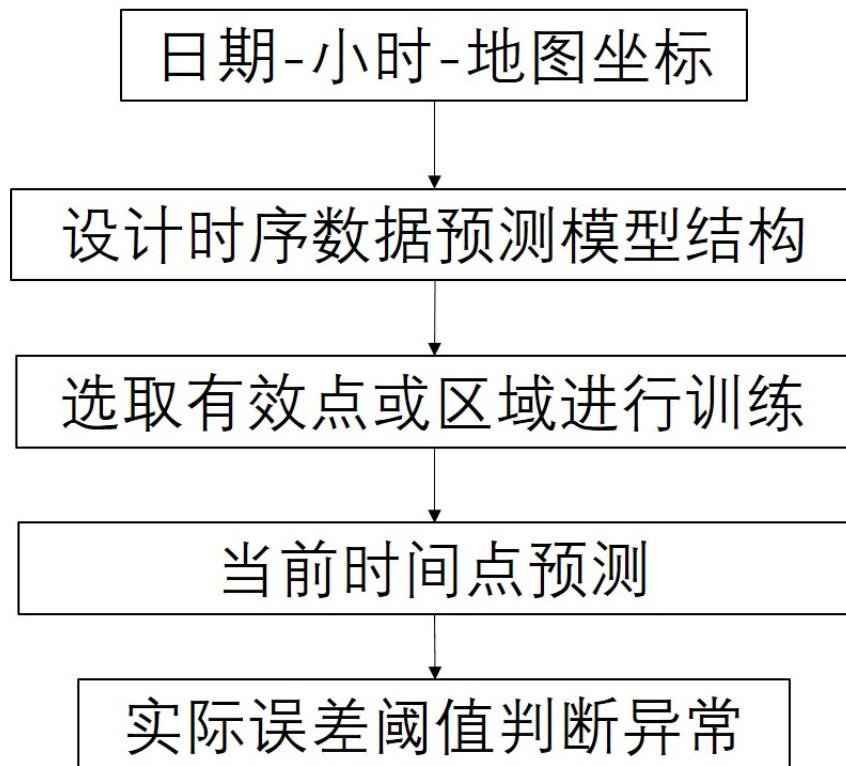


图 5-1 差分算法能够轻松检出异常时刻的数据陡变

在前序章节中我们曾提及本课题所研究的腾讯定位数据存在一些局部地理位置上的小时点突然异常，而第4章的方法更加适合检测整天整块区域的异常，这一章中我们更关注这些异常，通过预测的方法，预测当前时刻的定位数据值，并与实际值比较来检出异常，如图5-1所示。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

5.2.1 动态神经网络

神经网络是一种重要的机器学习技术。它是模仿生物神经网络（动物的中枢神经系统，特别是大脑）的结构和功能的数学模型或计算模型。它被用来估计或近似函数。神经网络是由大量的人工神经元连接来计算的。在大多数情况下，人工神经网络可以在外部信息的基础上改变内部结构，它是一个自适应系统。通过校正每个层的权重（学习）来创建模型的过程被称为自动学习过程（训练算法），通过训练样本的校正。由于网络结构和模型的不同，具体的学习方法不同，并使用反向传播算法（反向传播/反向传播/反向传播，使用差分增量规则来修改权值）来验证该方法。通过这样学习（训练与验证）的过程，它可以对目标函数进行相对完整的模拟，如图5-2所示。

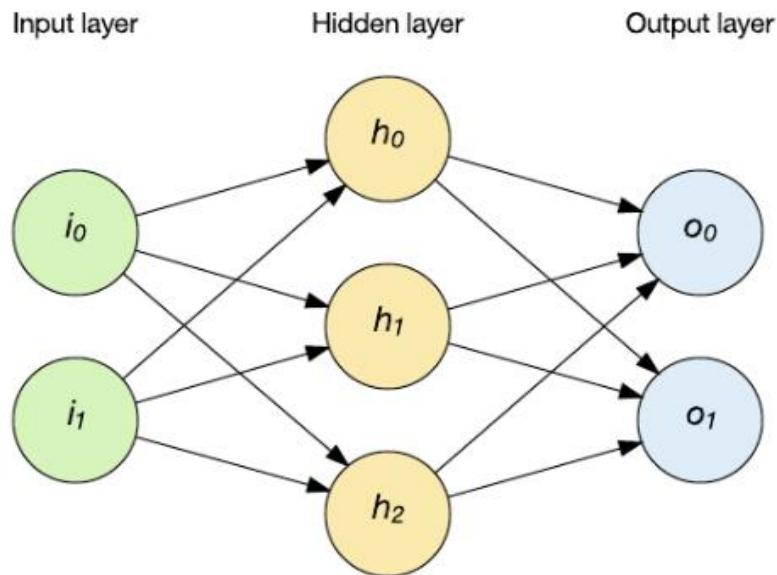


图 5-2 前馈神经网络示意

神经网络可分为静态神经网络和动态神经网络，根据其是否包含延迟或反馈，将具有延迟或反馈环节的神经网络称为动态神经网络。静态神经网络由前向传播的神经元组成，连接仅馈送到后续层，因为没有回馈的因素（即同一列神经元间没有相连的关系），其主要被用于静态判断或预测，例如卷积神经网络被用来识别静态图像。动态神经网络

中一个显著的特点就是神经元间的输出回馈到前序神经元中，神经元之间的联系使得动态神经网络又可以解决静态神经网络无法解决的时序问题，由于 RNN 包含循环，它们可以在处理新输入的同时存储信息。这种存储器使得它们非常适合于处理在输入之前必须考虑的任务，例如时间序列数据。在 4.4 中，我们曾用差分算法检测异常值，差分算法的本质在于通过相邻时间节点的浮动变化率对异常点进行判断，即根据变化趋势来检测是否存在异常。上述动态神经网络也需要使用训练出的模型并根据前几项时间节点上的值，来预测当前时间节点上的值，如图 5-3 所示。

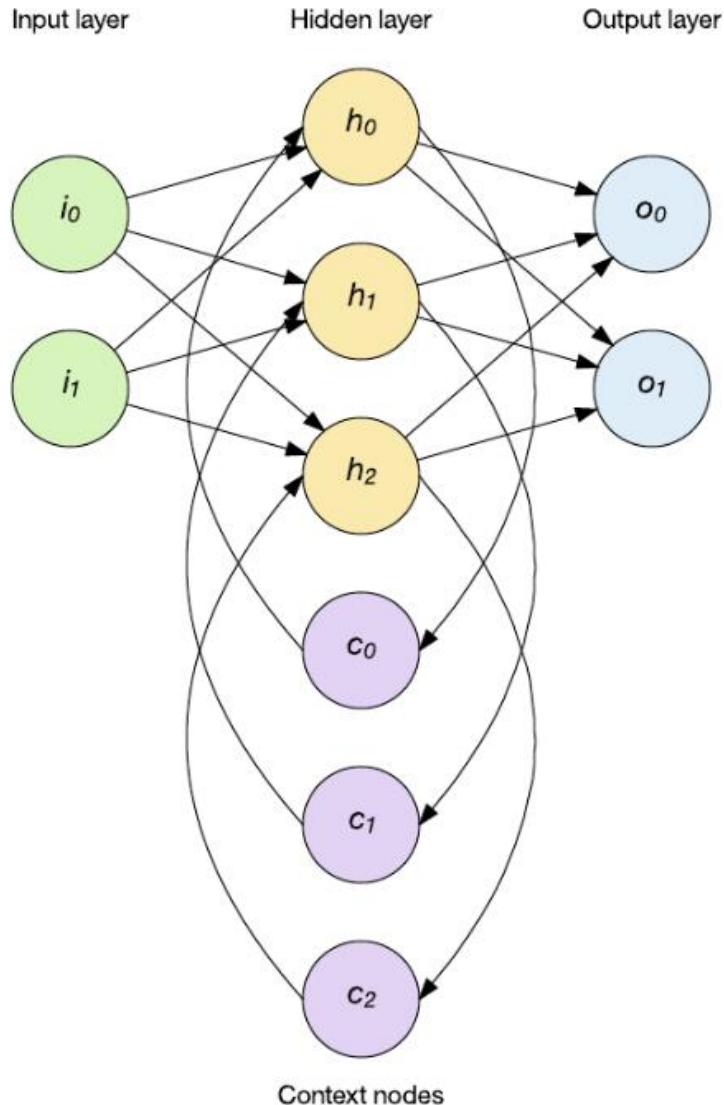


图 5-3 带有回馈的递归神经网络

对于本课题所讨论的定位数据异常检测，动态神经网络起到了先预测当前值的作用，而对于这个预测值，我们只需要和当前实际值进行比较，如果他们相差的幅度达到一定规模，则认为当前值存在异常，应当启动应急方案来应对。

5.2.2 定位数据预测

本章节将使用 MATLAB 软件中的神经网络工具箱，以某地区的过往定位数据时间序列为输入，该地区当前时间节点的定位为输出，依据输入的过往时空数据构建神经网络模型，以时间序列预测方法来进行当前定位数据量预测。借助神经网络的非线性问题处理能力，根据不同的实际情况，对数据进行训练模型，预测当前定位数据值并与实际定位数据对比，验证方法的可行性，如图 5-4 所示。

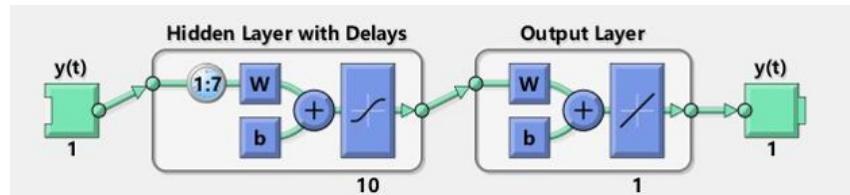


图 5-4 MATLAB 工具箱中非线性自回归模型

在 4.2 的极大似然估计中，我们提到异常的数据会对极大似然估计造成很大的误差，本章我们所讨论的定位数据预测同样需要考虑这个问题，应当将异常的数据进行剔除，使用正常的数据训练网络结构并使用它来预测当前时间节点上的值。接下来的部分和第 4 章一样，将经过预处理后的定位数据导入 MATLAB 中，使用 ntstool 命令进入时间序列工具箱进行训练。采用动态神经网络非线性自回归模型，网络训练时把数据分为三类：训练数据、验证数据和测试数据，三者比例设置为：70%、15%、15%。由于每一个自然周的变化大致相似，所以采用了前 7 个点作为预测的前序值。通过训练数据和验证数据来训练神经网络的模型并自动反向传播调整网络参数，如图 5-5 所示。

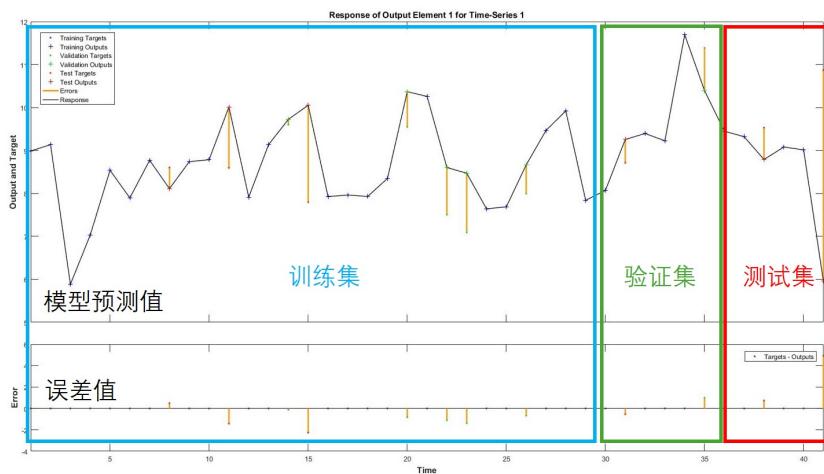


图 5-5 数据的不同部分被用于训练与验证网络

5.2.3 分析结果

如图 5-6 所示, 根据训练的模型, 对 9 月 23 日至 9 月 30 日中的午间实际定位数据进行预测, 可以观察到前几日的预测与实际值差较为相近, 而 9 月 30 日中出现了较大的变化, 预测值与实际值相差极大, 因被归类为异常。

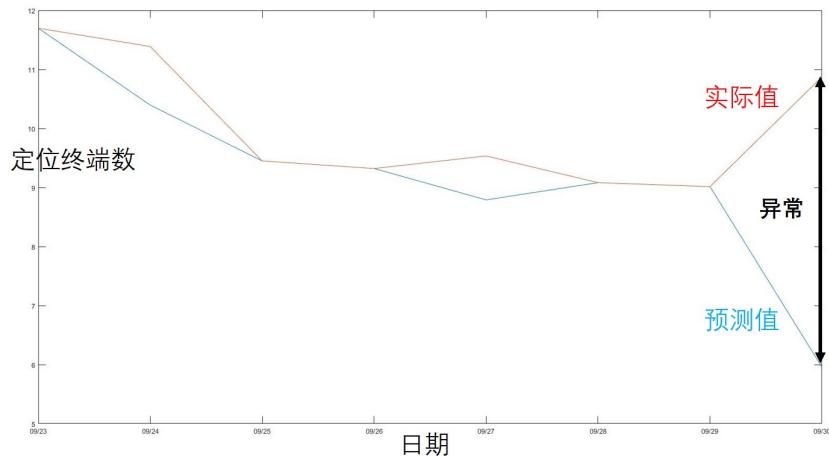


图 5-6 非线性自回归网络被用于预测小时时刻的值来检出异常

又如图 5-7 所示, 针对日期平均数据重新训练模型, 对 8 月 21 日至 8 月 27 日的平均定位数据进行预测, 可以观察到前几日的预测与实际值差较为相近, 而 8 月 23 日中出现了较大的变化, 预测值与实际值相差极大, 说明使用预测的方法也对类似于台风这样的区域性整天的异常有效。

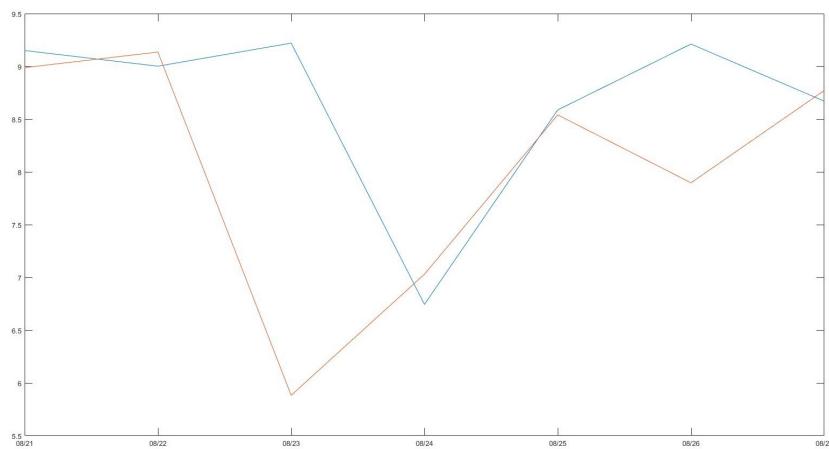


图 5-7 非线性自回归网络被用于预测整天整块区域的平均值来检出异常

本章的实验结果图是反复训练了多次模型才得到的, 可见神经网络对于小规模数据点的规律拟合较差, 当正常数据点的训练样本足够多时, 模型能够进行充分训练, 预测结果将更加优秀, 也更容易衡量实际值与预测值相差多少时被划定为异常的门限。

第六章 基于区域的定位数据异常检测

前述几章我们主要讨论了基于腾讯时序定位数据的异常检测及预测算法，可以对异常发生的整体区域进行分析，在第3章中我们也提及该定位数据中存在大量海面上的无效数据，实际上发生异常的可能只是部分区域。在这一章中，我们根据已知的异常时刻，与正常时刻的定位数据图进行比较，尝试去寻找具体发生异常的子区域。

6.1 基于图像的异常区域检测

6.1.1 相邻帧间差分法

相邻帧间差分法是一种视频分析上检测物体运动的方法。由于摄像机采集的视频序列具有连续性的特点。如果场景内没有运动目标，则连续帧的变化很微弱，如果存在运动目标，则连续的帧和帧之间会有明显地变化。帧间差分法(Temporal Difference)就是借鉴了上述思想。由于场景中的目标在运动，目标的影像在不同图像帧中的位置不同。该类算法对时间上连续的两帧或三帧图像进行差分运算，不同帧对应的像素点相减，判断灰度差的绝对值，当绝对值超过一定阈值时，即可判断为运动目标，从而实现目标的检测功能，如下图所示。



图 6-1 相邻帧间差分法原图

6.1.2 高浮动区域检测

对于本课题的定位数据，帧间差分法所实现的功能即是检测出由于异常导致的子区域定位数据的剧烈抖动。由于台风过境这类异常事件的产生，使得图中某一部分的定位数据相较于正常数据产生较大的变化，这些变化可以使定位数据产生的图像之间差分后



图 6-2 相邻帧间差分法差分图

在图中明显观察得到，之后再用滑动窗的方法去检测具体区域。正常情况下的数据帧间差分效果如图 6-3 所示：

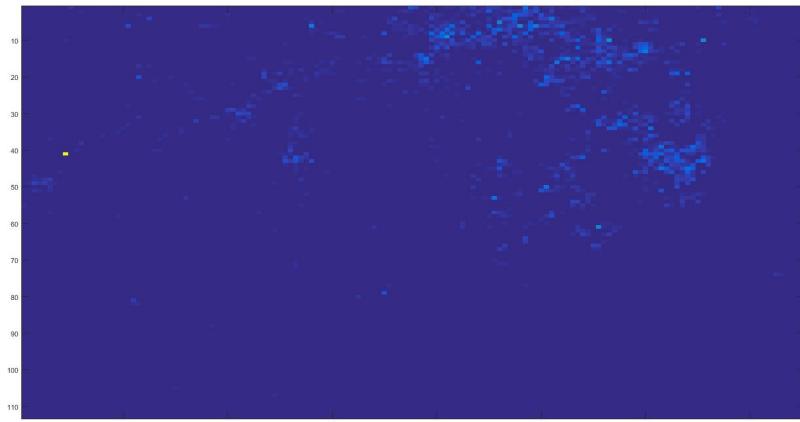


图 6-3 正常定位数据间相邻帧差分图

在第 5 章中我们检测出 9 月 30 日午间的定位数据存在异常，我们将其与 9 月 29 日的同时刻数据进行帧间差分，如图 6-4 所示：

由图可见，存在异常的时刻与正常时刻的定位数据图进行差分会在图中看到明显的“亮”区域，即变化幅值较大。我们再使用滑动窗的方法去检测高浮动区域，使用不同大小的正方形滑动窗对整个范围内进行搜索，自动标记出几个平均密度最高且超过阈值的区域，在图 6-5 中显示如下：

图中所标记的异常浮动值主要集中在城市区域，由于 9 月 30 日是国庆节前一天，可

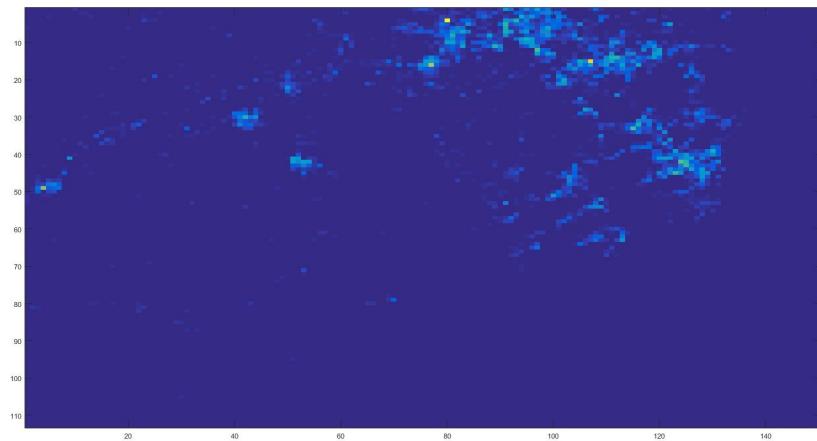


图 6-4 异常与正常定位数据间相邻帧差分图

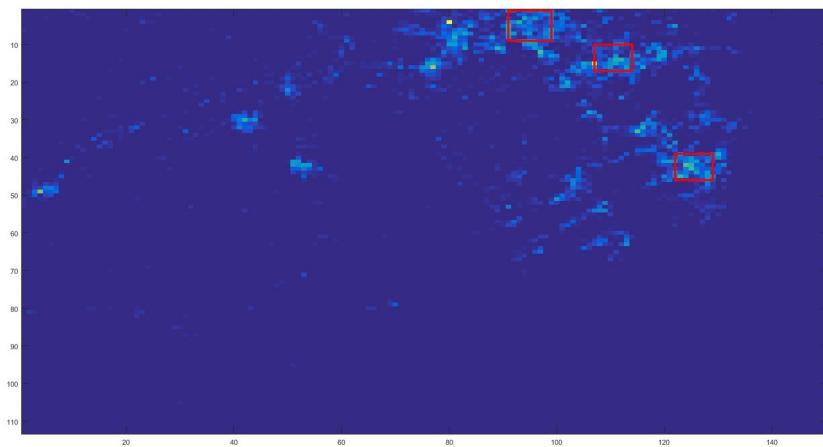


图 6-5 异常与正常定位数据间相邻帧差分图

能是游客的集中所造成的定位数据量增大。

第七章 总结与展望

7.1 内容总结

现代科学的发展，使得传感器网络越来越普及，让大数据分析成为可能。在这些数据之中，有一种数据尤为重要，其存在着很明显的时序性，描述着随着时间的推进中，某种事物的变化规律，这种数据常被称为时序数据。时序数据分析是对时间序列进行系统的分析并且简历合理的模型。其主要目的是考虑数据动态的波动情况并预报未来发生的事件。本文所讨论的定位数据就是基于某固定的空间坐标上的时序数据点，通过这种定位数据的时序分析，可以归纳总结出该地的人流变化特征。而对于存在明显规律的时序数据，我们可以对其进行建模并预测当前或未来时间节点上的值。但是，时序数据中经常会出现这样一些的观测点，它们的数据值较同性质时段上的数据存在着明显的偏差，我们称其为异常点。异常点的分析以及处理十分有意义：一方面异常点对于时间序列的建模来说是干扰很严重的噪音，较大的异常会使得拟合模型造成极大的偏差，需要将这些异常点检出并且删去使得模型的训练及预测更加精准；而另一方面，时序定位数据中的异常又反应出了该时间点上一定存在着某种外界干扰，比如定位数据急剧降低可能是由于台风过境人群大量躲在遮蔽物处而造成的。对于台风这种自然灾害，如果能根据时序定位数据的预测以及异常的判断，则能很好的做好预警，显得尤为重要。本文根据以上思路，基于腾讯的定位数据，对时序异常检测及相关内容完成了以下工作：

1. **数据分析：**对给定的腾讯定位数据进行处理后有了其基本的背景信息，以及为了后续的分析方便进行了预处理。给定的腾讯定位数据大致地理坐标为广东省珠海市沿海一带，而明确探查出的异常是某天的台风过境；由于地图上涵盖一大部分海面，而海面上的定位数据（定位终端数量）几乎为零，分析这些数据毫无意义，故通过最大值判断的方法进行了舍去；最后，通过数据的可视化，以及考虑到存在部分局部时刻异常的情况，我们得出了通过分析每天定位数据平均值来反应整天的定位数据信息，并统计分析图中所有地理坐标上的定位数据信息来确定某天是否存在异常的结论。
2. **基于曲线的异常检测分析：**我们将原本的整块区域异常天检测问题（本课题的异常为台风日）细化为各个有效定位数据点在数据范围日期内的一条曲线，并对整个定位数据点区域的每个点做曲线分析并统计结果；在曲线分析中，我们使用了基于统计、密度、差分、频域的几种曲线分析方法去检测异常并都成功将异常日期找出；最后，基于本课题的定位数据，我们分析了曲线异常检测的几种算法的特点并且讨论了本算法的局限性。
3. **基于曲线的时序数据预测：**由于前述曲线异常检测的算法对于检出局部时刻异常较为困难，我们采用预测并比较的方法去检测异常。神经网络能够经过数据的训

练来对网络结构中的参数进行调整，从而使网络近似趋近于时间序列的实际规律；我们对本课题所研究的腾讯时序定位数据中一部分进行了训练，使用该预测模型对国庆前的某一小时时刻的定位数据进行了预测及判断，预测与实际值有较大偏差，可以检出局部时刻的异常。

4. 基于图像的异常区域检测：本章基于上述两章的检测结果，研究在检测区域中哪一块小区域出现了大浮动从而导致了异常；使用了图像相邻帧间差分法，观察到异常时刻与正常时刻的定位数据差分图较正常时刻之间的差分图颜色更深，并使用了不同大小的矩形窗计算差分定位数据图中平均异常变化幅度，将最大的几块区域圈出。

7.2 未来展望

本文的实验表明，在基于腾讯地图的时序定位数据上，我们能够成功检出已知的台风天异常并建立预测模型达到实时异常检测的目的，并且能够根据异常的时间点找出异常子区域。但对于实际定位数据的分析来说，本文在以下方面可以进行改善：

1. 高地理精度定位数据的异常检测：本文所研究的课题所涉及的定位数据是大区域的粗精度数据，一个像素点约代表平方一千米以上的量级，算法对于分析大范围内的整体异常是有效的。但实际情况中有时研究者所接触到的定位数据是更高精度的，此时本文所讨论的对于整体区域的异常检测不再有效，高精度下应该更关注地理位置上的部分区域，应做更进一步探索。另外，本文对数据所做的简化处理也无法再适用于高精度的数据，无论是时空上或是地理上高精度的定位数据有更多信息可以挖掘，无法直接从统计角度进行整体分析，应当对数据进行更加细致的分析研究可实现的算法。
2. 大规模数据的曲线异常检测：本文所实现的曲线异常检测算法中，部分算法对于本课题所研究的小规模数据量表现不佳，例如极大似然估计法这类统计方法，当数据量足够大时可能会把数据模型的参数拟合的较好，从而能更好地判断数据是否异常。同样，对于第5章所讨论的神经网络时序数据预测，数据量小时可能会导致网络模型拟合不佳，对于数据的拓展预测不利。

7.3 特殊文本类型

7.3.1 脚注

社交媒体是一种供用户创建在线社群来分享信息、观点、个人信息和其它内容（如视频）的电子化交流平台，社交网络服务（social network service, SNS）和微博客（microblog-

ging) 都属于社交媒体的范畴^[?]，国外较为知名的有 Facebook¹、Instagram²、Twitter³、LinkedIn⁴等，国内较为知名的有新浪微博⁵。

在社交媒体的强覆盖下，新闻信息的传播渠道也悄然发生了变化。^[?]

7.3.2 定义、定理与引理等

定义 7.1 这是一条我也不知道在说什么的定义。^[?]

定理 7.1 这是一条我也不知道在说什么的定理。

公理 7.1 这是一条我也不知道在说什么的公理。

引理 7.1 这是一条我也不知道在说什么的引理。

命题 7.1 这是一条我也不知道在说什么的命题。

推论 7.1 这是一条我也不知道在说什么的推论。

7.3.3 中英文文献、学位论文引用

根据美国皮尤研究中心的 2017 年 9 月发布的调查结果^[?]，67% 的美国民众会从社交媒体上获取新闻信息，其中高使用频率用户占 20%。在国内，中国互联网信息中心《2016 年中国互联网新闻市场研究报告》^[?]也显示，社交媒体已逐渐成为新闻获取、评论、转发、跳转的重要渠道，在 2016 年下半年，曾经通过社交媒体获取过新闻资讯的用户比例高达 90.7%，在微信、微博等社交媒体参与新闻评论的比例分别为 62.8% 和 50.2%。社交媒体正在成为网络上热门事件生成并发酵的源头，在形成传播影响力后带动传统媒体跟进报道，最终形成更大规模的舆论浪潮。

在国内，新浪微博由于其发布方便、传播迅速、受众广泛且总量大的特点，成为了虚假信息传播的重灾区：《中国新媒体发展报告（2013）》^[?]显示，2012 年的 100 件微博热点舆情案例中，有超过 1/3 出现谣言；《中国新媒体发展报告（2015）》^[?]对 2014 年传播较广、比较典型的 92 条假新闻进行了多维度分析，发现有 59% 的虚假新闻首发于新浪微博。

此等信息的传播严重损害了有关公众人物的名誉权，降低了社交媒体服务商的商业美誉度，扰乱了网络空间秩序，冲击着网民的认知，极易对民众造成误导，带来诸多麻烦和经济损失，甚至会导致社会秩序的混乱。针对社交媒体谣言采取行动成为了有关部门、服务提供商和广大民众的共同选择。^[?]

¹<http://www.facebook.com/>

²<https://www.instagram.com/>

³<http://www.twitter.com/>

⁴<http://www.linkedin.com/>

⁵<http://www.weibo.com/>

7.4 图表及其引用

此处引用了表 附-1。

表 7-1 基于浏览器行为的特征

特征	描述	形式与理论范围
点赞量	微博的点赞数量	数值, \mathbb{N}
评论量	微博的评论数量	数值, \mathbb{N}
转发量	微博的转发数量	数值, \mathbb{N}

此处引用了一张图。图 附-1 表示的是一个由含有 4 个神经元的输入层、含有 3 个神经元的隐藏层和含有 4 个神经元的输出层组成的自编码器，+1 代表偏置项。

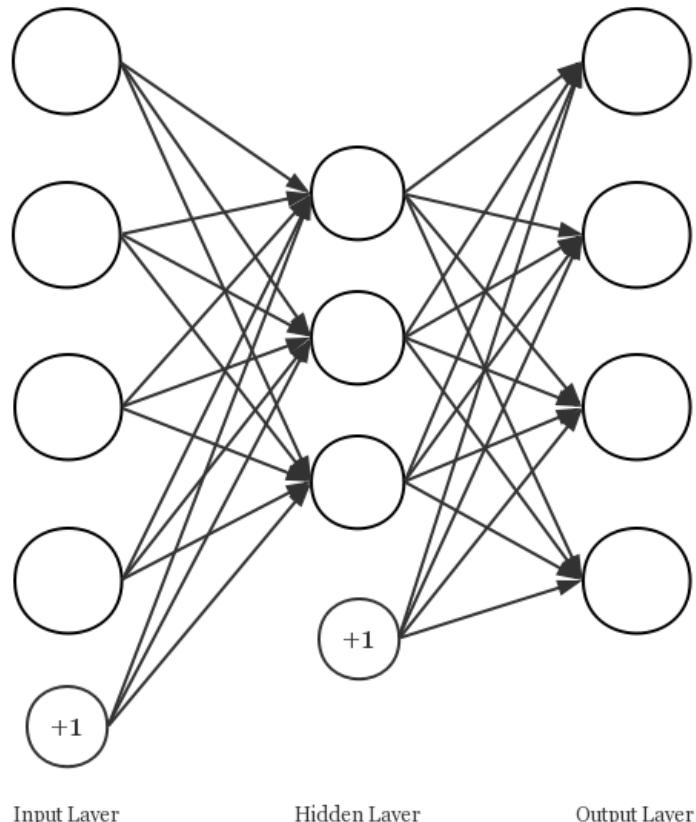


图 7-1 自编码器结构

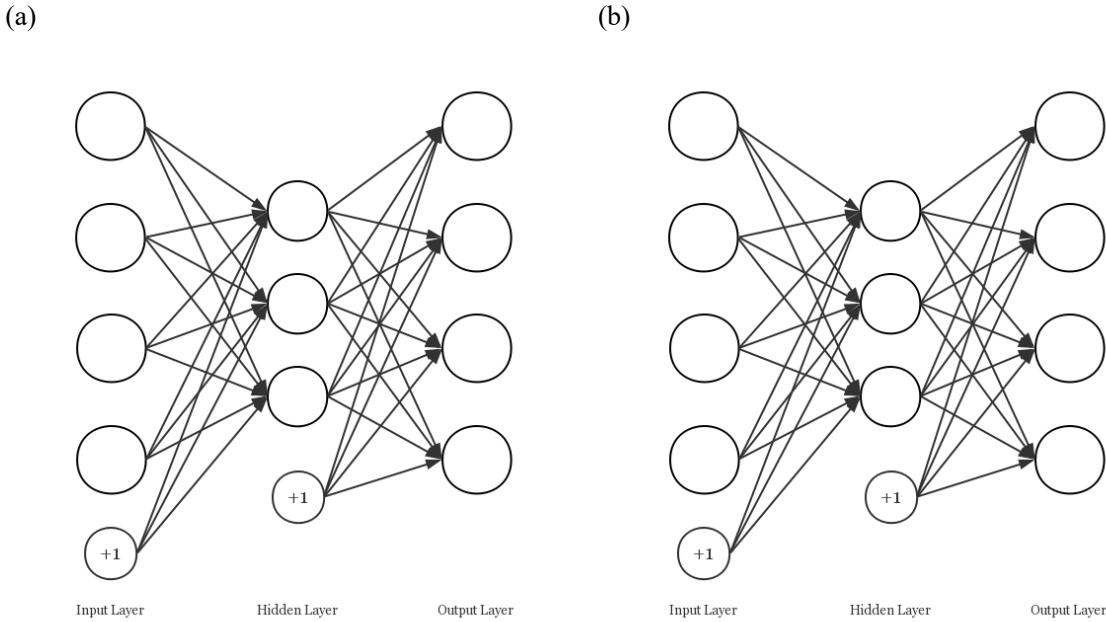


图 7-2 这是两个自编码器结构，我就是排一下子图的效果：(a)左边的自编码器，(b)右边的自编码器

7.5 公式与算法表示

7.5.1 例子：基于主成分分析

7.5.1.1 主成分分析算法

下面对主成分分析进行介绍。

主成分分析是一种简单的机器学习算法，其功能可以从两方面解释：一方面可以认为它提供了一种压缩数据的方式，另一方面也可以认为它是一种学习数据表示的无监督学习算法。^[?] 通过 PCA，我们可以得到一个恰当的超平面及一个投影矩阵，通过投影矩阵，样本点将被投影在这一超平面上，且满足最大可分性（投影后样本点的方差最大化），直观上讲，也就是能尽可能分开。

对中心化后的样本点集 $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_i, \dots, \mathbf{x}_m\}$ (有 $\sum_{i=1}^m \mathbf{x}_i = 0$)，考虑将其最大可分地投影到新坐标系 $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_i, \dots, \mathbf{w}_d\}$ ，其中 \mathbf{w}_i 是标准正交基向量，满足 $\|\mathbf{w}_i\|_2 = 1$, $\mathbf{w}_i^T \mathbf{w}_j = 0$ ($i \neq j$)。假设我们需要 d' ($d' < d$) 个主成分，那么样本点 \mathbf{x}_i 在低维坐标系中的投影是 $\mathbf{z}_i = (z_{i1}; z_{i2}; \dots; z_{id'})$ ，其中 $z_{ij} = \mathbf{w}_j^T \mathbf{x}_i$ ，是 \mathbf{x}_i 在低维

坐标系下第 j 维的坐标。对整个样本集，投影后样本点的方差是

$$\begin{aligned}
 & \frac{1}{m} \sum_{i=1}^m \mathbf{z}_i^\top \mathbf{z}_i \\
 &= \frac{1}{m} \sum_{i=1}^m (\mathbf{x}_i^\top \mathbf{W})^\top (\mathbf{x}_i^\top \mathbf{W}) \\
 &= \frac{1}{m} \sum_{i=1}^m \mathbf{W}^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{W} \\
 &= \frac{1}{m} \mathbf{W}^\top \mathbf{X} \mathbf{X}^\top \mathbf{W}
 \end{aligned} \tag{7-1}$$

由于我们知道新坐标系 \mathbf{W} 的列向量是标准正交基向量，且样本点集 \mathbf{X} 已经过中心化，则 PCA 的优化目标可以写为

$$\begin{aligned}
 \max_{\mathbf{W}} \quad & \text{tr}(\mathbf{W}^\top \mathbf{X} \mathbf{X}^\top \mathbf{W}) \\
 \text{s. t.} \quad & \mathbf{W}^\top \mathbf{W} = \mathbf{I}
 \end{aligned} \tag{7-2}$$

由于 $\mathbf{X} \mathbf{X}^\top$ 是协方差矩阵，那么只需对它做特征值分解，即

$$\mathbf{X}^\top \mathbf{X} = \mathbf{W} \Lambda \mathbf{W}^\top \tag{7-3}$$

其中 $\Lambda = \text{diag}(\boldsymbol{\lambda})$, $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \dots, \lambda_m\}$ 。

具体地，考虑到它是半正定矩阵的二次型，存在最大值，可对式（附-1）使用拉格朗日乘数法

$$\mathbf{X} \mathbf{X}^\top \mathbf{w}_i = \lambda_i \mathbf{w}_i \tag{7-4}$$

之后将求得的特征值降序排列，取前 d' 个特征值对应的特征向量组成所需的投影矩阵 $\mathbf{W}' = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{d'})$ ，即可得到 PCA 的解。PCA 算法的描述如算法1所示。

算法 1 主成分分析 (PCA)

输入： 样本集 $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_i, \dots, \mathbf{x}_m\}$, 低维空间维数 d'

输出： 投影矩阵 $\mathbf{W}' = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{d'})$

- 1: 对所有样本中心化 $\mathbf{x}_i \leftarrow \mathbf{x}_i - \frac{1}{m} \sum_{i=1}^m \mathbf{x}_i$
 - 2: 计算样本的协方差 $\mathbf{X} \mathbf{X}^\top$
 - 3: 对协方差矩阵 $\mathbf{X} \mathbf{X}^\top$ 做特征值分解
 - 4: 取最大的 d' 个特征值所对应的特征向量 $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{d'}$
-

7.5.1.2 主成分分析可信度评估方法

记待判定微博 \mathbf{w}_0 的经典特征向量为 \mathbf{f}_0^c ，它的发布者在 \mathbf{w}_0 前发布的 k 条微博为 $\mathbf{W} = \mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k$ ，这 k 条微博对应的经典特征向量集为 $\mathbf{F}_W^c = \{\mathbf{f}_1^c, \mathbf{f}_2^c, \dots, \mathbf{f}_k^c\}$ 。令

$label = 1$ 代表谣言, $label = 0$ 代表非谣言。算法的具体流程如算法2所示。

算法2 基于PCA的信息可信度评估

输入: $\mathbf{f}_0^c, \mathbf{F}_W^c$, 保留主成分数 n

输出: 标签 $label \in \{0, 1\}$

- 1: 对所有特征向量应用PCA, 保留前 n 个主成分 $\mathbf{o}_i^c \leftarrow PCA(\mathbf{f}_i^c, n)$ ($i = 0, 1, \dots, k$)
 - 2: 计算 \mathbf{F}_W^c 中各向量的平均距离 μ 和标准差 σ
 - 3: 计算阈值 $thr = \mu/\sigma$
 - 4: **if** $\min_{1 < j \leq k} \|\mathbf{o}_0^c - \mathbf{o}_j^c\|_2 > thr$ **then**
 - 5: $label \leftarrow 1$
 - 6: **else**
 - 7: $label \leftarrow 0$
 - 8: **end if**
-

7.6 代码表示

下面的代码7.1是用Python编写的加法函数。

代码 7.1 加法

```
1 def plus_func(a, b):
2     return a + b
```

7.7 列表样式

以下是使用圆点作为项目符号的列表样式。

- 第一章为基础模块示例, 是的, 就是本章。
- 第二章为不存在, 是的, 其实它不存在。

以下是使用数字作为项目符号的列表样式。

1. 第一章为基础模块示例, 是的, 就是本章。
2. 第二章为不存在, 是的, 其实它不存在。

以下是无项目符号(实际是可以自定义一些符号, 但我懒得加了)的列表样式, 它会顶格书写。

第一章为基础模块示例, 是的, 就是本章。

第二章为不存在, 是的, 其实它不存在。

致 谢

此处请写致谢的内容。

它可以有多段。

附录

附录 1 缩略语表

表 附-1 基于浏览器行为的特征

特征	描述	形式与理论范围
点赞量	微博的点赞数量	数值, \mathbb{N}
评论量	微博的评论数量	数值, \mathbb{N}
转发量	微博的转发数量	数值, \mathbb{N}

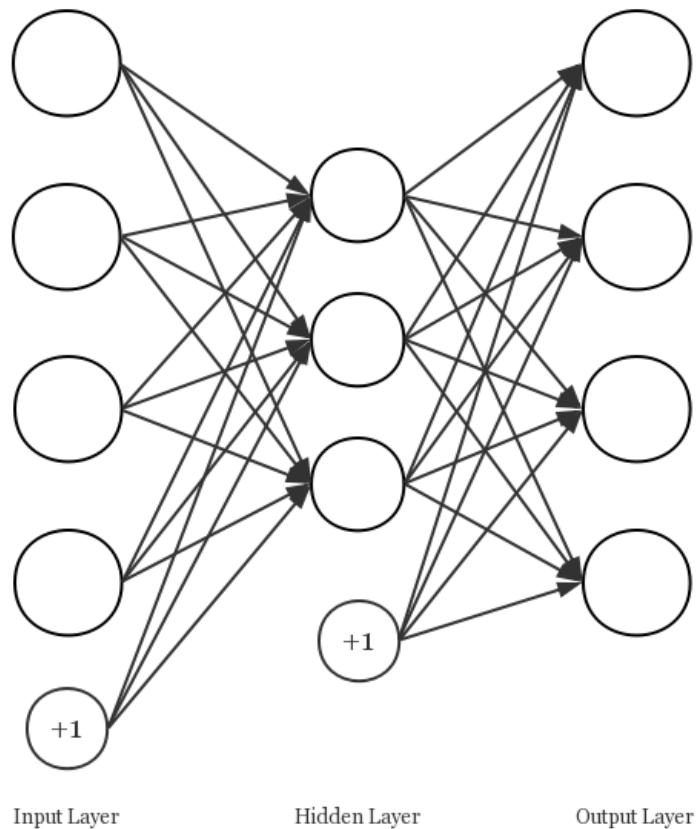


图 附-1 自编码器结构

$$\max_{\mathbf{W}} \operatorname{tr}(\mathbf{W}^T \mathbf{X} \mathbf{X}^T \mathbf{W})$$

式 (附-1)

附录 2 数学符号

数和数组

a	标量 (整数或实数)
\mathbf{a}	向量
$dim()$	向量的维数
A	矩阵
A^T	矩阵 A 的转置
I	单位矩阵 (维度依据上下文而定)
$diag(\mathbf{a})$	对角方阵, 其中对角元素由向量 \mathbf{a} 确定

外 文 译 文

真假新闻的在线传播

Soroush Vosoughi, Deb Roy, Sinan Aral

麻省理工学院

决策、合作、通信和市场领域的基础理论全都将对真实或准确度的概念化作为几乎一切人类努力的核心。然而，不论是真实信息还是虚假信息都会于在线媒体上迅速传播。定义什么是真、什么是假成了一种常见的政治策略，而不是基于一些各方同意的事实的争论。我们的经济也难免遭受虚假信息传播的影响。虚假流言会影响股价和大规模投资的动向，例如，在一条声称巴拉克·奥巴马在爆炸中受伤的推文发布后，股市市值蒸发了 1300 亿美元。的确，从自然灾害到恐怖袭击，我们对一切事情的反应都受到了扰乱。新的社交网络技术在使信息的传播速度变快和规模变大的同时，也便利了不实信息（即不准确或有误导性的信息）的传播。然而，尽管我们对信息和新闻的获取越来越多地收到这些新技术的引导，但我们仍然对他们在虚假信息传播上的作用知之甚少。尽管媒体对假新闻传播的轶事分析给予了相当多的关注，但仍然几乎没有针对不实信息扩散或其发布源头的大规模实证调查。目前，虚假信息传播的研究仅仅局限于小的、局部的样本的分析上，而这些分析忽略了两个最重要的科学问题：真实信息和虚假信息的传播有什么不同？哪些人类判断中的因素可以解释这些不同？

SOCIAL SCIENCE

The spread of true and false news online

Soroush Vosoughi,¹ Deb Roy,¹ Sinan Aral^{2*}

We investigated the differential diffusion of all of the verified true and false news stories distributed on Twitter from 2006 to 2017. The data comprise ~126,000 stories tweeted by ~3 million people more than 4.5 million times. We classified news as true or false using information from six independent fact-checking organizations that exhibited 95 to 98% agreement on the classifications. Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information, and the effects were more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information. We found that false news was more novel than true news, which suggests that people were more likely to share novel information. Whereas false stories inspired fear, disgust, and surprise in replies, true stories inspired anticipation, sadness, joy, and trust. Contrary to conventional wisdom, robots accelerated the spread of true and false news at the same rate, implying that false news spreads more than the truth because humans, not robots, are more likely to spread it.

Foundational theories of decision-making (1–3), cooperation (4), communication (5), and markets (6) all view some conceptualization of truth or accuracy as central to the functioning of nearly every human endeavor. Yet, both true and false information spreads rapidly through online media. Defining what is true and false has become a common political strategy, replacing debates based on a mutually agreed on set of facts. Our economies are not immune to the spread of falsity either. False rumors have affected stock prices and the motivation for large-scale investments, for example, wiping out \$130 billion in stock value after a false tweet claimed that Barack Obama was injured in an explosion (7). Indeed, our responses to everything from natural disasters (8, 9) to terrorist attacks (10) have been disrupted by the spread of false news online.

New social technologies, which facilitate rapid information sharing and large-scale information cascades, can enable the spread of misinformation (i.e., information that is inaccurate or misleading). But although more and more of our access to information and news is guided by these new technologies (11), we know little about their contribution to the spread of falsity online. Though considerable attention has been paid to anecdotal analyses of the spread of false news by the media (12), there are few large-scale empirical investigations of the diffusion of misinformation or its social origins. Studies of the spread of misinformation are currently limited to analyses of small, ad hoc samples that ignore two of the most important scientific questions: How do truth and falsity diffuse differently, and what factors of human judgment explain these differences?

Current work analyzes the spread of single rumors, like the discovery of the Higgs boson (13) or the Haitian earthquake of 2010 (14), and multiple rumors from a single disaster event, like the Boston Marathon bombing of 2013 (10), or it develops theoretical models of rumor diffusion (15), methods for rumor detection (16), credibility evaluation (17, 18), or interventions to curtail the spread of rumors (19). But almost no studies comprehensively evaluate differences in the spread of truth and falsity across topics or examine why false news may spread differently than the truth. For example, although Del Vicario *et al.* (20) and Bessi *et al.* (21) studied the spread of scientific and conspiracy-theory stories, they did not evaluate their veracity. Scientific and conspiracy-theory stories can both be either true or false, and they differ on stylistic dimensions that are important to their spread but orthogonal to their veracity. To understand the spread of false news, it is necessary to examine diffusion after differentiating true and false scientific stories and true and false conspiracy-theory stories and controlling for the topical and stylistic differences between the categories themselves. The only study to date that segments rumors by veracity is that of Friggeri *et al.* (19), who analyzed ~4000 rumors spreading on Facebook and focused more on how fact checking affects rumor propagation than on how falsity diffuses differently than the truth (22).

In our current political climate and in the academic literature, a fluid terminology has arisen around “fake news,” foreign interventions in U.S. politics through social media, and our understanding of what constitutes news, fake news, false news, rumors, rumor cascades, and other related terms. Although, at one time, it may have been appropriate to think of fake news as referring to the veracity of a news story, we now believe that this phrase has been irredeemably polarized in our current political and media climate. As politicians have implemented a political strategy of labeling news sources that do not

support their positions as unreliable or fake news, whereas sources that support their positions are labeled reliable or not fake, the term has lost all connection to the actual veracity of the information presented, rendering it meaningless for use in academic classification. We have therefore explicitly avoided the term fake news throughout this paper and instead use the more objectively verifiable terms “true” or “false” news. Although the terms fake news and misinformation also imply a willful distortion of the truth, we do not make any claims about the intent of the purveyors of the information in our analyses. We instead focus our attention on veracity and stories that have been verified as true or false.

We also purposefully adopt a broad definition of the term news. Rather than defining what constitutes news on the basis of the institutional source of the assertions in a story, we refer to any asserted claim made on Twitter as news (we defend this decision in the supplementary materials section on “reliable sources,” section S1.2). We define news as any story or claim with an assertion in it and a rumor as the social phenomena of a news story or claim spreading or diffusing through the Twitter network. That is, rumors are inherently social and involve the sharing of claims between people. News, on the other hand, is an assertion with claims, whether it is shared or not.

A rumor cascade begins on Twitter when a user makes an assertion about a topic in a tweet, which could include written text, photos, or links to articles online. Others then propagate the rumor by retweeting it. A rumor’s diffusion process can be characterized as having one or more cascades, which we define as instances of a rumor-spreading pattern that exhibit an unbroken retweet chain with a common, singular origin. For example, an individual could start a rumor cascade by tweeting a story or claim with an assertion in it, and another individual could independently start a second cascade of the same rumor (pertaining to the same story or claim) that is completely independent of the first cascade, except that it pertains to the same story or claim. If they remain independent, they represent two cascades of the same rumor. Cascades can be as small as size one (meaning no one retweeted the original tweet). The number of cascades that make up a rumor is equal to the number of times the story or claim was independently tweeted by a user (not retweeted). So, if a rumor “A” is tweeted by 10 people separately, but not retweeted, it would have 10 cascades, each of size one. Conversely, if a second rumor “B” is independently tweeted by two people and each of those two tweets is retweeted 100 times, the rumor would consist of two cascades, each of size 100.

Here we investigate the differential diffusion of true, false, and mixed (partially true, partially false) news stories using a comprehensive data set of all of the fact-checked rumor cascades that spread on Twitter from its inception in 2006 to 2017. The data include ~126,000 rumor cascades spread by ~3 million people more than 4.5 million times. We sampled all rumor cascades investigated by six independent fact-checking organizations

¹Massachusetts Institute of Technology (MIT), the Media Lab, E14-526, 75 Amherst Street, Cambridge, MA 02142, USA. ²MIT, E62-364, 100 Main Street, Cambridge, MA 02142, USA.

*Corresponding author. Email: sinan@mit.edu

(snopes.com, politifact.com, factcheck.org, truthfiction.com, hoax-slayer.com, and urbanlegends.about.com) by parsing the title, body, and verdict (true, false, or mixed) of each rumor investigation reported on their websites and automatically collecting the cascades corresponding to those rumors on Twitter. The result was a sample of rumor cascades whose veracity had been agreed on by these organizations between 95 and 98% of the time. We cataloged the diffusion of the rumor cascades by collecting all English-language replies to tweets that contained a link to any of the aforementioned websites from 2006 to 2017 and used optical character recognition to extract text from images where needed. For each reply tweet, we extracted the original tweet being replied to and all the retweets of the original tweet. Each retweet cascade represents a rumor propagating on Twitter that has been verified as true or false by the fact-checking organizations (see the supplementary materials for more details on cascade construction). We then quantified the cascades'

depth (the number of retweet hops from the origin tweet over time, where a hop is a retweet by a new unique user), size (the number of users involved in the cascade over time), maximum breadth (the maximum number of users involved in the cascade at any depth), and structural virality (23) (a measure that interpolates between content spread through a single, large broadcast and that which spreads through multiple generations, with any one individual directly responsible for only a fraction of the total spread) (see the supplementary materials for more detail on the measurement of rumor diffusion).

As a rumor is retweeted, the depth, size, maximum breadth, and structural virality of the cascade increase (Fig. 1A). A greater fraction of false rumors experienced between 1 and 1000 cascades, whereas a greater fraction of true rumors experienced more than 1000 cascades (Fig. 1B); this was also true for rumors based on political news (Fig. 1D). The total number of false rumors peaked at the end of both 2013 and 2015 and again at the

end of 2016, corresponding to the last U.S. presidential election (Fig. 1E). The data also show clear increases in the total number of false political rumors during the 2012 and 2016 U.S. presidential elections (Fig. 1E) and a spike in rumors that contained partially true and partially false information during the Russian annexation of Crimea in 2014 (Fig. 1E). Politics was the largest rumor category in our data, with ~45,000 cascades, followed by urban legends, business, terrorism, science, entertainment, and natural disasters (Fig. 1F).

When we analyzed the diffusion dynamics of true and false rumors, we found that falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information [Kolmogorov-Smirnov (K-S) tests are reported in tables S3 to S10]. A significantly greater fraction of false cascades than true cascades exceeded a depth of 10, and the top 0.01% of false cascades diffused eight hops deeper into the Twittersphere than the truth, diffusing to depths

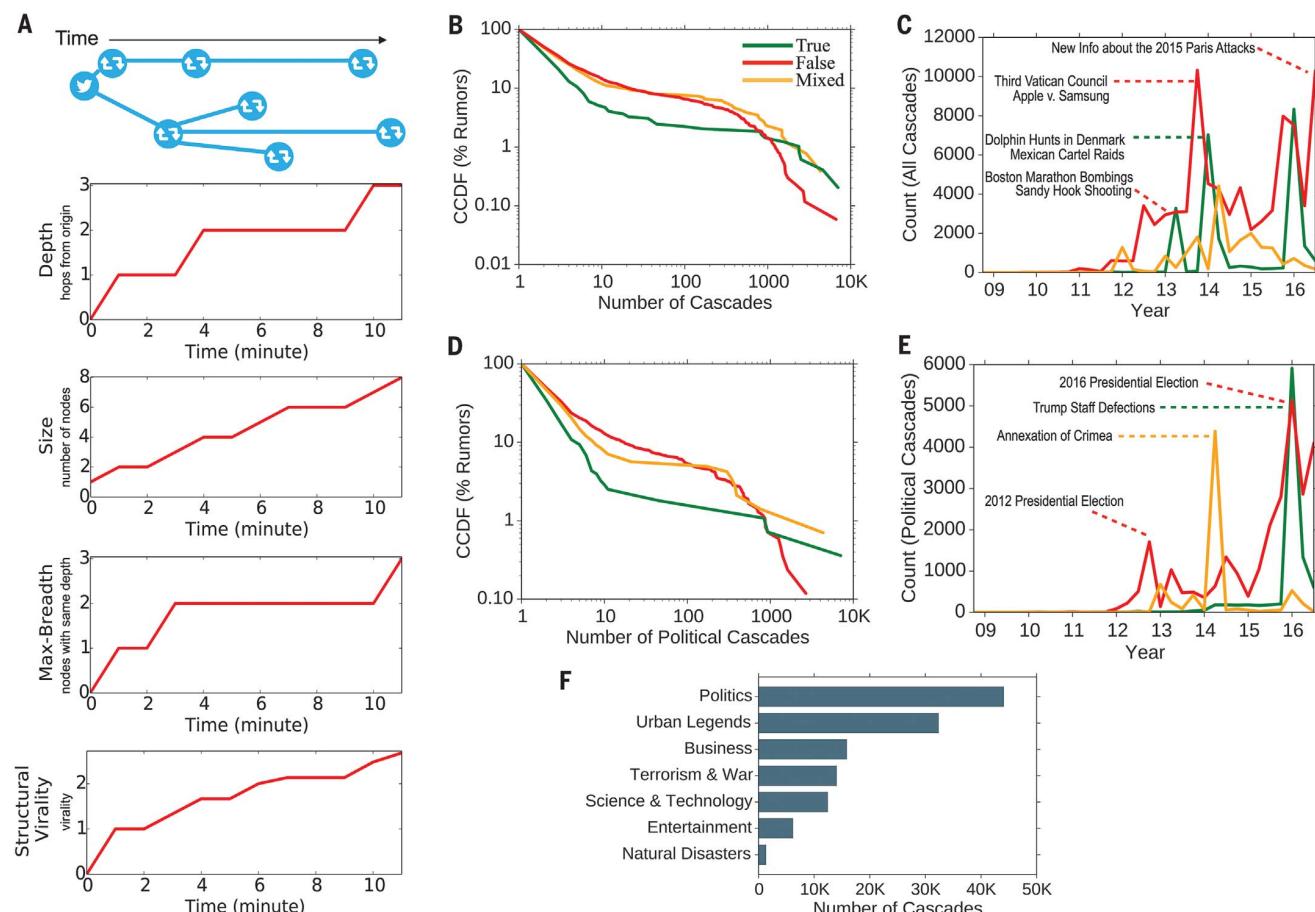


Fig. 1. Rumor cascades. (A) An example rumor cascade collected by our method as well as its depth, size, maximum breadth, and structural virality over time. “Nodes” are users. (B) The complementary cumulative distribution functions (CCDFs) of true, false, and mixed (partially true and partially false) cascades, measuring the fraction of rumors that exhibit a given number of cascades. (C) Quarterly counts of all true, false, and mixed rumor cascades

that diffused on Twitter between 2006 and 2017, annotated with example rumors in each category. (D) The CCDFs of true, false, and mixed political cascades. (E) Quarterly counts of all true, false, and mixed political rumor cascades that diffused on Twitter between 2006 and 2017, annotated with example rumors in each category. (F) A histogram of the total number of rumor cascades in our data across the seven most frequent topical categories.

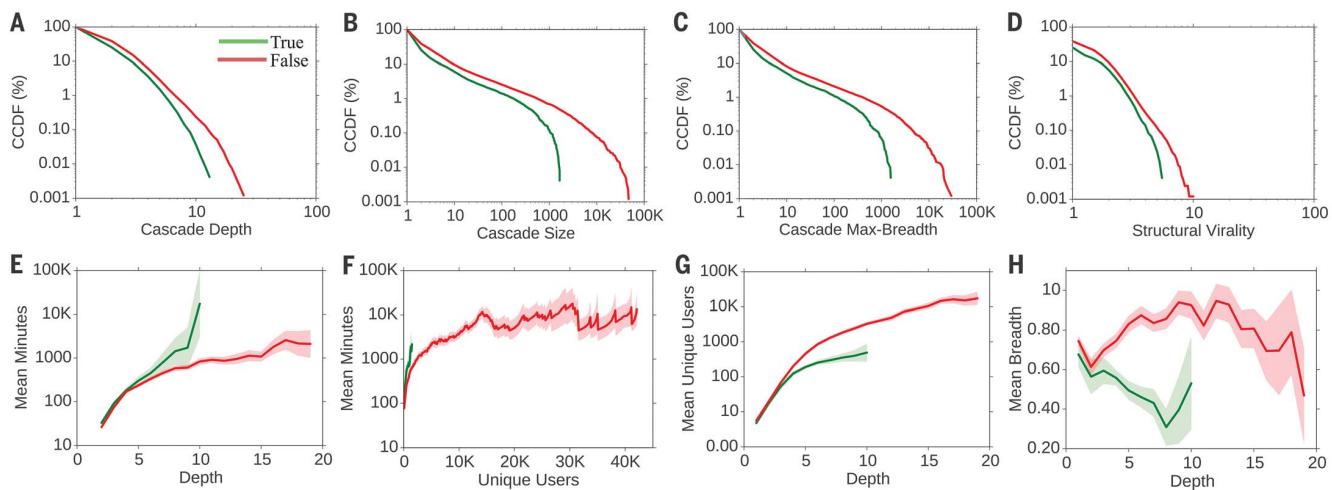


Fig. 2. Complementary cumulative distribution functions (CCDFs) of true and false rumor cascades. (A) Depth. (B) Size. (C) Maximum breadth. (D) Structural virality. (E and F) The number of minutes it takes for true and false rumor cascades to reach any (E) depth and (F) number of unique Twitter users. (G) The number of unique Twitter

users reached at every depth and (H) the mean breadth of true and false rumor cascades at every depth. In (H), plot is lognormal. Standard errors were clustered at the rumor level (i.e., cascades belonging to the same rumor were clustered together; see supplementary materials for additional details).

Downloaded from <http://science/science.org/> on March 10, 2018

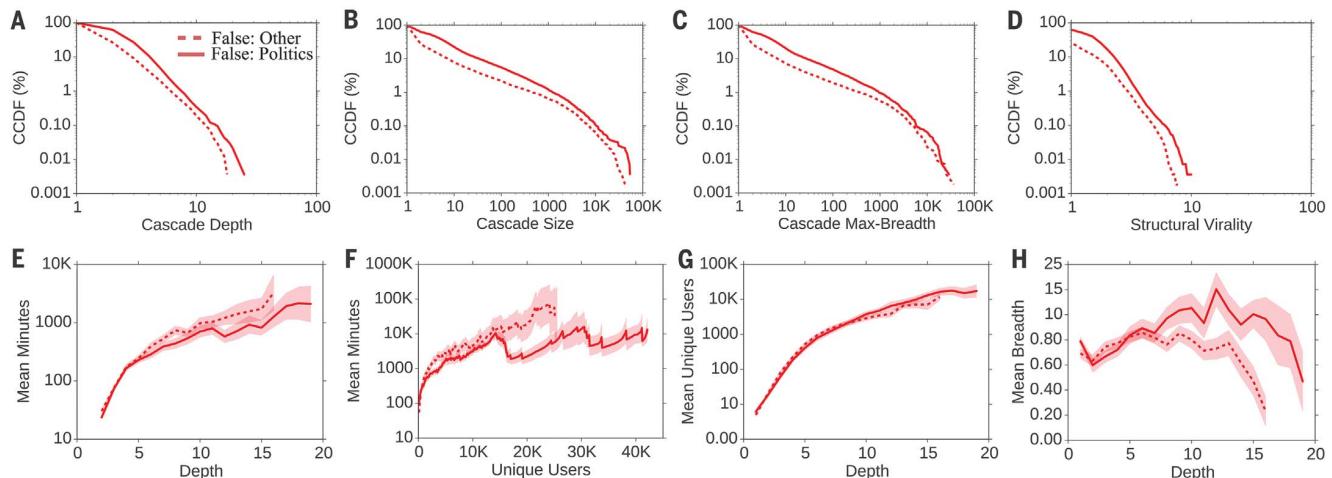


Fig. 3. Complementary cumulative distribution functions (CCDFs) of false political and other types of rumor cascades. (A) Depth. (B) Size. (C) Maximum breadth. (D) Structural virality. (E and F) The number of minutes it takes for false political and other false news cascades to reach

any (E) depth and (F) number of unique Twitter users. (G) The number of unique Twitter users reached at every depth and (H) the mean breadth of these false rumor cascades at every depth. In (H), plot is lognormal. Standard errors were clustered at the rumor level.

greater than 19 hops from the origin tweet (Fig. 2A). Falsehood also reached far more people than the truth. Whereas the truth rarely diffused to more than 1000 people, the top 1% of false-news cascades routinely diffused to between 1000 and 100,000 people (Fig. 2B). Falsehood reached more people at every depth of a cascade than the truth, meaning that many more people retweeted falsehood than they did the truth (Fig. 2C). The spread of falsehood was aided by its virality, meaning that falsehood did not simply spread through broadcast dynamics but rather through peer-to-peer diffusion characterized by a viral branching process (Fig. 2D).

It took the truth about six times as long as falsehood to reach 1500 people (Fig. 2F) and 20 times as long as falsehood to reach a cascade depth of 10 (Fig. 2E). As the truth never diffused beyond a depth of 10, we saw that falsehood reached a depth of 19 nearly 10 times faster than the truth reached a depth of 10 (Fig. 2E). Falsehood also diffused significantly more broadly (Fig. 2H) and was retweeted by more unique users than the truth at every cascade depth (Fig. 2G).

False political news (Fig. 3D) traveled deeper (Fig. 3A) and more broadly (Fig. 3C), reached more people (Fig. 3B), and was more viral than any other category of false information (Fig. 3D). False po-

litical news also diffused deeper more quickly (Fig. 3E) and reached more than 20,000 people nearly three times faster than all other types of false news reached 10,000 people (Fig. 3F). Although the other categories of false news reached about the same number of unique users at depths between 1 and 10, false political news routinely reached the most unique users at depths greater than 10 (Fig. 3G). Although all other categories of false news traveled slightly more broadly at shallower depths, false political news traveled more broadly at greater depths, indicating that more-popular false political news items exhibited broader and more-accelerated diffusion dynamics

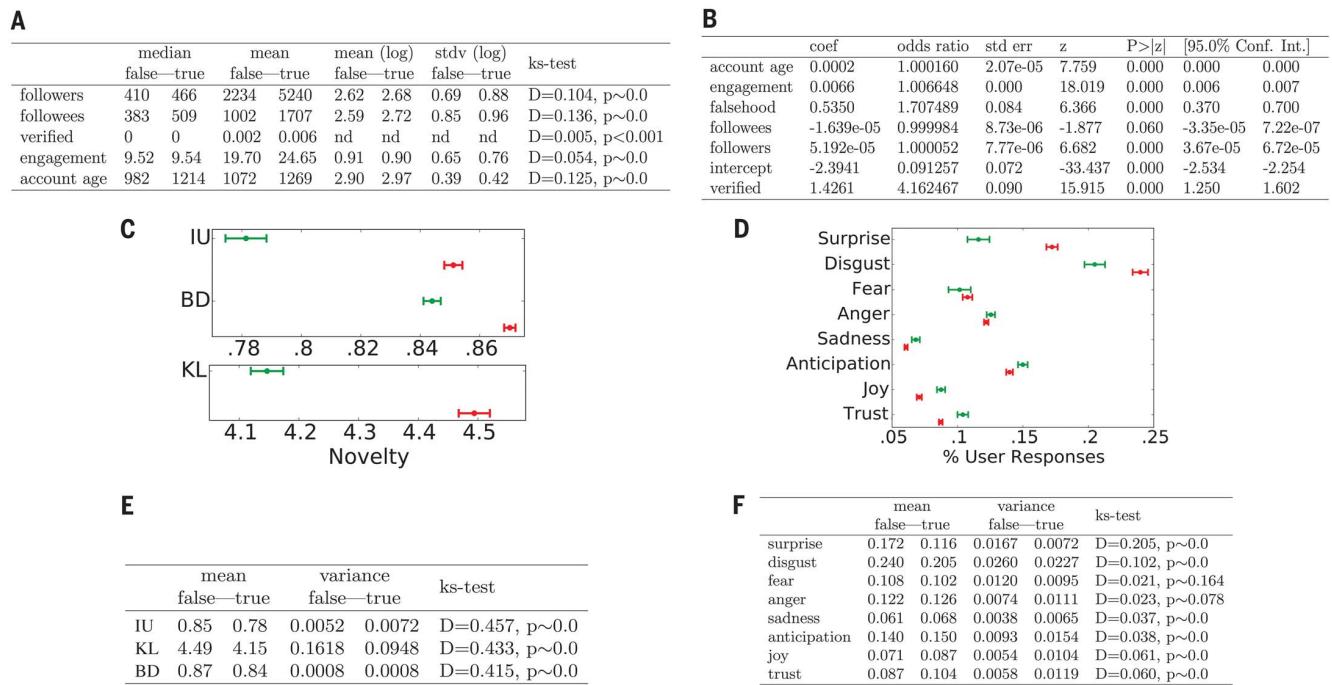


Fig. 4. Models estimating correlates of news diffusion, the novelty of true and false news, and the emotional content of replies to news.

(A) Descriptive statistics on users who participated in true and false rumor cascades as well as K-S tests of the differences in the distributions of these measures across true and false rumor cascades. (B) Results of a logistic regression model estimating users' likelihood of retweeting a rumor as a function of variables shown at the left. coeff, logit coefficient; z, z score. (C) Differences in the information uniqueness (IU), scaled Bhattacharyya distance (BD), and K-L divergence (KL) of true (green) and false (red) rumor tweets compared to the corpus of prior tweets the user was exposed to in the 60 days before retweeting the rumor tweet. (D) The emotional

content of replies to true (green) and false (red) rumor tweets across seven dimensions categorized by the NRC. (E) Mean and variance of the IU, KL, and BD of true and false rumor tweets compared to the corpus of prior tweets the user has seen in the 60 days before seeing the rumor tweet as well as K-S tests of their differences across true and false rumors. (F) Mean and variance of the emotional content of replies to true and false rumor tweets across seven dimensions categorized by the NRC as well as K-S tests of their differences across true and false rumors. All standard errors are clustered at the rumor level, and all models are estimated with cluster-robust standard errors at the rumor level.

(Fig. 3H). Analysis of all news categories showed that news about politics, urban legends, and science spread to the most people, whereas news about politics and urban legends spread the fastest and were the most viral in terms of their structural virality (see fig. S11 for detailed comparisons across all topics).

One might suspect that structural elements of the network or individual characteristics of the users involved in the cascades explain why falsity travels with greater velocity than the truth. Perhaps those who spread falsity "followed" more people, had more followers, tweeted more often, were more often "verified" users, or had been on Twitter longer. But when we compared users involved in true and false rumor cascades, we found that the opposite was true in every case. Users who spread false news had significantly fewer followers (K-S test = 0.104, $P \sim 0.0$), followed significantly fewer people (K-S test = 0.136, $P \sim 0.0$), were significantly less active on Twitter (K-S test = 0.054, $P \sim 0.0$), were verified significantly less often (K-S test = 0.004, $P < 0.001$), and had been on Twitter for significantly less time (K-S test = 0.125, $P \sim 0.0$) (Fig. 4A). Falsehood

diffused farther and faster than the truth despite these differences, not because of them.

When we estimated a model of the likelihood of retweeting, we found that falsehoods were 70% more likely to be retweeted than the truth (Wald chi-square test, $P \sim 0.0$), even when controlling for the account age, activity level, and number of followers and followees of the original tweeter, as well as whether the original tweeter was a verified user (Fig. 4B). Because user characteristics and network structure could not explain the differential diffusion of truth and falsity, we sought alternative explanations for the differences in their diffusion dynamics.

One alternative explanation emerges from information theory and Bayesian decision theory. Novelty attracts human attention (24), contributes to productive decision-making (25), and encourages information sharing (26) because novelty updates our understanding of the world. When information is novel, it is not only surprising, but also more valuable, both from an information theoretic perspective [in that it provides the greatest aid to decision-making (25)] and from a social perspective [in that it conveys so-

cial status on one that is "in the know" or has access to unique "inside" information (26)]. We therefore tested whether falsity was more novel than the truth and whether Twitter users were more likely to retweet information that was more novel.

To assess novelty, we randomly selected ~5000 users who propagated true and false rumors and extracted a random sample of ~25,000 tweets that they were exposed to in the 60 days prior to their decision to retweet a rumor. We then specified a latent Dirichlet Allocation Topic model (27), with 200 topics and trained on 10 million English-language tweets, to calculate the information distance between the rumor tweets and all the prior tweets that users were exposed to before retweeting the rumor tweets. This generated a probability distribution over the 200 topics for each tweet in our data set. We then measured how novel the information in the true and false rumors was by comparing the topic distributions of the rumor tweets with the topic distributions of the tweets to which users were exposed in the 60 days before their retweet. We found that false rumors were significantly more

novel than the truth across all novelty metrics, displaying significantly higher information uniqueness ($K\text{-}S$ test = 0.457, $P \sim 0.0$) (28), Kullback-Leibler ($K\text{-}L$) divergence ($K\text{-}S$ test = 0.433, $P \sim 0.0$) (29), and Bhattacharyya distance ($K\text{-}S$ test = 0.415, $P \sim 0.0$) (which is similar to the Hellinger distance) (30). The last two metrics measure differences between probability distributions representing the topical content of the incoming tweet and the corpus of previous tweets to which users were exposed.

Although false rumors were measurably more novel than true rumors, users may not have perceived them as such. We therefore assessed users' perceptions of the information contained in true and false rumors by comparing the emotional content of replies to true and false rumors. We categorized the emotion in the replies by using the leading lexicon curated by the National Research Council Canada (NRC), which provides a comprehensive list of ~140,000 English words and their associations with eight emotions based on Plutchik's (31) work on basic emotion—anger, fear, anticipation, trust, surprise, sadness, joy, and disgust (32)—and a list of ~32,000 Twitter hashtags and their weighted associations with the same emotions (33). We removed stop words and URLs from the reply tweets and calculated the fraction of words in the tweets that related to each of the eight emotions, creating a vector of emotion weights for each reply that summed to one across the emotions. We found that false rumors inspired replies expressing greater surprise ($K\text{-}S$ test = 0.205, $P \sim 0.0$), corroborating the novelty hypothesis, and greater disgust ($K\text{-}S$ test = 0.102, $P \sim 0.0$), whereas the truth inspired replies that expressed greater sadness ($K\text{-}S$ test = 0.037, $P \sim 0.0$), anticipation ($K\text{-}S$ test = 0.038, $P \sim 0.0$), joy ($K\text{-}S$ test = 0.061, $P \sim 0.0$), and trust ($K\text{-}S$ test = 0.060, $P \sim 0.0$) (Fig. 4, D and F). The emotions expressed in reply to falsehoods may illuminate additional factors, beyond novelty, that inspire people to share false news. Although we cannot claim that novelty causes retweets or that novelty is the only reason why false news is retweeted more often, we do find that false news is more novel and that novel information is more likely to be retweeted.

Numerous diagnostic statistics and manipulation checks validated our results and confirmed their robustness. First, as there were multiple cascades for every true and false rumor, the variance of and error terms associated with cascades corresponding to the same rumor will be correlated. We therefore specified cluster-robust standard errors and calculated all variance statistics clustered at the rumor level. We tested the robustness of our findings to this specification by comparing analyses with and without clustered errors and found that, although clustering reduced the precision of our estimates as expected, the directions, magnitudes, and significance of our results did not change, and chi-square ($P \sim 0.0$) and deviance (d) goodness-of-fit tests ($d = 3.4649 \times 10^{-6}$, $P \sim 1.0$) indicate that the models are well specified (see supplementary materials for more detail).

Second, a selection bias may arise from the restriction of our sample to tweets fact checked by the six organizations we relied on. Fact checking may select certain types of rumors or draw additional attention to them. To validate the robustness of our analysis to this selection and the generalizability of our results to all true and false rumor cascades, we independently verified a second sample of rumor cascades that were not verified by any fact-checking organization. These rumors were fact checked by three undergraduate students at Massachusetts Institute of Technology (MIT) and Wellesley College. We trained the students to detect and investigate rumors with our automated rumor-detection algorithm running on 3 million English-language tweets from 2016 (34). The undergraduate annotators investigated the veracity of the detected rumors using simple search queries on the web. We asked them to label the rumors as true, false, or mixed on the basis of their research and to discard all rumors previously investigated by one of the fact-checking organizations. The annotators, who worked independently and were not aware of one another, agreed on the veracity of 90% of the 13,240 rumor cascades that they investigated and achieved a Fleiss' kappa of 0.88. When we compared the diffusion dynamics of the true and false rumors that the annotators agreed on, we found results nearly identical to those estimated with our main data set (see fig. S17). False rumors in the robustness data set had greater depth ($K\text{-}S$ test = 0.139, $P \sim 0.0$), size ($K\text{-}S$ test = 0.131, $P \sim 0.0$), maximum breadth ($K\text{-}S$ test = 0.139, $P \sim 0.0$), structural virality ($K\text{-}S$ test = 0.066, $P \sim 0.0$), and speed (fig. S17) and a greater number of unique users at each depth (fig. S17). When we broadened the analysis to include majority-rule labeling, rather than unanimity, we again found the same results (see supplementary materials for results using majority-rule labeling).

Third, although the differential diffusion of truth and falsity is interesting with or without robot, or bot, activity, one may worry that our conclusions about human judgment may be biased by the presence of bots in our analysis. We therefore used a sophisticated bot-detection algorithm (35) to identify and remove all bots before running the analysis. When we added bot traffic back into the analysis, we found that none of our main conclusions changed—false news still spread farther, faster, deeper, and more broadly than the truth in all categories of information. The results remained the same when we removed all tweet cascades started by bots, including human retweets of original bot tweets (see supplementary materials, section S8.3) and when we used a second, independent bot-detection algorithm (see supplementary materials, section S8.3.5) and varied the algorithm's sensitivity threshold to verify the robustness of our analysis (see supplementary materials, section S8.3.4). Although the inclusion of bots, as measured by the two state-of-the-art bot-detection algorithms we used in our analysis, accelerated the spread of both true and false news, it affected their spread roughly equally. This suggests that false

news spreads farther, faster, deeper, and more broadly than the truth because humans, not robots, are more likely to spread it.

Finally, more research on the behavioral explanations of differences in the diffusion of true and false news is clearly warranted. In particular, more robust identification of the factors of human judgment that drive the spread of true and false news online requires more direct interaction with users through interviews, surveys, lab experiments, and even neuroimaging. We encourage these and other approaches to the investigation of the factors of human judgment that drive the spread of true and false news in future work.

False news can drive the misallocation of resources during terror attacks and natural disasters, the misalignment of business investments, and misinformed elections. Unfortunately, although the amount of false news online is clearly increasing (Fig. 1, C and E), the scientific understanding of how and why false news spreads is currently based on ad hoc rather than large-scale systematic analyses. Our analysis of all the verified true and false rumors that spread on Twitter confirms that false news spreads more pervasively than the truth online. It also overturns conventional wisdom about how false news spreads. Though one might expect network structure and individual characteristics of spreaders to favor and promote false news, the opposite is true. The greater likelihood of people to retweet falsity more than the truth is what drives the spread of false news, despite network and individual factors that favor the truth. Furthermore, although recent testimony before congressional committees on misinformation in the United States has focused on the role of bots in spreading false news (36), we conclude that human behavior contributes more to the differential spread of falsity and truth than automated robots do. This implies that misinformation-containment policies should also emphasize behavioral interventions, like labeling and incentives to dissuade the spread of misinformation, rather than focusing exclusively on curtailing bots. Understanding how false news spreads is the first step toward containing it. We hope our work inspires more large-scale research into the causes and consequences of the spread of false news as well as its potential cures.

REFERENCES AND NOTES

1. L. J. Savage, *J. Am. Stat. Assoc.* **46**, 55–67 (1951).
2. H. A. Simon, *The New Science of Management Decision* (Harper & Brothers Publishers, New York, 1960).
3. R. Wedgwood, *Noûs* **36**, 267–297 (2002).
4. E. Fehr, U. Fischbacher, *Nature* **425**, 785–791 (2003).
5. C. E. Shannon, *Bell Syst. Tech. J.* **27**, 379–423 (1948).
6. S. Bikhchandani, D. Hirshleifer, I. Welch, *J. Polit. Econ.* **100**, 992–1026 (1992).
7. K. Rapoza, "Can 'fake news' impact the stock market?" *Forbes*, 26 February 2017; www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/.
8. M. Mendoza, B. Poblete, C. Castillo, in *Proceedings of the First Workshop on Social Media Analytics* (Association for Computing Machinery, ACM, 2010), pp. 71–79.
9. A. Gupta, H. Lamba, P. Kumaraguru, A. Joshi, in *Proceedings of the 22nd International Conference on World Wide Web* (ACM, 2010), pp. 729–736.

10. K. Starbird, J. Maddock, M. Orand, P. Achterman, R. M. Mason, in *iConference 2014 Proceedings* (iSchools, 2014).
11. J. Gottfried, E. Shearer, "News use across social media platforms," Pew Research Center, 26 May 2016; www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/.
12. C. Silverman, "This analysis shows how viral fake election news stories outperformed real news on Facebook," *BuzzFeed News*, 16 November 2016; www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook/.
13. M. De Domenico, A. Lima, P. Mougel, M. Musolesi, *Sci. Rep.* **3**, 2980 (2013).
14. O. Oh, K. H. Kwon, H. R. Rao, in *Proceedings of the International Conference on Information Systems* (International Conference on Information Systems, ICIIS, paper 231, 2010).
15. M. Tambusco, G. Ruffo, A. Flammini, F. Menczer, in *Proceedings of the 24th International Conference on World Wide Web* (ACM, 2015), pp. 977–982.
16. Z. Zhao, P. Resnick, Q. Mei, in *Proceedings of the 24th International Conference on World Wide Web* (ACM, 2015), pp. 1395–1405.
17. M. Gupta, P. Zhao, J. Han, in *Proceedings of the 2012 Society for Industrial and Applied Mathematics International Conference on Data Mining* (Society for Industrial and Applied Mathematics, SIAM, 2012), pp. 153–164.
18. G. L. Ciampaglia et al., *PLOS ONE* **10**, e0128193 (2015).
19. A. Frigeri, L. A. Adamic, D. Eckles, J. Cheng, in *Proceedings of the International Conference on Weblogs and Social Media* (Association for the Advancement of Artificial Intelligence, AAAI, 2014).
20. M. Del Vicario et al., *Proc. Natl. Acad. Sci. U.S.A.* **113**, 554–559 (2016).
21. A. Bessi et al., *PLOS ONE* **10**, e0118093 (2015).
22. Frigeri et al. (19) do evaluate two metrics of diffusion: depth, which shows little difference between true and false rumors, and shares per rumor, which is higher for true rumors than it is for false rumors. Although these results are important, they are not definitive owing to the smaller sample size of the study; the early timing of the sample, which misses the rise of false news after 2013; and the fact that more shares per rumor do not necessarily equate to deeper, broader, or more rapid diffusion.
23. S. Goel, A. Anderson, J. Hofman, D. J. Watts, *Manage. Sci.* **62**, 180–196 (2015).
24. L. Itti, P. Baldi, *Vision Res.* **49**, 1295–1306 (2009).
25. S. Aral, M. Van Alstyne, *Am. J. Sociol.* **117**, 90–171 (2011).
26. J. Berger, K. L. Milkman, *J. Mark. Res.* **49**, 192–205 (2012).
27. D. M. Blei, A. Y. Ng, M. I. Jordan, *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
28. S. Aral, P. Dhillon, "Unpacking novelty: The anatomy of vision advantages," Working paper, MIT-Sloan School of Management, Cambridge, MA, 22 June 2016; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2388254.
29. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (Wiley, 2012).
30. T. Kailath, *IEEE Trans. Commun. Technol.* **15**, 52–60 (1967).
31. R. Plutchik, *Am. Sci.* **89**, 344–350 (2001).
32. S. M. Mohammad, P. D. Turney, *Comput. Intell.* **29**, 436–465 (2013).
33. S. M. Mohammad, S. Kiritchenko, *Comput. Intell.* **31**, 301–326 (2015).
34. S. Vosoughi, D. Roy, in *Proceedings of the 10th International AAAI Conference on Weblogs and Social Media* (AAAI, 2016), pp. 707–710.
35. C. A. Davis, O. Varol, E. Ferrara, A. Flammini, F. Menczer, in *Proceedings of the 25th International Conference Companion on World Wide Web* (ACM, 2016), pp. 273–274.
36. For example, this is an argument made in recent testimony by Clint Watts—Robert A. Fox Fellow at the Foreign Policy

Research Institute and Senior Fellow at the Center for Cyber and Homeland Security at George Washington University—given during the U.S. Senate Select Committee on Intelligence hearing on "Disinformation: A Primer in Russian Active Measures and Influence Campaigns" on 30 March 2017; www.intelligence.senate.gov/sites/default/files/documents/os-cwatts-033017.pdf.

ACKNOWLEDGMENTS

We are indebted to Twitter for providing funding and access to the data. We are also grateful to members of the MIT research community for invaluable discussions. The research was approved by the MIT institutional review board. The analysis code is freely available at <https://goo.gl/forms/AK1lZujpepxhNTy33>. The entire data set is also available, from the same link, upon signing an access agreement stating that (i) you shall only use the data set for the purpose of validating the results of the MIT study and for no other purpose; (ii) you shall not attempt to identify, reidentify, or otherwise deanonymize the data set; and (iii) you shall not further share, distribute, publish, or otherwise disseminate the data set. Those who wish to use the data for any other purposes can contact and make a separate agreement with Twitter.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/359/6380/1146/suppl/DC1
Materials and Methods
Figs. S1 to S20
Tables S1 to S39
References (37–75)

14 September 2017; accepted 19 January 2018
10.1126/science.aap9559

北京邮电大学

本科毕业设计（论文）开题报告

学院	信息与通信工程学院	专业	通信工程	班级	201421119
学生姓名	猜猜	学号	2014210999	班内序号	99
指导教师姓名	猜猜	所在单位	信息与通信工程学院	职称	教授
设计（论文） 题目	(中文) 猜猜看毕设题目是什么 (英文) Just Guess What On Earth My Title is				

一、选题背景及意义

社交多媒体(social multimedia)是多媒体数据(multimedia)与社交媒体(social media)相结合的新型媒体形式。它是互联网技术发展过程中，人们对多样的媒体内容和新型的交互模式的需求中产生的。其中，多媒体数据极大地丰富了纯文本内容，而社会媒体网络提供了快速交流、传播多媒体内容的高效平台，两者相互转化。全世界内，最引人注目的社交媒体平台当属微博客(Microblog)，其中以中文的新浪微博和英文的Twitter最为活跃，各平台每时每刻产生并流动着种类繁多的大量信息。

微博客平台有着发布方便、传播迅速、受众广泛且总量大的特点。这种特点使得更多的官方媒体将其作为资讯发布的重要平台，同时更多的普通用户将其作为获取热点信息的重要来源。然而，在加速真实信息的有效传播的同时，微博客平台也成了虚假消息的温床，这一现象在社会和科学健康类话题中表现突出：在重大事件、突发事件和灾害事故消息等社会类话题中，虚假信息的传播严重扰乱了网络空间秩序，冲击着网民的认知，有的甚至导致了社会秩序的混乱(如日本福岛核电站泄露事件发生后我国的食用盐哄抢事件)和事件走向的转变(如2016年的美国总统选举)；在科学健康类话题中，耸人听闻的食品安全曝光(如“塑料紫菜”、“棉花肉松”)、不科学的食品安全警告(如“柿子和酸奶一起吃会中毒致死”)和错误的医疗手段(如“一滴血就能验癌”)极易对人们的认知造成误导，进一步带来不必要的麻烦和相应的经济冲击。

二、研究的基本内容

对所提出算法进行性能的测试、比较和分析，针对结论面向未来发展方向进行探讨。

三、 研究方法及措施

从数据分布的角度上讲，检测谣言的这一类问题非常适合归入数据挖掘的经典问题——异常检测（anomaly detection）或离群点检测（outlier detection），一方面是因为谣言的种类繁多，若归入一大类，其与正常信息的边界可能会难以寻找；另一方面是即便虚假信息被认为泛滥成灾，但谣言在微博空间中仍是少数，可获取的谣言和非谣言比例失衡。

四、 研究工作的步骤与进度

2018.1.1 ~ 2018.2.10 完成领域内容调研，模板对应部分撰写。

2018.2.28~2018.4.15 完成相关模板研究，设计模板。

2018.4.16~2018.4.30 进行模板设计评估和比较分析。

2018.5.1~2018.5.15 模板整体撰写。

五、 主要参考文献

Zubiaga A, Aker A, Bontcheva K, et al. Detection and Resolution of Rumours in Social Media: A Survey[J]. ACM Computing Surveys (CSUR), 2018, 51(2): 32.

Savage D, Zhang X, Yu X, et al. Anomaly detection in online social networks[J]. Social Networks, 2014, 39(1):62-70.

Castillo C, Mendoza M, Poblete B. Information credibility on twitter[C]// International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April. DBLP, 2011:675-684.

Jin Z, Cao J, Guo H, et al. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs[C]//Proceedings of the 2017 ACM on Multimedia Conference. ACM, 2017: 795-816.

指导教师签字		日期	年 月 日
--------	--	----	-------

注：可根据开题报告的长度加页。

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	2014211199				
学生姓名	猜猜				
学号	2014210999				
指导教师姓名					
所在单位					
职称					
设计（论文）题目	(中文) 猜猜看毕设题目是什么				
	(英文) Just Guess What On Earth My Title is				
目前已完成任务	<p>截至中期检查前夕，本课题已经完成的工作如下：</p> <p>完成有关实验。</p> <p>实验结束后，对整体准确率（Accuracy）进行了统计，还得到了谣言和非谣言的精度（Precision）、召回率（Recall）和F1值（F1-Score）。</p>				
	是否符合任务书要求进度 是				
尚需完成的任务	<ul style="list-style-type: none"> • 完成整体架构和论文书写任务。 • 完成外文文献的翻译。 				
	能否按期完成设计（论文） 能				
存在问题和解决办法	存在	模型中存在一些欠缺讨论分析的细节，如阈值选择等。			
	拟采取的办法	拟补充部分实验和查阅领域经验进行讨论分析。			
指导教师签字		日期	年 月 日		

检查小 组意见	
负责人签字: 年 月 日	

注: 可根据长度加页。