

# 大学生论文检测系统

## 文本复制检测报告单 (全文标明引文)

№: ADBD2022R\_2022053017300120220530174008466367518827

检测时间: 2022-05-30 17:40:08

篇名: 96-180110704-段裕-《新冠疫情时空数据的收集与建模分析》-计算机科学与技术学院-冯山山

作者: 段裕

指导教师:

检测机构: 哈尔滨工业大学(深圳)

提交论文IP: 219.\*\*\*.\*\*\*.\*\*\*

文件名: 96-180110704-段裕-《新冠疫情时空数据的收集与建模分析》-计算机科学与技术学院-冯山山.docx

检测系统: 大学生论文检测系统

检测类型: 大学生论文

检测范围: 中国学术期刊网络出版总库

中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库

中国重要会议论文全文数据库

中国重要报纸全文数据库

中国专利全文数据库

图书资源

优先出版文献库

大学生论文联合比对库

互联网资源(包含贴吧等论坛资源)

英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)

港澳台学术文献库

互联网文档资源

源代码库

CNKI大成编客-原创作品库

机构自建比对库

时间范围: 1900-01-01至2022-05-30

### 检测结果

去除本人文献复制比: 14.1%

跨语言检测结果: 0%

去除引用文献复制比: 10.8%

总文字复制比: 14.1%

单篇最大文字复制比: 3.2% (传染病传播模型综述)

重复字数: [3686]

总段落数: [5]

总字数: [26182]

疑似段落数: [5]

单篇最大重复字数: [842]

前部重合字数: [1393]

疑似段落最大重合字数: [1651]

后部重合字数: [2293]

疑似段落最小重合字数: [76]



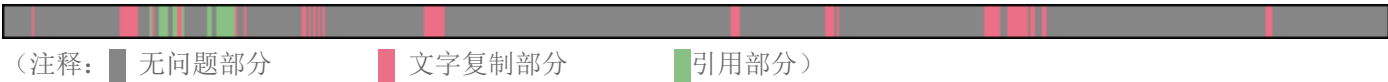
指标: ☐ 疑似剽窃观点 ☒ 疑似剽窃文字表述 ☐ 疑似整体剽窃 ☐ 过度引用

相似表格: 0 相似公式: 没有公式 疑似文字的图片: 0

3.4% (76) 3.4% (76) 中英文摘要等 (总2241字)

28.1% (1651) 28.1% (1651) 第1章绪论 (总5868字)

7.7% (562) 7.7% (562) 第2章新冠疫情时空数据的获取与可视化 (总7336字)



指导教师审查结果

指导教师:

审阅结果:

审阅意见: 指导老师未填写审阅意见

1. 中英文摘要等			总字数: 2241
相似文献列表			
去除本人文献复制比: 3.4%(76)      文字复制比: 3.4%(76)      疑似剽窃观点: (0)			
1	非贵金属基3DOM（双）钙钛矿型催化剂催化炭烟燃烧性能研究	3.4% (76)	
	梅雪垒(导师: 韦岳长) - 《中国石油大学(北京) 硕士学位论文》- 2020-06-30	是否引证: 否	

原文内容	
新冠疫情时空数据的收集与建模分析 段裕 学院: 计算机科学与技术学院专业: 计算机科学与技术 学号: 180110704 指导教师: 冯山山 2022年6月 哈尔滨工业大学深圳校区 毕业设计（论文） 题目新冠疫情时空数据的收集与建模分析 姓名段裕 学号 180110704 学院计算机科学与技术学院 专业计算机科学与技术专业 指导教师冯山山 答辩日期 2022年06月09日 摘要 2019年末爆发的新型冠状病毒肺炎对人们的生产生活和身心健康造成了极大的影响和危害，因此展开对中国新冠疫情时空数据的收集与分析十分必要。项目收集和整理了自2020年01月至2022年03月期间中	
题目	新冠疫情时空数据的收集与建模分析
姓名	段裕
学号	180110704
学院	计算机科学与技术学院
专业	计算机科学与技术专业
指导教师	冯山山
答辩日期	2022年06月09日

国的新冠疫情数据，同时展开了数据集可视化工作，并从不同的角度对数据集进行了建模分析。本文首先从传统的流行病动力学角度进行建模，使用到了SEIR系列模型；接着考虑数据的时序特点，从长短期记忆网络的角度进行了研究；最后进一步从时空点过程的角度进行了实验分析。实验结果表明，新冠疫情的发展趋势总体较为符合SEIR系列模型的假设，可以通过SEIR系列模型来模拟和预测疫情爆发初期的发展趋势；在训练样本数据充足时，长短期记忆网络模型能够达到更优的性能，但在训练样本不足时，长短期记忆网络无法有效的学习到疫情发展的趋势；通过条件强度函数能够完全地描述一个时空点过程模型，时空点过程分析法适合用于处理疫情时空数据。

关键词: 新型冠状病毒；流行病动力学模型；长短期记忆网络；时空点过程；疫情数据可视化；预测模型

Abstract

The outbreak of novel coronavirus pneumonia at the end of 2019 has caused great impact and harm to people's production, life and physical and mental health. Therefore, it is necessary to collect and analyze the spatial-temporal data of COVID-19 in China. This project has collected and sorted out the data of COVID-19 in China from January 2020 to March 2022. At the same time, it has carried out the visualization of data sets, and

conducted modeling and analysis of data sets from different perspectives. In this thesis, the traditional epidemic dynamics model is utilized, and SEIR series models are used. Then, considering the time series characteristics of data, the research is carried out from the perspective of short-term and long-term memory network. Finally, the experimental analysis is carried out from the point of view of spatiotemporal point process. The experimental results show that the development trend of COVID-19 is generally in line with the hypothesis of SEIR series models, and SEIR series models can be used to simulate and predict the development trend at the initial stage of the outbreak. When the training sample data is sufficient, the long-term and short-term memory network model can achieve better performance, but when the training sample is insufficient, the long-term and short-term memory network cannot effectively learn the trend of epidemic development. The conditional intensity function can effectively describe a spatiotemporal point process model, and the spatiotemporal point process analysis method is suitable for processing spatiotemporal data of epidemic.

Keywords: novel coronavirus, epidemic dynamics model,  
long-short-term memory network, spatial-temporal point process  
visualization of epidemic data, prediction model

目录

摘要 ..... I

ABSTRACT ..... II

第1章绪论..... 1

1.1 课题背景及研究的目的和意义..... 1

1.2 疫情趋势预测模型及其相关理论的发展概况..... 1

1.2.1 传染病动力学模型分类..... 1

1.2.2 经典仓室模型及相关研究工作的发展..... 3

1.2.3 LSTM模型及相关研究工作的发展..... 3

1.3 时空点过程模型的发展概况及其在流行病学中的应用..... 4

1.3.1 时空点过程模型分析法概述..... 4

1.3.2 点过程模型的分类与例述..... 5

1.3.3 时空点过程模型在流行病时空传播分析的应用..... 6

1.4 本文的主要研究内容..... 6

1.5 本文的组织结构..... 6

第2章新冠疫情时空数据的获取与可视化 ..... 8

2.1 引言..... 8

2.2 网络爬虫的一般概念与工作流程..... 8

2.2.1 爬虫技术的概念与工作流程..... 8

2.2.2 数据爬取预处理系统的模型建立..... 8

2.3 新冠疫情数据集的获取..... 10

2.3.1 新冠疫情数据集总体情况说明..... 10

2.3.2 省市级疫情实时数据的爬取..... 11

2.3.3 深圳市卫健委病例数据的爬取..... 13

2.3.4 中国各级行政区划信息的爬取..... 14

2.4 新冠疫情数据可视化系统..... 15

2.4.1 中国实时疫情信息的可视化分析..... 15

2.4.2 深圳市疫情信息的可视化分析..... 16

2.5 本章小结..... 18

第3章新冠疫情数据的预测分析方法 ..... 19

3.1 引言..... 19

3.2 SEIR 系列模型原理详述..... 19

3.3 基于深圳市数据的SEIR系列模型分析..... 22

3.4 基于长短期记忆网络的时序分析..... 24

3.4.1 RNN网络模型中存在的问题..... 25

3.4.2 LSTM网络的工作原理与改进点..... 25

3.4.3 使用LSTM网络分析疫情数据..... 26

3.5 模型间的比较分析..... 27

3.6 本章小结..... 28

第4章基于时空点过程的新冠疫情数据分析 ..... 29

4.1 引言..... 29

4.2 时空点过程的事件建模法..... 29

4.2.1 时空点过程的一般形式..... 29

4.2.2 几种点过程模型的探讨..... 30

4.3 基于点过程模型的新冠疫情数据分析..... 31

4.3.1 数据的预处理..... 31

4.3.2 点过程模型的部署与训练..... 32

4.3 本章小结.....33
结论 .....34
参考文献.....35
原创性声明.....37
致谢.....38

指 标
疑似剽窃文字表述

1. harm to people’s production, life and physical and mental health. Therefore

2. 第1章绪论
总字数：5868

相似文献列表

去除本人文献复制比：28.1%(1651)
文字复制比：28.1%(1651)
疑似剽窃观点：(0)

Table with 3 columns: Index, Title, and Similarity/Quotation Ratio. It lists 17 references related to epidemic models and public health, including titles like '传染病传播模型综述' and '新冠肺炎疫情趋势预测模型'.



## 原文内容

## 第1章绪论

## 1.1 课题背景及研究的目的和意义

2019年12月,新型冠状病毒肺炎疫情(corona virus disease 2019, COVID-19)在中国湖北省武汉市爆发。新型冠状病毒在爆发初期就表现出传播速度快、传染性强的特点,短时间内即造成大范围的感染。为控制疫情,武汉市于2020年1月23日实施封城管理,此后全国多个省市相继启动突发公共卫生事件I级响应机制。2020年1月31日,世界卫生组织宣布COVID-19疫情(以下简称“新冠疫情”或“疫情”)构成“国际关注的突发公共卫生事件”,后于当地时间2020年3月11日宣布新冠疫情成为全球大流行。一时之间,疫情的发展趋势成为全国乃至世界人民共同关注的问题,通过对疫情发展趋势的有效预测以及时空特征分析等可对人们提前制定应对措施提供帮助。从疫情爆发初期开始,世界范围内就有诸多学者对疫情情况进行了研究分析[1-4]。

一方面当前正是经济全球化的时代,货物、人口流动频繁,人类社会经济文化空前发展,另一方面这也加剧了各种传染病的产生、爆发和流行的可能。因此,在新冠疫情逐渐常态化的今天,展开对中国新冠疫情数据的收集工作,以及利用数学建模方法与机器学习模型对新冠疫情传播过程进行描述、分析、预报和控制具有重要的意义。同时,国家卫生健康委员会、地方政府会对每日的新冠疫情情况进行通报,部分网络平台会依据上述通报上线实时播报平台,这是本文研究的数据基础。

## 1.2 疫情趋势预测模型及其相关理论的发展概况

从系统科学的角度来看,传染病流行是在社会人群中发生的一个复杂扩散过程,针对这一扩散过程对传染病传播流行机理进行分析表达的理论方法即为传染病动力学模型[5]。另一方面,整理得到的疫情数据中包含时序数据,可以使用LSTM网络进行一定的分析。以模型为基础对传染病进行分析和预测,有助于理解传染病的流行机理,为人工干预措施的选择提供理论依据。

## 1.2.1 传染病动力学模型分类

对传染病扩散过程的建模已经有很长的历史,期间发展出了多种模型范式,这些模型各有所长。单一群体法基于传统的仓室模型,将整个人群视为一个封闭系统,利用微分方程组刻画人群状态的变化。随着社会流动性的增强,传染病的空间扩展呈现出新的模式,主要是范围更大、局部聚集等特性,使得单一群体法很难处理,而来源于空间生态学的复合群体(meta-population)方法在大空间尺度的传染病传播建模中得到了应用。近年来,随着复杂性交叉科学的兴起,建模趋于细化,将每一个个体作为研究对象,与社会网络和深度学习相结合,发展出基于网络的微观个体建模方法,为认识传染病的传播规律提供了新的途径。

因此,目前建立传染病传播模型方法主要有三种[5]:

(a) 单一人群仓室建模模型将整个人群视为一个整体,并把这个整体放到一个虚拟的封闭仓室中研究,在人群内部进行类型划分,流行过程表现在易感者,感染者等各类人员集计量的变化。

(b) 多子群耦合建模将人群划分为多个子群(subpopulation),各子群体之间因人员流动而耦合,形成复杂动态系统

(c) 微观个体网络建模建模出发点是人群中的个体,每一个个体都有各自的属性和行为规则,个体视为网络节点,个体与个体的接触视为网络边,形成网络上的传播动力学过程。

三种建模方法的基本思想如图1-1所示:

$PI[s(t), \dots], \dots, PI[s(t), \dots], \dots S(t), I(t), \dots$

人群1人群

人群3人群2

(a) 单一人群仓室建模

(b) 多子群耦合建模

$PI[s_1, s_2, \dots], \dots, PI[s_1, s_2, \dots], \dots$

(c) 微观个体网络建模

图1-1 传染病动力学模型结构的分类

图1-1中展示了三种传染病传播模型的核心思想。图(a)是单一人群仓室模型,将人群看作一个整体,直接划分为S(易感者)、I(感染者)等仓室,由于仓室人数随时间变化,因此分别记为 $S(t)$ 、 $I(t)$ 等。图(b)是多子群耦合模型,一个实际的庞大的群体往往存在空间局部聚集性,模型据此将人群划分为多个子群,在子群内部进行不同仓室的划分。图(c)表示微观个体网络模型,每一个个体包含不同的属性,不同属性的个体使用不同颜色表示,个体间的连线表示两个个体发生了接触联系,因而微观个体网络模型研究的对象更细致,考虑的因素也更加复杂。

## 1.2.2 经典仓室模型及相关研究工作的发展

仓室模型采用宏观视角建模,关注整个人群状态的变化,所有处于相同状态的人构成一个仓室,随着状态的变化,人员在仓室之间移动,一般采用(偏)微分方程描述。最经典的仓室模型(compartmental model)是SIR模型,由Kermack等于1927年提出[6],正如模型名称所示,该模型把人群划分为三个仓室易感者(Susceptible, S)、感染者(Infected, I)、康复者(Recovered, R)。

在SIR模型的基础上,仓室模型主要在以下几个方面进行了扩展:一是采用不同的仓室设置;二是考虑人口动力学,特别是将人口年龄结构看作重要因素;三是考虑更多的因素,如随机性、人口、空间的异质性等。

考虑其中针对仓室设置的扩展,有如下说明:

仓室对应着传染病流行过程中的可能状态,仓室的划分首先与传染病特性有关。经典SIR模型适合治愈后获得终生免疫的疾病。某些疾病治愈后不能获得免疫力,则可用SIS模型。若康复后能够获得一定的免疫力,但免疫力会逐渐消失,则可用SIRS模型。某些疾病有潜伏期(暴露, exposed),则可用SEIR模型。某些疾病新生儿可从母亲获得被动免疫,但一段时间后免疫力消失,则可采用MSEIR模型(M表示从母亲获得被动免疫的人群)[5]。与此类似,还有许多仓室设置方式,一般根据仓室和转移路线称为SI、SEI、SEIRS、MSEIRS等。

### 1.2.3 LSTM模型在疫情数据上的分析

循环神经网络(Recurrent Neural Network, RNN)采用循环链式结构处理问题输入,其中的循环重复节点称为闭合隐藏中间单元。RNN的闭合隐藏神经单元结构较为简单,在数据处理的过程中会出现累乘算式,容易引起梯度消失或梯度爆炸问题[7]。长短期记忆网络(Long Short Term Memory network, LSTM)是一种特殊的RNN,相对于RNN,其闭合隐藏中间单元更加复杂,可以很好地解决长时依赖问题[7][8]。具体的,LSTM主要通过引入门控机制来控制信息的累积速度,包括有选择地加入新的信息,并有选择地遗忘之前累积的信息[8],由此缓解长序列训练过程中的梯度消失和梯度爆炸问题。

新冠疫情数据中包含时序性的数据序列,可使用LSTM模型进行分析。赵永翼等人针对2020年2月22日—7月13日的新冠疫情新增数据等进行了序列分析预测[9],实验基于python的keras框架和tensorflow深度学习库进行,根据前3天预测第4天累计确诊人数。实验表明LSTM网络适合用于做疫情序列分析预测。

### 1.3 时空点过程模型的发展概况及其在流行病学中的应用

#### 1.3.1 时空点过程模型分析法概述

点过程的概念本身就包含着时空的因素,点是指定义在有界连续的时间域或空间域(例如一个欧几里得平面或三维的欧几里得空间等)中的事件(events),而过程(process)则蕴含着时间推进的思想,因而点过程可以更为也称为时空点过程。

时空点过程建模分析法能对众多真实场景中产生的数据建模。例如在1978年Ogata教授等人基于时空点过程研究了地震爆发的位置和震级数据,预测余震发生的位置和时间,论文采用的预测方法是通过仿真来生成未来的余震事件,并提出了一种基于Shedler-Lewis 细化算法(Shedler-Lewis thinning algorithm)的改进细化算法[10]。Stoyan教授将时空点过程建模方法应用于林业,对森林科学数据展开分析,指出点过程的理论统计方法和随机模型的快速发展在某种程度上完全是受林业应用的启发,特别是现代已开发的单树模型[11]。

地震数据、森林数据以及疫情数据等都是多维异步事件数据,它们互相影响并在连续时间域上呈现出复杂的动态规律。基于时空点过程的序列分析法,一个重要的应用就是对未来的事件进行预测。图1-2给出了多维异步事件序列的一个示意图,三个用户的行为被表征为不同的事件序列,不同类型的事件由不同形状表征,虚线箭头代表事件间前后关联。

图1-2 类别标记的连续时间域异步事件序列及其相互影响

图1-2中,箭头(激励关系)都是指向时间轴的正方向,表明历史事件会影响未来事件的时空过程特征。同时,一部分箭头从一个用户的事件指向另一个用户事件,另一部分则从一个用户的事件指向同一个用户的另一个事件,这体现着时空点过程研究对象(事件)的多维异步性。

#### 1.3.2 点过程模型的分类与例述

以机器学习的视角来看,时序点过程的发展可以分为两个方向:传统统计点过程和深度点过程。统计点过程有多种点过程函数形式,其建模的成功往往依赖于正确的模型选择。由于其函数形式和参数表示一般具有一定的物理意义,统计点过程具有更强的解释性,对于样本数量的依赖也较小。相比之下,深度点过程(亦被称为神经点过程)利用神经网络的强大容量,通过大规模数据,试图学习拟合能力更强的模型,并减少对先验知识的依赖,而一定程度上牺牲了模型解释性。

典型的点过程模型包括[12]:

(a) 泊松过程(Poisson process):泊松过程又分为齐次泊松过程和非齐次泊松过程。在齐次泊松过程中,强度函数值 $\lambda$ 是一个恒定的值,独立于历史事件。与齐次泊松过程相比,非齐次泊松过程则认为 $\lambda(t)$ 是随时间变化的函数。而Cox教授提出了经典的重随机泊松过程[13],将强度函数看成是一个受到各种外部因素影响的随机变量。

(b) 自校正过程(self-correcting process):自校正过程假设强度函数随着时间稳步提升,当一个事件发生后,强度函数会以指数的形式回落[12]。

(c) 自/互激励过程(self/mutual-exciting process):自激励过程,亦称为霍克斯过程(Hawkes processes)[14],其内部机制表示发生的历史事件对于未来事件的发生有激励作用,并且历史事件的影响以累加的形式进行叠加。

#### 1.3.3 时空点过程模型在流行病时空传播分析的应用

在2012年Meyer教授等发表的基于烈性脑膜炎球菌病流行过程的条件强度时空模型分析一文中[15],Meyer教授及其团队提出了基于条件强度函数(conditional intensity function)的时空点过程模型,它提供了一个参数框架,用于在疾病监测中典型的时空点过程中进行前瞻性变化点分析:在随机过程控制的框架内,例如,可以使用似然比检测器来监控时间需要包含流行病成分来描述观察到的数据的点。这在想法上与同质时空泊松过程设置相对应。

Schoenberg教授使用2014西非埃博拉疫情的数据来评估简单的霍克斯点过程模型在预测埃博拉病毒在几内亚、塞拉利昂和利里亚的传播预测的效果[16]。为了比较,使用SEIR模型在相同数据集上使用相同的指标进行评估。为了测试每个模型的预测能力,论文通过使用前75%的数据进行估计和随后的25%的数据进行评估来模拟在实际爆发期间进行近乎实时的预测的能力。

RTQ Chen教授在研究神经点过程的相关工作中[17],使用美国新泽西州自2020年03月15日至2020年08月01日共139日的数据库,并从时空点过程分析的视角对多个模型泊松过程、自校正过程、霍克斯过程、神经霍克斯过程等等进行了研究分析。

### 1.4 本文的主要研究内容

本文对几类流行病动力学模型包括SEIR模型以及改进的SEIR模型进行了模拟和仿真实验,结合深圳市早期疫情爆发时的历史疫情数据集对SEIR模型与改进SEIR模型进行了参数拟合。同时,结合最新研究的长短期记忆网络在与SEIR系列模型相同的数据集上进行了训练和预测,结合均方根误差(RMSE)和平均绝对误差(MAE)等误差标准对SEIR系列模型以及LSTM模型的预测结果进行了评估。结果表明,在训练样本容量足够时,LSTM模型的预测结果最好,预测值与真实值相当吻合,改进的SEIR模型预测结果其次,SEIR模型预测结果误差略大。此外,本文还从点过程分析法的视角,对疫情的扩散过程进行了研究,并采用RTQ Chen教授在研究时空点过程中使用的方法[17],对中国各省市历史新冠确诊数据集进行了研究和分析。此外,针对前期工作中收集整理得到的数据集,本文从不同的角度对其进行了描绘和可视化。

### 1.5 本文的组织结构

本文一共分为四章对研究内容进行展开叙述。

第一章为绪论:首先对本文的研究对象中国新冠疫情时空数据、研究目的和意义进行了简要概括。其次对本文将要使用到的模型及相关研究工作进行了叙述和说明,包括基于(偏)微分方程组的仓室系列模型及相关理论的研究概况。此外,近些年兴起的机器学习相关理论实践也为疫情数据的研究提供了新的视角,例如基于LSTM网络进行分析以及基于时空点过程研究其时



空特征。

第二章为新冠疫情时空数据的获取与可视化，详细说明了中国新冠疫情时空数据、深圳市患者个案数据以及中国各地级市人口、面积、经纬度等相关数据的获取方法，展示爬虫的工作流程和部分数据样例。同时介绍了基于已有的数据集进行的多方面可视化工作，完成了以Python与Flask+MySQL为基本框架的展示系统，系统使用到了jquery、ajax、echarts等前端技术。

第三章为新冠疫情数据的预测分析方法，详细阐释了SEIR系列模型和LSTM模型的原理，并对其中部分模型进行了仿真，同时以深圳市早期疫情数据集为基础，对SEIR模型、mSEIR模型和LSTM模型进行了训练，将各自的预测结果进行误差分析和比较。

第四章为从时空点过程的角度分析疫情数据，针对RTQ Chen教授等人研究的时空点过程方法对中国历史疫情数据进行了分析。主要完成了泊松过程、自校正过程和霍克斯过程的模型训练，阐述了极大似然估计的方法原理，利用对数似然值对结果进行了探讨分析。

指 标		
疑似剽窃文字表述		
<div>1. 新型冠状病毒在爆发初期就表现出传播速度快、传染性强的特点，短时间内即造成大范围的感染。为控制疫情，武汉市于2020年1月23日实施封城管理，此后全国多个省市相继启动突发公共卫生事件Ⅰ级响应机制。</div> <div>2. 一时之间，疫情的发展趋势成为全国乃至世界人民共同关注的问题，通过对疫情发展趋势的有效预测以及时空特征分析等可对人们提前制定应对措施提供帮助。</div> <div>3. 模型的分类与例述 以机器学习的视角来看，时序点过程的发展可以分为两个方向：传统统计点过程和深度点过程。统计点过程有多种点过程函数形式，其建模的成功往往依赖于正确的模型选择。</div> <div>4. 在齐次泊松过程中，强度函数值 <math>\lambda</math> 是一个恒定的值，独立于历史事件。与齐次泊松过程相比，</div>		
3. 第2章新冠疫情时空数据的获取与可视化		总字数：7336
相似文献列表		
去除本人文献复制比：7.7%(562)      文字复制比：7.7%(562)      疑似剽窃观点：(0)		
1	高性能网络爬虫:研究综述 周德懋;李舟军; - 《计算机科学》- 2009-08-15	5.1% (374) 是否引证: 否
2	网站内容安全管理平台的设计与实现 斯鹏(导师: 韩臻) - 《北京交通大学硕士论文》- 2010-06-01	5.0% (364) 是否引证: 否
3	高性能网络爬虫研究综述 - 道客巴巴 - 《互联网文档资源 ( <a href="http://www.doc88.com">http://www.doc88.com</a> )》- 2019	4.1% (300) 是否引证: 否
4	高性能网络爬虫 - 《互联网文档资源 ( <a href="http://wenku.baidu.c">http://wenku.baidu.c</a> )》- 2012	3.3% (242) 是否引证: 否
5	高性能网络爬虫研究综述-百度文库 - 《互联网文档资源 ( <a href="http://wenku.baidu.c">http://wenku.baidu.c</a> )》- 2012	3.3% (242) 是否引证: 否
6	网络爬虫技术研究 于成龙;于洪波; - 《东莞理工学院学报》- 2011-06-15	3.1% (228) 是否引证: 否
7	1604854502_单昌旺_基于Java的停车场管理系统 单昌旺 - 《大学生论文联合比对库》- 2020-04-22	2.3% (169) 是否引证: 否
8	基于可视化技术的数据分析平台的设计与实现 屈成莉 - 《大学生论文联合比对库》- 2020-04-28	2.3% (169) 是否引证: 否
9	基于python的政务网站分析与监督系统 韩飞 - 《大学生论文联合比对库》- 2020-06-07	2.3% (169) 是否引证: 否
10	数据可视化技术研究与初探 肖媛 - 《大学生论文联合比对库》- 2020-04-01	2.3% (169) 是否引证: 否
11	WEB大屏可视化设计与实现 刘春宇 - 《大学生论文联合比对库》- 2020-04-14	2.3% (169) 是否引证: 否
12	基于地图检索的景区资源管理系统设计与实现 王俊玲 - 《大学生论文联合比对库》- 2020-05-08	2.3% (169) 是否引证: 否

13	200521003006   11603990829_邢津萍_行业大数据集的动态构建与可视化应用（医疗器械领域） 邢津萍 - 《大学生论文联合比对库》 - 2020-05-22	2.1% (157) 是否引证：否
14	11603990829_邢津萍_行业大数据集的动态构建与可视化应用（医疗器械领域）(1) 邢津萍 - 《大学生论文联合比对库》 - 2020-05-27	2.1% (157) 是否引证：否
15	唐莉201704136103-肺炎疫情小程序设计 唐莉 - 《高职高专院校联合比对库》 - 2020-03-27	2.1% (155) 是否引证：否
16	201704136103唐莉2毕业设计作品 唐莉 - 《高职高专院校联合比对库》 - 2020-05-11	2.1% (155) 是否引证：否
17	基于爬虫的网站文本信息分类与评估系统设计与实现-金明渊 无 - 《高职高专院校联合比对库》 - 2020-06-07	2.1% (155) 是否引证：否
18	计算机科学与技术学院_计算机1601班_160800316_金明渊 班 - 《大学生论文联合比对库》 - 2020-06-01	2.1% (155) 是否引证：否
19	基于Smart Card数据的乘客公共交通出行分析与可视化 李金灿 - 《大学生论文联合比对库》 - 2020-04-18	2.1% (155) 是否引证：否
20	客户管理信息系统设计与实现 王顺 - 《大学生论文联合比对库》 - 2020-05-06	2.1% (155) 是否引证：否
21	多模态情感分析 徐宇超 - 《大学生论文联合比对库》 - 2020-05-16	2.1% (155) 是否引证：否
22	何海洋_160000000160_小区垃圾服务平台 何海洋 - 《大学生论文联合比对库》 - 2020-05-21	1.8% (132) 是否引证：否
23	何海洋_160000000160_小区垃圾服务平台设计与实现 何海洋 - 《大学生论文联合比对库》 - 2020-06-02	1.8% (132) 是否引证：否
24	基于Python的农业气象数据可视化设计 龚雪敏 - 《高职高专院校联合比对库》 - 2020-04-09	1.3% (98) 是否引证：否
25	基于Echarts和GeoServer的学术地图发布技术研究 魏张鉴 - 《大学生论文联合比对库》 - 2018-06-04	1.2% (87) 是否引证：否

原文内容

第2章新冠疫情时空数据的获取与可视化

2.1 引言

真实可信的数据是后续展开建模分析与可视化工作的基础。目标数据往往散布在离散的网页上，而网页是一种半结构化的数据结构，可以包含文字、图片、音视频等信息，因而其内容往往是存在大量冗余的，需要进行过滤和预处理。一般而言过滤和预处理过程是多层次的，其最终目的是得到程序可处理的数据输入。这一系列工作即定义了本文使用的数据爬取预处理系统，另一方面，为了更好的展示实时爬取的数据，本文还设计实现了疫情实时数据可视化系统。

2.2 网络爬虫的一般概念与工作流程

2.2.1 爬虫技术的概念与工作流程

作为搜索引擎的基础构件之一，网络爬虫 (Crawler) 直接面向互联网，它是搜索引擎的数据来源，决定着整个系统的内容是否丰富、信息能否得到及时更新。它的性能表现直接影响整个搜索引擎的效果。一个高性能的Crawler需要从以下几个方面来考虑[18]：

(a) 可伸缩性。能胜任海量数据的抓取，并可通过增加硬件资源使性能得到线性提高。

(b) 分布式。集中式的Crawler架构已经不能满足目前互联网的规模，因此支持分布式的爬行，处理和协调好各结点之间的交互，也是一个重要议题。

(c) 限制爬行。Crawler不能在短时间内大数据量地集中访问同一个主机下的网页，否则会影响普通用户对其的访问，进而可能被对方限制访问。

(d) 可定制性。可根据不同的爬行任务和特定的主题定制相应的功能模块，使功能插件化，打造个性化Crawler。

在本文中，给出一个可能的定义：网络爬虫(Web Crawler)是从指定的初始url队列开始，对于每一个队列中的url，通过网络资源请求获取到网络页面，并结合网页解析技术提取有效信息同时把过滤得到的新url添加到初始url队列尾并继续爬取页面的自动化过程，这一过程在达到用户指定的结束条件时终止。

2.2.2 数据爬取预处理系统的模型建立

一个良好运作的系统一般要满足功能模块化层次化、高内聚低耦合的特点，本文设计的数据爬取预处理系统将数据提取工作细化，并分别交给爬取模块、数据预处理模块和数据存取模块。这三个模块的说明如下：

(1) 爬取模块

爬取模块负责爬取原始网页，将网页的目标内容存入文档，是数据爬取预处理系统的关键模块。需要获取的目标数据属于文本数据，因而文档格式采用逗号分隔文件(csv文件)，优点是文档属于纯文本文档，占用空间小，同时又可以借助python的



pandas库进行存取，操作类似处理表格时按行按列操作。

爬取模块地核心网络爬虫程序，程序必须按照2.2.1中对爬虫的定义严格执行，同时按照需求尽可能满足高性能的Crawler的特性：

首先对于可扩展性与分布式的要求。由于本系统仅针对几个特定的平台或发布页进行信息获取，数据量并不大，初始url队列较小。因此，可以暂时不考虑这两个层次的要求。

其次，爬虫必须有限制地爬行，否则可能面临拒绝访问的风险。同时与通常的程序不同，爬虫程序是一种网络程序，需要考虑网络的波动性与服务器的负载能力，例如，需要编写异常处理模块以应对请求页面超时等问题，需要在连续获取页面时插入一段随机等待时间，需要为网络请求函数封装请求头部，并随机更换等等。

最后针对可定制性的要求，在使用Python语言编写爬虫模块的背景下，可以将其封装为一个类，例如仅针对类似发布页信息爬取的爬虫类，它可以自动的翻页和获取文章列表，同时获取文章具体内容。

(2) 数据预处理模块

数据预处理模块负责按照模型训练的数据输入要求对爬取得到的csv文件中的原始数据进行过滤和转换，是数据爬取预处理系统的重要模块。这一步往往是多层次的，需要一步步的过滤以及转化，其中过滤过程用到的主要方法是直接字符串匹配法、字符串正则匹配法等；转化过程一方面需要将一些非结构化的信息转为结构化的信息，例如对于提取出的地点信息，应当结合有关地图API得到经纬度的位置信息，另一方面则需要进行一定的计算处理，例如对于日期信息，应该确定起止日期，并将起止日期以及之间的所有日期转为距起始日期的时间间隔等。

(3) 数据存取模块

数据存取模块负责将处理好的数据存入数据库进行固化，本文采用MySQL数据库，借助MySQL脚本进行数据存取，因此这一部分的工作主要交由MySQL数据服务器处理。

如图2-1展示了数据爬取预处理系统的主要工作流程：

图2-1 数据爬取预处理系统工作流程

图2-1中，整个系统的工作流程被划分为三个部分：

其一是爬取模块，模块维持一个URL队列，队列初始化为初始URL集合，此后爬虫程序每次从URL队列中取一个URL并访问得到页面，一方面将页面数据交给数据预处理模块，另一方面依据主题过滤得到页面中的目标URL并放入队列。

其二是数据预处理模块，模块将接收到的页面数据依据主题进行过滤和转化得到适应模型的输入数据。

最后是数据存取模块，模块将适应模型的输入数据存入MySQL数据库以满足后续工作的数据存取需求。

2.3 新冠疫情数据集的获取

2.3.1 新冠疫情数据集总体情况说明

为了满足后续模型训练的需要，本文利用设计的数据爬取预处理系统爬取了必要的基本数据集。首先是中国的疫情情况数据，包括每天各省市的新增确诊病例数等，时间范围确定为2020年2月疫情爆发初期一直到2022年3月，来源首选国家以及地方卫生健康局官网的疫情通报。其次，在时空点过程模型分析中，本文的主要研究对象是市区，需要得到中国各级行政区划的面积等信息。最后，本文还讨论了深圳市新冠肺炎确诊患者更详细的情况，包括性别、年龄等。因此至少需要收集三个基本数据集，表2-1展示了这三个数据集的基本情况：

表2-1 新冠疫情数据集总体情况说明

数据集名称	主键	数据集大小	主要字段
中国各省市疫情数据集	日期、城市行政代码	381, 156	新增确诊、新增治愈、累计确诊、累计治愈等
中国行政区划数据集	城市（乡镇）行政代码	3, 071	人口数、占地面积、政府所在地名称、政府所在地经度、政府所在地纬度等
深圳市病例数据集	病例编号	1, 084	确诊时间、性别、年龄、常居地名称、常居地经度、常居地纬度

数据集名称主键数据集大小主要字段

中国各省市疫情数据集日期、城市行政代码 381, 156 新增确诊、新增治愈、累计确诊、累计治愈等

中国行政区划数据集城市（乡镇）行政代码 3, 071 人口数、占地面积、政府所在地名称、政府所在地经度、政府所在地纬度等

深圳市病例数据集病例编号 1, 084 确诊时间、性别、年龄、常居地名称、常居地经度、常居地纬度

表2-1中，中国各省市疫情数据集数据量最大，包含了自2020年2月疫情爆发初期到2022年5月精确到二级或三级行政区划（例如：广东省-深圳市属于二级行政区划，北京市朝阳区属于三级行政区划）的疫情数据共约38万条。中国行政区划数据集则包含中国各级行政区划的信息情况，包括后续模型训练中需要考虑到的区域面积信息。深圳市病例数据集则是一份较小的数据集，数据条目以患病个体个体为描述对象，主要包含确诊时间和地点信息等。

2.3.2 省市级疫情实时数据的爬取

本部分的数据采集主要依赖于疫情实时播报平台，此时网页更像是一个静态的容器，装入的数据是实时更新的，无法按照爬取网页然后利用HTML解析器来解析静态网页的方式获取数据。因此，需要考虑一种动态的方式，从请求网页资源开始，监视服务器端回复的所有格式的文件和数据包，通常数据存放在json格式的文件中。通过浏览器的开发者工具，抓取目标json数据包的api地址，获取和分析更加格式化的json数据。

经过数据对比分析，本课题选择网易的实时发布平台作为爬取对象，将其中有关中国各省市级的病例json数据爬取下来，目前累计收集清洗数据量约为38万条。如图2-2所示，为香港2022年3月13日的疫情情况。可以看到，3月13日香港单日新增确诊8163例，需要得到严格的管控，另外毗邻香港的深圳市单日新增确诊60例，形势依然严峻，从更大的时空范围看，疫情的爆发的确具有时间和空间的局部性，而基于疫情数据的时空点过程模型的研究，正是把这一种局部性用更形式化的语言加以表征。

图2-2 数据爬取预处理系统

表2-2 2022年3月13日的疫情情况（部分）

经过预处理后，将数据存储在MySQL数据库中，主要字段包括LastUpdate（更新日期）、CityShortName（城市名称）、incrConfirmed（新增病例数）、incrHealed（新增治愈数）、incrDeath（新增死亡数）、totalConfirmed（累计确诊数）、totalHealed（累计治愈数）、totalDeath（累计死亡数）等。其中2022年3月13日的部分数据样例如表2-2所示：

更新日期	城市名称	新增病例数	新增治愈数	新增死亡数
2022-03-13	长春	831	0	0
2022-03-13	青岛	179	0	0
2022-03-13	深圳	60	0	0
2022-03-13	宝鸡	33	0	0

更新日期城市名称新增病例数新增治愈数新增死亡数

2022-03-13 长春 831 0 0  
2022-03-13 青岛 179 0 0  
2022-03-13 深圳 60 0 0  
2022-03-13 宝鸡 33 0 0

整个省市级新冠疫情数据集包含了自2020年02月25日至今（2022年03月31日），同时爬虫部署为每天下午16:00执行一次，因此数据集仍在扩充。本文针对2020年02月25日至2022年03月31日共765日约32万条数据的数据集进行研究和分析。

2.3.3 深圳市卫健委病例数据的爬取

为了进一步对深圳市新冠肺炎确诊患者数据进行分析，包括本文后续将要开展的基于患者的时空点过程模型分析的研究，本文收集整理了深圳市本土确诊患者数据集。这一数据集粒度更细，精确到每一位患者，需要得到每位患者的详细信息，因此数据预处理过程更加复杂。

首先从深圳市人民政府疫情通报发布页爬取疫情通报，主要包括文章发布时间、文章标题以及文章内容，其中文章内容以字符串列表的形式存放，按照HTML中正文内容安排依次存放其中每一个文本标签（p标签）中的文字，这样处理有利于后续病例信息的提取。

接下来是将每一个本土确诊病例的详细信息从通报中分离出来，主要包含确诊日期、病患性别、病患年龄以及常居地点。这一步首先依据是否含有类似“本土确诊”等字样过滤出含有本土确诊病例详细信息的通报，然后再过滤得到病患的详细信息。最后一步是要把非结构化的常居地转为可处理的结构化数据，也即经纬度，这里利用了百度地图API，通过查询得到每个地点的经纬度。如表2-3所示为深圳市新冠本土确诊患者的详细情况：

确诊日期	性别	年龄	经度	纬度
2022-02-27	1	49	114.03911	22.56227
2022-02-27	0	9	114.06407	22.53181
2022-02-27	0	37	114.06407	22.53181
2022-02-27	1	44	113.93721	22.50655

确诊日期性别年龄经度纬度

2022-02-27 1 49 114.03911 22.56227  
2022-02-27 0 9 114.06407 22.53181  
2022-02-27 0 37 114.06407 22.53181  
2022-02-27 1 44 113.93721 22.50655

表2-3 深圳市新冠疫情患者的具体信息（部分）

数据集合计约有1000条记录，每一条记录都代表了一个确诊患者，记录中包含的字段包括确诊日期、病例性别（0表示女性、1表示男性）、病例年龄以及常居地经纬度。确诊患者的性别比例为女性：男性 = 0.9056：1，可以看到患者中男女比例大致相当，但男性占比稍高。

2.3.4 中国各级行政区划信息的爬取

表2-4 中国行政区划信息一览（部分，仅北京市）

在时空点过程建模分析的研究中，需要考虑到各个省市的面积这一影响因素，本文收集整理了中国各省、各地级市最新的行政区划信息，例如面积、人口、行政区划代码等等，这些信息可以在中华人民共和国民政部全国行政区划信息查询平台中找到。为了获取到所有地区的信息，本文首先查询并存储了包含34个省级行政区划的网页（也就是无法自动翻页），随后将离线页面经过页面解析和过滤预处理得到最终的数据结果。其中北京市的数据如表2-4所示：

行政区划代码	地区	面积 单位：km2	人口 单位：万人	政府部门所在地	邮政编码
110000	北京市	16, 418	1, 392	通州区	（空）
110101	东城区	42	99	景山街道	100010
110102	西城区	51	150	金融街街道	100032
110105	朝阳区	465	214	朝外街道	100020
110106	丰台区	306	116	丰台街道	100071
110107	石景山区	84	39	鲁谷街道	100043
110108	海淀区	431	239	海淀街道	100089
110109	门头沟区	1, 448	25	大峪街道	102300
110111	房山区	1, 995	84	拱辰街道	102400
110112	通州区	906	80	北苑街道	101100
110113	顺义区	1, 020	65	双丰街道	101300
110114	昌平区	1, 342	65	城北街道	102200
110115	大兴区	1, 036	73	兴丰街道	102600
110116	怀柔区	2, 123	29	龙山街道	101400
110117	平谷区	948	41	滨河街道	101200
110118	密云区	2, 226	44	鼓楼街道	101500

110119	延庆区	1, 995	29	儒林街道	102100
--------	-----	--------	----	------	--------

行政区划代码地区面积

单位: km2 人口

单位: 万人政府部门所在地邮政编码

110000 北京市 16, 418 1, 392 通州区 (空)  
110101 东城区 42 99 景山街道 100010  
110102 西城区 51 150 金融街街道 100032  
110105 朝阳区 465 214 朝外街道 100020  
110106 丰台区 306 116 丰台街道 100071  
110107 石景山区 84 39 鲁谷街道 100043  
110108 海淀区 431 239 海淀街道 100089  
110109 门头沟区 1, 448 25 大峪街道 102300  
110111 房山区 1, 995 84 拱辰街道 102400  
110112 通州区 906 80 北苑街道 101100  
110113 顺义区 1, 020 65 双丰街道 101300  
110114 昌平区 1, 342 65 城北街道 102200  
110115 大兴区 1, 036 73 兴丰街道 102600  
110116 怀柔区 2, 123 29 龙山街道 101400  
110117 平谷区 948 41 滨河街道 101200  
110118 密云区 2, 226 44 鼓楼街道 101500  
110119 延庆区 1, 995 29 儒林街道 102100

表2-4给出了北京市及其下级行政单位的基本信息, 包括行政区划代码、对应行政区名称以及面积、人口等信息, 可以看到北京市一共有15个下级行政区划, 占地面积16, 418km<sup>2</sup>, 人口共约1, 392万人, 人口密度平均约为848人/km<sup>2</sup>。

2.4 新冠疫情数据可视化系统

2.4.1 中国实时疫情信息的可视化分析

为了更直观的对当日的疫情情况作一个直观的了解, 本文实现了一个简单的中国疫情信息实时可视化平台。如图2-4所示, 疫情数据可视化平台主要由中国疫情实时地图、中国疫情的累计趋势和新增趋势折线统计图和新增确诊较多的城市横向条形图构成。

界面左边是两张折线图, 左上方是近一周全国累计趋势, 其中的三条折线从上至下依次是累计确诊、累计治愈、累计死亡的趋势, 左下方则为近一周全国新增趋势, 其中的三条折线从上至下依次是新增确诊、新增治愈、新增死亡的趋势, 可以看出, 无论是累计趋势还是新增趋势, 确诊所在折线均远高于其它折线, 对比来看, 累计确诊数字和新增确诊数字均有较大增幅, 形式不容乐观。

区域中部上方精确展示了截至05月16日全国累计确诊、累计治愈、新增确诊、新增治愈的数字, 中部中下方用一张中国地图展示中国各个省级行政区划实时的新增病例情况, 新增病例数越多颜色越深。右边则为横向的直方图, 从上到下分别表示新增确诊数、新增治愈数最多的十个省市, 按照数量递减的顺序进行排列。

图2-4 中国疫情实时播报平台概览图 (2022-05-16)

这些图表的展示主要运用到了ECharts技术。ECharts, 一个使用 JavaScript 实现的开源可视化库, 可以流畅的运行在 PC 和移动设备上, 兼容当前绝大部分浏览器 (IE8/9/10/11, Chrome, Firefox, Safari等), 底层依赖矢量图形库 ZRender, 提供直观, 交互丰富, 可高度个性化定制的数据可视化图表。

本项目首先引入ECharts库, 以疫情地图地实现为例, 需要将库中的地图组件插入到前端的HTML文档中, 接着从后台查询得到当日的疫情数据, 需要每个省的新增确诊病例数, 并划分严重等级, 在前台用由浅到深的颜色进行渲染和绘制得到最终的疫情实时地图如图2-5所示:

图2-5 中国疫情新增病例实时地图 (2022-05-16)

图2-5展示了2022年5月16日全国各省地新冠疫情新增确诊情况, 大部分地区处于零新增状态, 出现新增病例的省份大多数位于沿海位置, 且新增病例数大部分为个位数, 其中较严重的是四川省和北京市等地区, 疫情的动态清零工作仍需继续。

2.4.2 深圳市疫情信息的可视化分析

针对深圳市新冠病患详细信息数据集, 本文也进行了一定的可视化分析工作。如图2-3展示了深圳市新冠肺炎患者的年龄分布情况, 其中年龄划分以国际常用标准将人的年龄划分为婴幼儿 (0~2岁)、儿童 (3~10岁)、少年 (11~17岁)、青壮年 (18~35岁)、中年 (36~60岁) 和老年 (大于60岁) 六个阶段。

图2-3 深圳市新冠患者年龄分布 (截至2022年03月31日)

图2-3表明, 在各年龄段中, 患病占比最多的是青壮年和中年两个年龄段, 尽管年龄划分不均衡, 该图仍能说明新冠疫情患病与个体年龄存在一定的联系。一方面观察到青壮年时期为17年, 从婴幼儿一直到少年阶段结束时期也为17年, 而青壮年占比要高于婴幼儿一直到少年阶段占比的和, 同时中年时期有24年, 但其占比只是略微高于青壮年的占比, 因此总体来看, 青壮年更易患病。这一年龄段的人群一般处于就学或就业的环境, 流通性和聚集性较强, 是导致这一现象出现的可能因素之一。

如图2-4展示了深圳市新冠疫情 (2020-02至2022-03) 新冠疫情趋势, 其中图(a)为深圳市疫情累计趋势, 图(b)为深圳市疫情新增趋势。

(b) 深圳市新冠疫情新增趋势 (a) 深圳市新冠疫情累计趋势

图2-4 深圳市新冠疫情 (2020-02至2022-02) 新冠疫情趋势

方程组 (3)

可以看到深圳市有三轮较大的疫情爆发, 分别是疫情扩散初期也即2020年1月到4月、2021年7月份以及2022年1月底这三次

2.5 本章小结

本章主要说明了本文使用到的数据来源以及获取方法，介绍了获取网络数据的基本流程，设计实现了数据爬取预处理系统，其中最关键的模块是数据爬取模块，其中的主要程序被称为爬虫。一个高性能爬虫一般具备易扩展、有限制、分布式和可定制四个特点，对于本文数据量相对较少的场景，爬虫只需具备有限制和可定制的特点。同时数据预处理也是一个重要的环节，对于爬取的原始数据，需要依据实验需求过滤筛选、加工转换以得到合适的的数据输入。

另一方面，要对爬取到的每日省市数据进行直观地分析，一种好的方法就是将数据进行可视化，并以较丰富的形式呈现，例如折线图、条形图、地图等。

指 标	
疑似剽窃文字表述	
1. 能胜任海量数据的抓取，并可通过增加硬件资源使性能得到线性提高。 (b) 分布式。集中式的Crawler架构已经不能满足目前互联网的规模，因此支持分布式的爬行，处理和协调好各结点之间的交互，也是一个重要议题。 (c) 限制爬行。Crawler不能在短时间内大数据量地集中访问同一个主机下的网页，否则会影响普通用户对其的访问，进而可能被对方限制访问。 (d) 可定制性。可根据不同的爬行任务和特定的主题定制相应的功能模块，使功能插件化，打造个性化Crawler。	
4. 第3章新冠疫情数据的预测分析方法	
总字数：6081	
相似文献列表	
去除本人文献复制比：20.5%(1244)      文字复制比：20.5%(1244)      疑似剽窃观点：(0)	
1	简单理解LSTM神经网络 - 博客频道 - CSDN.NET - 《网络（ <a href="http://blog.csdn.net">http://blog.csdn.net</a> ）》 - 2016 7.6% (461) 是否引证：否
2	通信1601 201608030127 方娜妮 - 《大学生论文联合比对库》 - 2020-05-26 7.3% (441) 是否引证：否
3	基于低分辨率采样数据的非侵入式电力负荷分解方法研究 罗霄 - 《大学生论文联合比对库》 - 2019-05-21 6.0% (364) 是否引证：否
4	基于注意力的机器翻译 刘锋 - 《大学生论文联合比对库》 - 2018-05-25 5.9% (357) 是否引证：否
5	pm39_王锦坤_基于开源社区的bug修复者推荐 王锦坤 - 《大学生论文联合比对库》 - 2018-05-03 5.7% (349) 是否引证：否
6	基于TensorFlow的神经机器翻译 刘锋 - 《大学生论文联合比对库》 - 2018-05-29 5.7% (345) 是否引证：否
7	142720_基于低分辨率采样数据的非侵入式负荷分解研究 吴娣 - 《大学生论文联合比对库》 - 2019-05-21 5.6% (341) 是否引证：否
8	基于脑电信号的抬腿跨障意图在线预测方法研究 姚程远 - 《大学生论文联合比对库》 - 2019-05-26 5.5% (336) 是否引证：否
9	基于脑电信号的抬腿跨障意图在线预测方法研究 姚程远 - 《大学生论文联合比对库》 - 2019-05-31 5.5% (336) 是否引证：否
10	滚动轴承全寿命周期数据分析及趋势预测方法研究 黄逸飞 - 《大学生论文联合比对库》 - 2020-05-22 5.3% (320) 是否引证：否
11	2-王锦坤 王锦坤 - 《大学生论文联合比对库》 - 2018-05-03 5.1% (312) 是否引证：否
12	基于深度学习的船舶智能经济航速优化模型研究 展月 - 《大学生论文联合比对库》 - 2020-05-08 4.7% (287) 是否引证：否
13	信息学院-10153188-邹飞 信息学院 - 《大学生论文联合比对库》 - 2019-05-24 4.7% (285) 是否引证：否
14	信息学院-10153188-邹飞 信息学院 - 《大学生论文联合比对库》 - 2019-05-30 4.7% (285) 是否引证：否



15	172258_1621840567872 田镕昊 - 《大学生论文联合比对库》 - 2021-05-24	4.7% (284) 是否引证: 否
16	基于双语词向量的跨语言情感分析 张允哲 - 《大学生论文联合比对库》 - 2021-05-23	4.7% (284) 是否引证: 否
17	1501532822_陈冠恒_基于数据分析的济南泉水水位预测 祝静 - 《高职高专院校联合比对库》 - 2019-05-11	4.6% (279) 是否引证: 否
18	基于卷积神经网络的机械设备状态评估与故障诊断技术 肖扬 - 《大学生论文联合比对库》 - 2020-05-22	4.5% (275) 是否引证: 否
19	基于混合神经网络对金融序列预测的研究 陈基彤(导师: 陈浩;CHEN Zhan) - 《湖南大学硕士论文》 - 2019-04-15	4.1% (247) 是否引证: 否
20	信管电商_01621535_曹帅 信管电商 - 《大学生论文联合比对库》 - 2020-04-24	4.0% (242) 是否引证: 否
21	新冠肺炎疫情趋势预测模型 甘雨;吴雨;王建勇; - 《智能系统学报》 - 2021-04-01 09:29	4.0% (242) 是否引证: 否
22	0161121535_曹帅_计算机学院 曹帅 - 《大学生论文联合比对库》 - 2020-05-18	3.3% (198) 是否引证: 否
23	基于LSTM的中国古诗自动生成方法研究 李俊潮 - 《大学生论文联合比对库》 - 2020-01-01	3.2% (197) 是否引证: 否
24	不同深度学习算法在中文语音识别中的准确性比较 毛汝勇 - 《大学生论文联合比对库》 - 2018-06-01	2.6% (160) 是否引证: 否
25	基于EEMD-SSA组合模型的短期电力负荷预测研究 齐少拴(导师: 曹广毕;朱雷) - 《东北石油大学硕士论文》 - 2021-06-30	2.6% (157) 是否引证: 否
26	自适应学习率梯度下降的优化算法 宋美佳;贾鹤鸣;林志兴;卢仁盛;刘庆鑫; - 《三明学院学报》 - 2021-12-20	1.6% (98) 是否引证: 否
27	基于脑电信号分析的癫痫发作预测 朱柠(导师: 魏海坤) - 《东南大学硕士论文》 - 2021-05-29	1.5% (92) 是否引证: 否
28	基于卷积神经网络的《黑暗之魂》游戏AI设计 王靖; - 《数字通信世界》 - 2022-02-20	1.4% (86) 是否引证: 否

原文内容

第3章新冠疫情数据的预测分析方法

3.1 引言

在做好充分的文献调研以及数据准备后，需要进一步对疫情数据进行建模分析，主要是对包括SEIR模型、改进的SEIR模型和LSTM模型在内的各类模型进行验证，分析预测疫情的发展趋势并比较各模型的性能优劣。

SEIR模型以及改进的SEIR模型都属于传统的流行病动力学建模中的单一人群模型，这一类模型还包括SI模型、SIR模型等。SI模型没有设置康复者（R）仓室，适用于感染者不可康复的情况，而SIR模型则没有考虑潜伏者（E）的存在，这两种模型均与新冠疫情实际医疗实践情况不符，因而本章着重讨论SEIR系列模型。另一方面，本章从时序序列分析的角度采取LSTM网络模型进行了模型验证，并对以上三种模型进行了比较。

3.2 SEIR系列模型原理详述

单一人群模型将整个人群作为研究对象，将人群划分为不同的仓室，通过研究各仓室的人群数量变化规律来预测疫情的发展趋势，其中SEIR模型是最经典的传染病模型，它将人群分为4类（4个仓室）：

(a) 易感者(susceptible)。此前未接触该病毒，健康但易于感染病毒的人，总人数为S。

(b) 潜伏者(exposed)，已感染病毒但仍未出现病症的人，总人数为E。

(c) 感染者(infected)，已感染病毒并出现病症的确诊患者，总人数为I。

(d) 康复者(recovered)，感染病毒后因病死亡或成功治愈的人，总人数为R。

四类人群之间的转化关系如图3-1所示：

图3-1 SEIR模型传播动力图

图3-1中包含四类人群和三类转化关系，是从易感者S到康复者R的单向转化过程。其中 $\beta$ 为感染者I将病毒传染给易感者S的概率， $\alpha$ 为潜伏者E转化为感染者I的概率， $\gamma$ 为感染者I转化为康复者R的概率。方程组为SEIR的动力学方方程组（1）程：

方程组（（3-4）（3-3）（3-2）（3-1）1）是一个微分方程组，其中模型假设人群总数固定不变为N，即有 $S+E+I+R=N$ ，同时注意到，这也说明人群的总数不会变化。另外观察到式（3-1）恒为负值，这表现为易感者S的数量会持续下降，式（3-4）的值恒为正值，这表现为康复者R的数量会不断上升。SEIR的动力学方程组（1）表明传染病的流行过程是一个单向的过程，从每

一个方程式等号右边的算式可以看出，E是S和I的中间状态，S减少的部分（被感染的部分）转移到E，E减少的部分（感染后确诊的部分）转移到I，I减少的部分（恢复健康并具有免疫力的部分）最后转移到R。设定初值  $N=1000$ ， $\beta=0.272$ ， $\alpha=0.0715$ ， $\gamma=0.091$ ，借助数值积分得到SEIR方程组的数值解如图3-2所示：

图3-2 SEIR模型传播动力图

方程组3-2

图3-2中横轴表示距疫情爆发以来的天数，纵轴表示四类人群S（易感者）、E（潜伏者）、I（感染者）、R（康复者）的人数。可以看出，易感者人数持续降低，康复者人数持续升高，两者均在150天左右到达稳定状态。而潜伏者和感染者的人数走势类似，都是先升高达到峰值后回落，区别在于：潜伏者峰值较高且出现时刻更早。

方程组（2）针对新冠疫情的实际情况，首先人群中存在潜伏者E，其次医疗实践表明潜伏者E也具备一定的感染性，据此对方程组（1）作出修正得到方程组（2）：

相对于方程组（1），方程组（2）中将参数 $\beta$ 替换为 $\beta'$ ，其中参数 $\beta'$ 表征了潜伏者E对易感者I的感染率。修正后的SEIR模型各类人群转化关系如图3-3所示：

图3-3 修正SEIR模型的传播动力图

在修正SEIR模型的传播动力图中，四类人群仍然是单向转化，不同之处在于易感者可以因接触潜伏者而有的几率转化为潜伏者。在方程组3-2中设定初值 $N=1000$ ， $\beta=0.127$ ， $\beta'=0.145$ ， $\alpha=0.0715$ ， $\gamma=0.091$ ，借助数值积分得到修正的SEIR方程组的数值解如图3-4所示：

图3-4 修正SEIR模型传播动力图

相对于SEIR模型，考虑到潜伏者也会感染易感者的修正SEIR模型（以下简称mSEIR模型）其各类人群数量的变化趋势并没有变化，但此时潜伏者、感染者峰值数量均有所升高，峰值出现时间也有所提前，较好地符合了修正后的假设。

### 3.3 基于深圳市数据的SEIR系列模型分析

SEIR系列模型假设人群的每个个体相同，人群均匀混合（homogeneous mixing）；接触是瞬时的，接触与历史无关；每个仓室的人群数量足够大，在传染病流行过程中，感染率、恢复率为常数。因此这一类方法一般适用于疫情爆发初期短时间内某个城市的疫情趋势（走向）的预测和判断，本文以深圳市的疫情数据为例展开对SEIR系列模型的验证分析。

首先，需要从中国各省市疫情情况数据集中查询得到深圳市2020年02月到2022年02月的疫情数据，包括每日的新增\累计本土确诊人数、新增\累计死亡人数、新增\累计治愈人数，截取第一次爆发的数据来对SEIR模型、mSEIR模型的参数进行拟合，时间范围从2020年01月20日到2020年04月15日。同时，对于实际的易感者、潜伏者、感染者和康复者数量，可以确定的是两个，康复者数量也就是累计治愈数（Heal），而感染者数量用累计确诊数（Confirm）减去累计治愈数再减去累计死亡数（Dead）得到。计算公式如方程组（3）所示：

同时，潜伏者数量 $E(t)$ 一般可以用无症状感染者数量代替，但由于早期检测手段和技术还不够成熟，又由于潜伏者本身就意味着没有症状，难以发现，检测代价较大，因此没有相应的数据。然而，如果从实际疫情防控需要的角度出发，对于峰值点的预测是其核心，包括何时会到达疫情传播高峰期以及此时感染者的数量，对于其它三类人群的关注程度更少。

对于方程参数确定问题，首先通过参数的实际意义结合医疗实践对部分参数进行，包括 $\alpha$ 感染率和 $\gamma$ 治愈率，依据隔离14天的标准取 $\alpha=1/7$ ， $\gamma=1/14$ ， $\beta$ 定义一个合理初值例如0.08。使用均方差函数和梯度下降算法对 $\alpha$ 和 $\gamma$ 和 $\beta$ 进行调整，最后得到 $\alpha=0.2$ ， $\beta=0.0035$ ， $\gamma=0.04$ 。最终的拟合图像如图3-6所示：

图3-6 深圳市新冠疫情SEIR模型的参数拟合

图3-6中两条带点的曲线为实际数据，另外两条不带点的光滑曲线为拟合曲线，可以看到对于感染者数据预测效果较好，峰值出现日期基本一致，在2月8日左右；另一方面峰值点高度预测值低于实际值，出现这一现象的原因在于实际情况与SEIR模型假设存在一定程度的不符，例如，这一时期内会有境外输入导致感染率升高，与感染率恒定的假设不符；疫情爆发初期人们需要一定的反应时间，医药方法的研究还未及时跟进，治愈率相对较低等。

针对mSEIR模型，由于使用的数据集相同，且 $\alpha$ 和 $\gamma$ 具备实际意义，从而可以将这两个参数固定下来，对 $\beta$ 和 $\beta'$ 进行参数拟合，同样使用均方差作为误差函数，使用梯度下降算法进行拟合。结果如图3-7所示：

图3-7 深圳市新冠疫情mSEIR模型的参数拟合

在改进的SEIR模型参数拟合图像中，感染者预测曲线在到达峰值前与实际值较原SEIR模型更加吻合。两个模型都能较好的预测峰值来临的日期，而峰值点人数预测值均偏低，在峰值点来临后，相较于实际患病人数曲线，预测曲线下降更缓慢；另一方面观察康复者曲线，预测曲线与实际曲线均呈现单调上升的状态，两曲线之间有一个交点，在相交之前，预测值偏高，而在相交之后，预测值又偏低，随着时间推移，两曲线又再次接近并趋向稳定。

从上述描述和分析可以得出：首先SEIR系列模型能够在一定程度上预测实际新冠疫情初期的发展趋势，能够作为实际医疗实践和人工干预措施的科学参考。其次由于模型较为严苛的假设，预测曲线与实际曲线又存在偏差，一个合理的猜想是感染率和治愈率并不是一直为常数，在疫情爆发起始时，由于人工干预还未到位以及对新冠病毒的医疗实践不够，新冠疫情感染率较高而治愈率较低，导致初始时感染者不断累积滞留，很快到达一个较高的峰值，此时康复者数量较少，上升缓慢；此后随着人工干预措施例如隔离等的实施，使得感染率大幅降低，感染者累积速度大大放缓，另一方面康复率变高，两者结合导致感染者数量下降较为迅速，康复者数量上升也较快，最后随着感染者的持续减少，康复者的持续增多直到趋于稳定，表明一轮疫情传播的结束。

### 3.4 基于长短期记忆网络的时序分析

#### 3.4.1 RNN网络模型中存在的问题

循环神经网络（Recurrent Neural Networks, RNN）或称循环神经网络，是在深度学习处理时序问题最常使用的模型之一。RNN之所以在时序数据上有着优异的表现是因为RNN在 $t$ 时间片时会把 $t-1$ 时间片的隐节点作为当前时间片的输入[8]。

梯度消失和梯度爆炸是困扰RNN模型训练的关键原因之一，产生梯度消失和梯度爆炸是由于RNN的权值矩阵循环相乘导致的，相同函数的多次组合会导致极端的非线性行为。梯度消失和梯度爆炸主要存在RNN中，因为RNN中每个时间片使用相同的权值矩阵。

另一方面在深度学习领域中，长期依赖（Long Term Dependencies）问题是普遍存在的。长期依赖产生的原因是当神经网



络的节点经过许多阶段的计算后，之前比较长的时间片的特征已经被覆盖，这导致了信息的损失。

处理梯度爆炸可以采用梯度截断的方法。所谓梯度截断是指将梯度值超过阈值的梯度手动降到阈值范围内。虽然梯度截断会一定程度上改变梯度的方向，但梯度截断的方向依旧是朝向损失函数减小的方向。对比梯度爆炸，梯度消失不能简单的通过类似梯度截断的阈值式方法来解决，因为长期依赖的现象也会产生很小的梯度。

3.4.2 LSTM网络的工作原理与改进点

LSTM(Long Short Time Memory)或称长短期记忆网络是一种特殊的RNN，针对原始的RNN结构存在的问题，LSTM对循环节点结构做出了改进，相对于RNN更加复杂，如图3-8所示为LSTM的网络结构[8]：

图3-8 LSTM网络的链式结构式(3-1)

理解LSTMs的关键就是图中的矩形方框，被称为memory block（记忆块），主要包含了三个门（forget gate、input gate、output gate）与一个记忆单元（cell）[8]。LSTM第一步是用来决定什么信息可以通过cell state。这个决定由“遗忘门通过sigmoid来控制，它会根据上一时刻的输出和当前输入来产生一个0到1 的值（式(3-1)），来决定是否让上一时刻学到的信息通过或部分通过。

式(3-4)式(3-3)式(3-2)

第二步是产生我们需要更新的新信息，这一步包含两部分，第一个是输入门通过sigmoid来决定哪些值用来更新（式(3-2)），第二个是一个tanh层用来生成新的候选值，它作为当前层产生的候选值可能会添加到cell state中（式(3-3)）。我们会把这两部分产生的值结合来进行更新（式(3-4)），此时就完成了长期记忆和部分遗忘的工作。

式(3-5)式(3-6)

最后一步是决定模型的输出，首先是通过sigmoid层来得到一个初始输出（式(3-5)），然后使用tanh将值缩放到-1到1间，再与sigmoid得到的输出逐对相乘，从而得到模型的输出（式(3-6)）。

3.4.3 使用LSTM网络分析疫情数据式(3-6)

LSTM网络可以借助python的依赖库pytorch中的nn.Module类进行实现，整个过程分为训练和预测两个部分，训练时采用MSE作为误差函数，MSE即Mean Square Loss，表示平方平均误差，基本形式为：

同时使用Adam算法进行优化。Adam 算法，即一种对随机目标函数执行一阶梯度优化的算法，该算法基于适应性低阶矩估计。Adam 算法很容易实现，并且有很高的计算效率和较低的内存需求。

在进行模型训练时，首先定义训练比（ratio），也即用于训练的数据占比，考虑两种情况：一是使用一个通常情况下较高的训练比，例如0.60；二是针对实际新冠疫情数据预测而言，如果只将峰值之前的数据用于训练，模型能否预测到峰值点的到来，此时训练比为0.11。

在两种情况下，LSTM模型训练结果如图3-9所示：

- (a) 训练比=0.60时LSTM的工作情况
- (b) 训练比=0.11时LSTM的工作情况

图3-9 不同训练比情况下LSTM的工作情况

图3-9表明，在有充分的训练数据时，LSTM网络能够很好的学习到实际疫情发展的趋势，预测值与实际值十分吻合；而在训练样本容量较小时，LSTM模型完全没有学习到康复者数量变化的趋势，但感染者的总体变化趋势预测是正确的，LSTM网络预测到了峰值的出现，而不是持续的单调上升或者放平。因此，当得到充分的训练样本数据时，LSTM网络模型的表现较好。

3.5 模型间的比较与分析

本文使用了统一的误差评判标准RMSE（均方根误差）和MAE（平均绝对误差）对几种模型的预测结果进行了计算和评判，结果如表3-1所示：

表3-1 各类模型在深圳市早期新冠疫情数据集上的预测情况

模型	RMSE	MAE
SEIR	34.48688	28.051678
mSEIR	34.33434	27.573588
LSTM(ratio=0.60)	15.64497	12.75899
LSTM(ratio=0.11)	196.01667	169.45608

模型 RMSE MAE  
SEIR 34.48688 28.051678  
mSEIR 34.33434 27.573588  
LSTM(ratio=0.60) 15.64497 12.75899  
LSTM(ratio=0.11) 196.01667 169.45608

整体来看，LSTM（ratio=0.60）模型预测值与实际值最为接近，在训练数据足够时，LSTM模型能够学习到实际疫情的发展趋势，另一方面LSTM（ratio=0.11）模型预测值与实际值偏差最大，在训练数据不足时，LSTM模型不足以学习到实际疫情的发展趋势，这说明了LSTM网络模型是不依靠先验知识或预先假设的，而是依赖于数据并从中进行学习。而SEIR模型和mSEIR模型总体误差接近，都表现较好，而mSEIR模型略胜一筹，这两个模型依赖于数学假设，与实际情况难免有偏差，但另一方面其内在规律又大体能够描述疫情发展趋势，因此能够较好地预测到感染者峰值点到来的日期。

3.6 本章小结

本章主要讨论了SEIR系列模型和LSTM模型的基本假设与设计原理，预置SEIR系列模型的参数进行了仿真，值得注意的地方是感染者数量曲线，它是先上升后下降的，从而有一个峰值点，这是本章研究的重点。同时，本章还阐述了LSTM模型的相对于原始RNN网络的几点改进之处，并对SEIR模型的假设提出了修正，主要是加入了“潜伏者有几率接触感染易感者”的因素，同时对方程也做了相应的修正。

最后，为了进一步探讨模型的实际性能，本章还抽取了深圳市疫情爆发初期的数据，对以上三类模型进行了训练和参数拟

合，并将结果绘制为图表的形式，评价了他们的优劣与适用性。

指 标	
疑似剽窃文字表述	
<div>1. 感染者I将病毒传染给易感者S的概率，<math>\alpha</math> 为潜伏者E转化为感染者I的概率，<math>\gamma</math> 为感染者I转化</div> <div>2. 梯度消失和梯度爆炸是困扰RNN模型训练的关键原因之一，产生梯度消失和梯度爆炸是由于RNN的权值矩阵循环相乘导致的，相同函数的多次组合会导致极端的非线性行为。梯度消失和梯度爆炸主要存在RNN中，因为RNN中每个时间片使用相同的权值矩阵。</div> <div>3. 长期依赖产生的原因是当神经网络的节点经过许多阶段的计算后，之前比较长的时间片的特征已经被覆盖，</div> <div>4. 处理梯度爆炸可以采用梯度截断的方法。所谓梯度截断是指将梯度值超过阈值的梯度手动降到阈值范围内。虽然梯度截断会一定程度上改变梯度的方向，但梯度截断的方向依旧是朝向损失函数减小的方向。对比梯度爆炸，梯度消失不能简单的通过类似梯度截断的阈值式方法来解决，因为长期依赖的现象也会产生很小的梯度。</div> <div>5. Adam 算法，即一种对随机目标函数执行一阶梯度优化的算法，该算法基于适应性低阶矩估计。Adam 算法很容易实现，并且有很高的计算效率和较低的内存需求。</div>	
5. 第4章基于时空点过程的新冠疫情数据分析	总字数：4656
相似文献列表	
去除本人文献复制比：3.3%(153)      文字复制比：3.3%(153)      疑似剽窃观点：(0)	
1 基于蒙特卡罗的移动节点定位算法研究 马征征(导师：陈嘉兴) - 《河北师范大学硕士论文》 - 2013-03-04	2.0% (94) 是否引证：否
2 最大似然估计log likelihood - Chloezhao的专栏 - CSDN博客 - 《网络 ( <a href="http://blog.csdn.net">http://blog.csdn.net</a> ) 》 - 2017	1.9% (89) 是否引证：否
3 16级川大金融曾珏栋论文初稿 (3) 曾珏栋 - 《大学生论文联合比对库》 - 2020-05-19	1.7% (78) 是否引证：否
4 模式识别 - Xing的专栏 - CSDN博客 - 《网络 ( <a href="http://blog.csdn.net">http://blog.csdn.net</a> ) 》 - 2017	1.6% (75) 是否引证：否
5 Xing的专栏 - CSDN博客 - 《网络 ( <a href="http://blog.csdn.net">http://blog.csdn.net</a> ) 》 - 2017	1.6% (75) 是否引证：否
6 机器学习 - shulixu的博客 - CSDN博客 - 《网络 ( <a href="http://blog.csdn.net">http://blog.csdn.net</a> ) 》 - 2017	1.6% (75) 是否引证：否
7 Polaris - CSDN博客 - 《网络 ( <a href="http://blog.csdn.net">http://blog.csdn.net</a> ) 》 - 2017	1.6% (75) 是否引证：否
8 基于模糊隶属度的贝叶斯网络模型 (FBM) 构建与物种生境适宜性分析 张权中(导师：卫海燕) - 《陕西师范大学硕士论文》 - 2019-05-01	1.5% (72) 是否引证：否
9 概率统计在产品质量验收抽样方案确定中的应用 杜志高 - 《大学生论文联合比对库》 - 2019-05-14	1.5% (71) 是否引证：否
10 李博文_120151504_人工胰腺系统模型参数自适应方法设计 李博文 - 《大学生论文联合比对库》 - 2019-05-27	1.5% (69) 是否引证：否
11 属性网络中可解释性社区发现方法研究 赵琪琪(导师：马慧芳) - 《西北师范大学硕士论文》 - 2021-05-01	1.3% (60) 是否引证：否
12 基于TMS320C6701EVM的嵌入式数字锁定放大器 钱黎平(导师：程德福) - 《吉林大学硕士论文》 - 2005-05-10	1.3% (60) 是否引证：否
13 李博文_120151504_人工胰腺系统模型参数自适应方法设计 李博文 - 《大学生论文联合比对库》 - 2019-05-29	1.2% (57) 是否引证：否
14 最大似然估计和贝叶斯参数估计.md - 我们的征途是星辰大海 - CSDN博客 - 《网络 ( <a href="http://blog.csdn.net">http://blog.csdn.net</a> ) 》 - 2017	1.2% (57) 是否引证：否



15	图像高斯噪声估计的极大似然算法	王蓓;张根耀;李智; - 《计算机技术与发展》- 2014-04-24 0	1.2% (55)	是否引证: 否
16	例谈概率统计教学中培养学生数学应用意识	熊淑艳; - 《考试周刊》- 2015-06-16	1.2% (55)	是否引证: 否
17	基于偏正态分布和逆尺度因子 $\gamma$ 偏正态分布的Tobit回归模型	王聪(导师:王通会) - 《西北农林科技大学硕士论文》- 2016-05-01	1.2% (55)	是否引证: 否
18	机电-1000150215-蓝昕	机电 - 《大学生论文联合比对库》- 2014-05-30	1.2% (55)	是否引证: 否
19	王大鹏-0911123043-统计学中各种似然方法的探讨	王大鹏 - 《大学生论文联合比对库》- 2013-06-18	0.9% (43)	是否引证: 否
20	曹子琦-基于差分进化算法的TDOA定位技术0521	曹子琦 - 《大学生论文联合比对库》- 2021-05-21	0.9% (43)	是否引证: 否

原文内容

第4章基于时空点过程的新冠疫情数据分析

4.1 引言

在第3章中，本文主要针对疫情发展趋势的分析和预测这一问题分别从传统的单一人群仓室模型和近年来在深度学习中应用广泛的时序序列分析模型LSTM网络两个角度进行了研究和讨论。另一方面，由于模型假设的限制，在第3章中使用到的数据集是中国各省市历史疫情数据的子集，包括空间范围限制在深圳市而时间范围限制在2020年2月第一轮疫情爆发到4月扩散基本结束，因而数据利用不够充分。

本章采用最新流行的时空点过程（Spatial-Temporal Point Process）视角对中国各省市历史疫情数据集进行了分析和研究，时空点过程分析方法的研究对象是一系列分布在连续时空域的抽象点（或称事件）[12]，每一个事件由一个时空坐标唯一确定，事件也可以有附加属性。顾名思义，时空点过程分析方法下有两个子研究分支，包括时间点过程分析法和空间点过程分析法，本章主要围绕Poisson点过程、Hawkes点过程、神经Hawkes点过程等模型展开对中国的疫情数据的研究分析。

4.2 时空点过程的事件建模法

4.2.1 时空点过程的一般形式

令表示一个事件序列，其中，是对应的空间坐标，同时，令表示发生在t时刻之前的所有事件。如式(4-1)所示，一个时空点过程可以使用一个条件概率强度函数完整刻画[17]：

式(4-1)

式4-1中的一个定义在上的以x为中心以 $\Delta x$ 为邻域半径的带心领域（相对于去心邻域）。注意到必须是一个正数（概率的实际意义），不妨将其简记为，也即。在时间间隔[0, T]内对于观测值的联合对数似然值由式(4-2)给出[20]：

式(4-2)

式(4-2)中包含多重积分，因此想要以此为极大似然函数训练一个一般性的STPP（Spatial-Temporal Point Process）模型是十分困难的。为了保证模型仍然能够使用较高精度的似然估计函数，本文采用RTQ Chen的方法，利用Neural ODE（Neural Ordinary Differential Equations）框架对STPP模型进行参数化[20]，这样就简化了计算过程。

式(4-3)

如式(4-3)所示简化过程的第一步是将进行分解：

式(4-4)式(4-5)式(4-6)

式(4-3)中的\*符号仍然表示某事件在条件下的概率，那么表示相对应的时间点过程的偏概率强度函数，表示在条件下在时刻t存在事件在位置x发生的条件概率，因此式(4-2)可以简化为：

因此，一个时空点过程可以使用一个和函数表示，一部分是时间点过程条件强度函数，另一部分是空间点过程的推荐强度函数，此时整个方程组中只包含一个一重积分（对时间）。

4.2.2 几种点过程模型的探讨

(1) 泊松过程（Poisson Process）

在4.1.1中的讨论中可以知道，使用一个条件强度函数可以完全地表示一个点过程。泊松点过程认为是独立于历史事件的，其中同质泊松过程（homogeneous）认为是一个常数，即；而非同质泊松过程则认为是随时间变化的，即。

(2) 自校正过程（Self-Correcting Process）

式(4-7)

自校正过程认为强度函数的趋势是一直在增大，但是当有一个事件发生后，会先减小，式(4-7)为自校正过程的强度函数：

自校正过程的强度函数是以欧拉数为底数的指数函数的形式，其中指数部分的 $\mu t (\mu > 0)$ 表示强度函数随着时间的推移持续的增大，而部分则表明每当有事件发生时，强度函数的值会减小。

(3) 霍克斯点过程（Hawkes Point Process）

式(4-7)

霍克斯过程是一种强度函数值依赖于历史事件的自激励点过程，其特征是其历史事件的影响以时间累加的形式进行，霍克斯过程的强度函数定义如式(4-7)所示[21]。其中表示截止到时间t的所有历史事件，是给定的确定性函数， $\phi$ 是核函数，描述了过去事件对当前事件的影响。也就是说，这是对在时间点t之前的所有事件都进行了考虑。为了描述这些影响随着时间的衰减效应，尤其是最新事件对未来事件的最大影响，核函数一般采用类似于指数函数的形式随着时间变化的强度函数的实例如图4-1所示。

图4-1 核函数的变化曲线

图4-1展示了霍克斯过程的强度函数曲线随时间推移发生的变化，可以看到历史事件对未来事件的影响是单调指数递减，然后以累加形式进行叠加。其中横轴标注了各个事件发生的时间，曲线上的空心圆点则代表着事件，每当有新事件发生时，事件产生一个正向激励，使强度函数呈跳跃式增长。

以上三种点过程属于传统点过程，都具有较强的解释性，需要一定的先验知识，并利用一定的数学假设和公式对条件强度函数进行刻画。然而，传统点过程对强度函数有着上述设定，很有可能不符合实际情况，比如历史事件对强度函数的影响并不一定是累加的；此外，如果有多种事件类型的话，还需作出各个事件类型是互相独立的假设，并且对每个事件类型求强度函数。

深度点过程则只需使用神经网络这样的非线性函数模拟强度函数，这样一个黑盒子无需设定任何先验知识。本文研究的一个深度点过程是基于霍克斯过程的神经霍克斯点过程，采用Mei H教授的训练方法[22]利用神经网络模拟了霍克斯过程的核函数。

4.3 基于点过程模型的新冠疫情数据分析

4.3.1 数据的预处理

此处需要用到的数据集中国各省市历史疫情数据的全部有效数据，也即需要过滤出所有包含新增本土确诊病例的记录，每条记录被日期和城市唯一确定，最终过滤得到的数据一共约8000条。

其次，利用城市代码在中国行政区划信息表中查询得到面积，利用百度地图API获取得到经纬度，以2020年02月01日为起点，将日期转为距起始日期的天数days这样就得到了初始输入，如表4-1所示：

表4-1 预处理后得到的初始输入

天数	地区	新增病例数	面积Km2	经度	纬度
0	恩施州	105	24, 111	109.49459	30.27794
0	黔南州	1	26, 195	107.52840	26.26062
0	黔东南州	2	30, 339	107.98945	26.58970
0	福州	5	11, 974	119.30347	26.08043
0	泉州	5	11, 015	118.68245	24.87995
0	莆田	2	4, 132	119.01452	25.45987
0	厦门	1	1, 701	118.09644	24.48541
0	漳州	1	12, 882	117.65358	24.51893
0	三明	1	22, 965	117.64552	26.26974
0	温州	24	11, 784	120.70648	28.00109
0	杭州	12	16, 596	120.21551	30.25308

天数地区新增病例数面积Km2 经度纬度

0 恩施州 105 24, 111 109.49459 30.27794  
0 黔南州 1 26, 195 107.52840 26.26062  
0 黔东南州 2 30, 339 107.98945 26.58970  
0 福州 5 11, 974 119.30347 26.08043  
0 泉州 5 11, 015 118.68245 24.87995  
0 莆田 2 4, 132 119.01452 25.45987  
0 厦门 1 1, 701 118.09644 24.48541  
0 漳州 1 12, 882 117.65358 24.51893  
0 三明 1 22, 965 117.64552 26.26974  
0 温州 24 11, 784 120.70648 28.00109  
0 杭州 12 16, 596 120.21551 30.25308

4.3.2 点过程模型的部署与训练

模型的建立和训练主要依赖于python的第三方库pytorch，并使用CUDA工具借助GPU进行加速，使用对数似然值对模型的结果进行评价分析。

最大似然估计法的思想在于：在已经得到试验结果的情况下，我们应该寻找使这个结果出现的可能性最大的那个作为真值x的估计，因而问题转化为求极大似然函数的最大值点，而由于往往是乘积的形式并且对数函数是单调增函数，取对数后求解更简单。对关于X求导数，并命其等于零，得到的方程组称为似然方程组。解方程组得到X的最大似然估计。

结合式4-5和式4-6，通过python的scipy库中的数值积分包进行计算，得到数据结果如表4-2所示：

表4-2 几种点过程模型在中国疫情数据集上的训练结果

model	Spatial log-likelihood	temporal log-likelihood
Poisson	-3.19491±1.23	0.17915±0.023
self-correcting	-2.19491±1.24	0.25971±0.022
Hawkes	-0.39191±0.012	0.14253±0.023
neural Hawkes	-0.38132±0.022	0.04132±0.013

model Spatial log-likelihood temporal log-likelihood  
Poisson -3.19491±1.23 0.17915±0.023  
self-correcting -2.19491±1.24 0.25971±0.022  
Hawkes -0.39191±0.012 0.14253±0.023  
neural Hawkes -0.38132±0.022 0.04132±0.013

结果表明，总体来看，几个过程的空间对数似然值差别不大，表现都比较好。在时间对数似然值上，泊松过程模型和自校正模型表现较差，其关于强度函数的假设与实际新冠疫情爆发的时空过程有较大偏差，而基本的霍克斯过程实验结果已经表现

相当良好,但基于神经的神经 STPP 可以实现更好的空间可能性,这也说明了神经网络的优势在于不需要或者只需要很少的先验知识和假设,而是通过网络来模拟其条件强度函数。

#### 4.4 本章小结

本章首先介绍了时空点过程的一般形式,指明条件强度函数能够完全地刻画一个时空点过程,给出了时空点过程的一般形式。然而,由于其对数似然函数包含多重积分,想要直接进行训练是困难的,因而可以采取参数化的方法将对数似然函数简化以适用于训练。其次介绍了几个典型的点过程模型的方法原理,包括泊松点过程模型、自校正点过程模型和霍克斯点过程模型,并基于中国各省市历史疫情数据集对以上三个模型以及神经霍克斯过程模型进行了训练。最后利用对数似然值对模型进行评价分析,相较于其他模型,霍克斯模型以及神经霍克斯模型在中国的历史疫情数据集上有着更为出色的表现。

#### 结论

本文收集了较为详细的中国各省市的历史疫情数据,同时还包括了我国各行政区的基本信息,形成了一个较为全面且准确的数据库,并针对性的完成了一系列的可视化工作,包括疫情趋势走向折线图、疫情地图等等。

本文首先从传统的仓室模型中选取了SEIR模型进行了模拟,同时对从“潜伏者也可能接触并感染易感者”这一角度对SEIR模型进行修正后的模型也进行了模拟,结果表明感染者曲线呈现先升后降的趋势,其顶点位置的预测成为工作的重点。随后本文利用深圳市的真实数据集对SEIR、mSEIR模型进行了参数拟合,在训练过程中,本文创新性的将感染率、恢复率等参数通过医疗实践知识进行预确定,从而可以舍弃不确定的潜伏者数据,使其余参数的拟合过程依赖于确定的数据。

此外,本文创新性的跳出传统的传染病动力学的分析角度,将疫情数据视为时序数据,并针对LSTM模型进行了实验,结果表明在相同的深圳市早期疫情数据集上,当训练样本容量充足时,LSTM网络模型预测值与真实值最为接近,同时SEIR系列模型也能够较好的预测新冠疫情的初期发展趋势。进一步的,将空间坐标也考虑进来,本文创新性的从时空点过程分析方法在中国历史疫情数据集上展开了几个时空点过程模型的训练,并利用对数似然值对模型进行了评价分析。

针对本文的工作,还可以从以下角度进行进一步的研究分析:

(1) SEIR模型往往只能模拟一轮疫情爆发,且其中在疫情散播过程中不会变,能否借助深度学习模型识别新一轮疫情爆发,清空旧的参数,在一轮疫情扩散过程中对参数进行自适应的调整;

(2) 针对SEIR模型本身,还可以做许多其它的修正,例如加入复阳者人群,他们在康复后可能复发;在开发出疫苗后,一部分人群,他们提前接种了疫苗,可以考虑加入预接种人群等;

(3) 针对时空点过程,还可以做更深入的分析,例如基于深圳市详细病例数据集的研究,同时需要做更深入的探讨使结果更清晰易懂。

#### 参考文献

- [1] Santosh, K.C. COVID-19 Prediction Models and Unexploited Data. J Med Syst 44, 170 (2020)
- [2] Alazab M, Awajan A, Mesleh A, et al. COVID-19 prediction and detection using deep learning[J]. International Journal of Computer Information Systems and Industrial Management Applications, 2020, 12(June): 168-181.
- [3] 裴韬,王席,宋辞,等. COVID-19 疫情时空分析与建模研究进展[J]. 地球信息科学学报, 2021, 23(2): 188-210.
- [4] Zareie B, Roshani A, Mansournia M A, et al. A model for COVID-19 prediction in Iran based on China parameters[J]. Archives of Iranian medicine, 2020, 23(4): 244-248.
- [5] 张发,李璐,宣慧玉. 传染病传播模型综述[J]. 系统工程理论与实践, 2011, 31(9): 1736-1744.
- [6] Kermack W O, McKendrick A G. Contributions to the mathematical theory of epidemics, part I [J]. Proceedings of the Royal Society of London A, 1927, 115: 700-721.
- [7] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network[J]. Physica D: Nonlinear Phenomena, 2020, 404: 132306.
- [8] Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural computation, 2019, 31(7): 1235-1270.
- [9] 赵永翼,王菲,申莹. 基于长短期记忆网络的 COVID-19 疫情趋势序列分析预测[J]. 沈阳师范大学学报(自然科学版), 2020.
- [10] Ogata Y. Space-time point-process models for earthquake occurrences[J]. Annals of the Institute of Statistical Mathematics, 1998, 50(2): 379-402.
- [11] Stoyan D, Penttinen A. Recent applications of point process methods in forestry statistics[J]. Statistical science, 2000: 61-78.
- [12] Junchi Yan. Recent Advance in Temporal Point Process: from Machine Learning Perspective[R]. Think Lab 2019.
- [13] David R Cox. Some statistical methods connected with series of events. Journal of the Royal Statistical Society. Series B (Methodological), pages 129 - 164, 1955.
- [14] A. G Hawkes. Spectra of some self-exciting and mutually-exciting point processes. Biometrika, 58, 1971
- [15] Sebastian Meyer, Johannes Elias, and Michael Hohle. A space-time conditional intensity model for invasive meningococcal disease occurrence. Biometrics, 68(2): 607 - 616, 2012.
- [16] Junhyung Park, Adam W Chaffee, Ryan J Harrigan, and Frederic Paik Schoenberg. A non-parametric Hawkes model of the spread of ebola in west africa. 2019.
- [17] RTQ Chen, Amos B, Nickel M. Neural spatio-temporal point processes[J]. arXiv preprint arXiv:2011.04583, 2020.
- [18] 孙立伟,何国辉,吴礼发. 网络爬虫技术的研究[J]. 电脑知识与技术, 2010 (15): 4112-4115.
- [19] Xue H, Huynh D Q, Reynolds M. SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018: 1186-1194.

[20] Adrian Baddeley, Imre Bar' any, and Rolf Schneider. Spatial point processes and their applications. 'Stochastic Geometry: Lectures Given at the CIME Summer School Held in Martina Franca, Italy, September 13-18, 2004, pp. 1-75, 2007.

[21] Daryl J Daley and David Vere-Jones. An introduction to the theory of point processes, volume 1:Elementary theory and methods. Verlag New York Berlin Heidelberg: Springer, 2003.

[22] Mei H, Eisner J M. The neural hawkes process: A neurally self-modulating multivariate point process[J]. Advances in neural information processing systems, 2017, 30.

指 标

疑似剽窃文字表述

1. 似然值对模型的结果进行评价分析。  
最大似然估计法的思想在于：在已经得到试验结果的情况下，我们应该寻找使这个结果出现的可能性最大的那个作为真值x的估计，因而问题转化为求极大似然函数的最大值点，而由于往往是乘积的形式并且对数函数是单调增函数，

说明：1. 总文字复制比：被检测论文总重合字数在总字数中所占的比例

2. 去除引用文献复制比：去除系统识别为引用的文献后，计算出来的重合字数在总字数中所占的比例

3. 去除本人文献复制比：去除作者本人文献后，计算出来的重合字数在总字数中所占的比例

4. 单篇最大文字复制比：被检测文献与所有相似文献比对后，重合字数占总字数的比例最大的那一篇文献的文字复制比


5. 复制比：按照“四舍五入”规则，保留1位小数

6. 指标是由系统根据《学术论文不端行为的界定标准》自动生成的

7. 红色文字表示文字复制部分；绿色文字表示引用部分；棕灰色文字表示系统依据作者姓名识别的本人其他文献部分

8. 本报告单仅对您所选择的比对时间范围、资源范围内的检测结果负责



 [amlc@cnki.net](mailto:amlc@cnki.net)

 <https://check.cnki.net/>

知网查重系统