湖南師範大學

硕士学位论文

基于机器学习的武警警务数据研究

学　科　专　业　计算机技术

学　位　类　型　□科学学位　☑专业学位

研　究　生　姓　名　周聪

指导教师姓名、职称　罗迅　讲师

论　文　编　号

湖南师范大学学位评定委员会办公室

二〇一九年八月

分 类 号_____ 密 级_____

学校代码_____10542_____ 学 号_____

基于机器学习的武警警务数据研究
Research on Data of PAP Based on Machine Learning

研 究 生 姓 名 周聪

导师姓名、职称 罗迅 讲师

学 科 专 业 计算机技术

研 究 方 向 _____

湖南师范大学学位评定委员会办公室

二〇一九年八月

摘要

随着改革强军的稳步推进，部队信息化正在加速发展，各大战区、军兵种、军委机关的数据量迅速增长。武警警务数据作为部队数据之一，虽然具有广阔的应用前景但由于保密规定限制，敏感度高，不易从互联网获取等特点，一直未被各级重视和合理利用，本文在深入了解国内外警务数据分析研究的基础上，合理地采用 LSTM 神经网络模型,对处理后的武警警务数据进行训练，用来预测未来一段时间内某个地区发生危及社会稳定的重大事件的数量。具体研究内容如下：

1.传统的武警警务数据存储方式粗放多样，复杂而繁琐，数据源涉及众多领域，类别多，结构各异，有着大量的无用信息。本文对警务原始数据进行了整理和分类，采用删除法和填充法处理数据中的缺失值，对于离散型数据运用 one-hot 编码来处理属性分类问题，并运用数据概化方法来提升数据的一致性，通过将现有数据中的信息进行清洗，筛选，将数据的储存变得标准化，归一化，达到预测系统可以应用的水平。

2.本文对原始武警警务数据进行收集和处理后，将之用于训练 LSTM 神经网络模型。本文设置 24 种不同的参数，进行不同的测试，最终得到了一个相对有价值的预测模型。

关键词：武警警务数据，机器学习，MAE 算法，LSTM 神经网络

Abstract

With the steady progress of reforming and strengthening the army, the informationization of the army is accelerating, and the data volume of the major war zones, arms and Military Commission organs is growing rapidly. As one of the army data, armed police data has a broad application prospect, but because of confidentiality restrictions, high sensitivity and difficult access to the Internet, it has not been paid attention to and rationally utilized at all levels. Based on in-depth understanding of domestic and foreign police data analysis and research, this paper reasonably uses LSTM neural network. The network model, which trains the processed data of the armed police, is used to predict the number of major events that endanger social stability in a certain area in the future. Specific research contents are as follows:

1.The traditional data storage methods of armed police affairs are extensive and diverse, complex and cumbersome. Data sources involve many fields, have many categories and different structures, and have a lot of useless information. In this paper, the police data is sorted and classified, and the missing values in the data are processed by the deletion method and the filling method, one-hot coding is used to deal with attribute classification for discrete data, and data generalization method is used to improve data consistency. By cleaning and filtering the

information in existing data, data storage is standardized and normalized. The level at which the prediction system can be applied.

2. This paper collects and processes the original data of the Armed Police and applies it to the training of LSTM neural network model. In this paper, 24 different parameters are set up, and different tests are carried out. Finally, a relatively valuable prediction model is obtained.

Key words: Armed Police Data, Machine Learning, MAE Algorithms, LSTM Neural Network

目录

第一章　绪论

1.1 课题的研究背景与意义

随着实现中华民族伟大复兴中国梦的伟大工程不断推进，部队发展建设越来越受到党和国家领导人以及社会各界人士的重视[1]。一百多年前，中华民族所遭受的那一场深重的灾难和一幕幕屈辱的历史还依稀可见，中华民族要想屹立于世界民族之林，不仅要在经济上有大的发展和飞越[2]，综合国力也要稳步提升，而强大的军队将是经济腾飞和综合国力跃升的有力支撑[3]，习主席在十八大、十九大报告中都提出了改革强军的战略设想，通过这些年对中国的军事力量和组成结构地改革重构，人民解放军已经插上强军的翅膀，实现了脱胎换骨的变化[4]。

随着部队改革的推进和自身岗位的变化，作者感受到部队每天所产生的数据越来越多，比如人员信息、训练数据、身高体重数据，还有我们武警相关的集访数据、群体性事件数据、反恐怖数据、个人极端事件数据、劫持人质事件数据、有组织犯罪数据等等[5]。这些每天都在生成的数据无疑是一座有价值的宝库，然而，由于这些数据比较零散，而且统计的方式比较粗放、多样化，涉及部门也比较多，所以利用起来比较困难，长期处于废弃状态。

武警部队官兵每天都在进行常态化的执勤处突，反恐维稳，抢险救援，海上维权等任务[6]，所以武警官兵每天都处在敌对势力和犯罪团伙压力之下，目前，敌对势力和犯罪团伙压力的联络方式和破坏手段越来越便捷和隐蔽[7]。例如恐怖分子的袭击、贩卖毒品、暴力攻击、非法走私和军火交易等[8]。这些犯罪团伙通过网络来秘密筹划犯罪活动，散播谣言，制造恐怖舆论[9]。传统的犯罪活动，由于获取消息的延时性和封闭性，为公安机关的破案带来了极大的困难，也导致武警人员不能行之有效地进行预防和抓捕[10]。而如今网络在各行各业都得到了快速的发展[11]，这为犯罪分子的活动提供了巨大空间的同时，也为武警获取犯罪分子的情报提供了数据来源[12]。在大数据时代下，要是能够对这些数据进行合理地分析，就会进一步掌握犯罪分子在网上进行活动的主要特点，就可以提前对这些犯

罪活动进行有效的防控[13]。通过把有效的数据进行挖掘，就会更进一步的获取犯罪分子的逃跑路线，这样不仅可以提前布控[14]，实行有效的逮捕，也能节省大量的财力物力，使得维稳行动变得更加便利[15]。

由此可知，当今犯罪分子对网络的依赖性以及数据收集的重要性。那么从大量数据中筛选出有用的数据，进行分组训练，以便于找到有效的信息就是其中的关键步骤[16]。如今随着科技的进步，对数据的分析也得到了较好的发展。我们可以使用大数据分析工具和技术，结合机器学习来进行建模，虽然犯罪活动并没有什么规律可言，但是这并不代表我们对犯罪活动无法进行预测[17]。如今大数据和数据科技，通过改进协作和数据分析，并通过机器学习进行合适的建模，减低了情报调查过程的繁琐程度，使机构更轻易地检测到犯罪活动的逃跑路线和人员活动情况，也可以做很多其他有价值的研究，比如通过对训练成绩数据和部队人员身高体重数据的分析来评判训练方法的有效性，通过群体性事件发生的时间地域概率来调整全省机动兵力配置等，这些都可为提升部队战斗力服务[18]。

1.2 国内外研究现状

随着信息采集技术的蓬勃发展和数据处理技术的提升，大数据越来越表现出惊人的优势，美国一些选举机构利用舆情分析成功预测总统大选结果，还能预测大部分参议员和众议员的选举结果；超市利用顾客的购买行为数据，高效的设置超市物品摆放策略，从而刺激顾客的消费欲望和增加购买行为，提升利润空间[29]；GOOGLE 利用互联网上抓取的动态数据，成功预测了甲型 H1N1 病毒爆发的地域分布图，比传统卫生部门要快 2-3 周；一家名为 Farecast 的科技公司为了帮助消费者抓住最佳购买时机，对机票价格的走势以及增降幅度进行预测[20]，而在此之前还没有其他网站能让消费者获得这些信息。上述这些都是大数据在思维变革、商业模式变革等各个场景的应用，由于保密，信息集采方法粗放单一，部门之间协调不畅等原因，武警警务数据一直是学术研究的空白领域[21]，当然，武警警务数据和公安警务数据有诸多相似的地方，比如一些恶性案件数据、反恐维稳数据等方面数据模式基本相同[22]；同时武警警务数据又有其独有的一些数据，比如各

类训练数据、现役军人身高体重数据、抢险救援数据等等，所以这要求我在借鉴公安警务数据研究的基础上推陈出新，专注于服务部队战斗力提升这个根本落脚点用力[23]。

1.2.1 国内研究现状

近些年以来，全国各地公安机关也在不断地探索大数据服务于实战的应用。从 2012 年开始，苏州市公安局从美国引进了犯罪预测系统，先在市范围内进行犯罪预测实验[24]。各地政府也紧随时代潮流，在民意调查，社会管理，铁路运输、民生服务、农林牧业等方面开展各种预测方法[25]。2007 年，中国人民公安大学教授梅建明，做了一些反恐数据的挖掘工作。目前，钦州市公安局应用大数据技术对 DNA 信息进行关联分析，取得了初步成效。虽然各种大数据预测方法层出不穷，但在实际应用中仍没有高效将全部赢无数据用于服务实战[26]，各级公安机关中的数据信息仍处于摸索尝试阶段。武警警务大数据也面对着和公安警务大数据同样的一些问题，有些甚至比公安警务大数据更加棘手，有很多国内顶尖的大数据相关学者很少将精力放到研究武警警务大数据上[27]，一是由于保密因素，武警警务数据都是密级数据，不能在互联网上公开，更别说合理利用了；二是武警部队对武警警务数据的研究分析还不够重视，没有调集人员将基础的信息集采工作流程理顺，导致很多数据都需要手工录入和人工处理，这也是武警警务数据一直未被开发利用的重要原因；三是武警警务数据没有统一的管理系统，且数据分布散乱，造成了大数据技术的部队人员无法接触到武警警务大数据，能接触到武警警务大数据的部队人员大部分又缺乏相关的工程技术能力[28]。虽然我国在机器学习方向也取得了一定的成绩，但大多数研究集中在数据挖掘层面，很少有将研究成果应用于军事领域。部队得数据库建设发展一直没有得到足够重视，一些对国防和军队具有价值的理论和设想甚至经常被忽略，除此之外，其余有的学科则聚集了大量的人力、物力和财力[29]。这种形式使我国在部队数据应用方面处于落后的局面。

随着社会的高速发展，每天产生大量的数据，为了提高数据的利用率，挖掘

数据潜在的价值，如今各大高校都建立了大数据相关专业。同时，南部战区、中部战区、陆军机关都在着手建立大数据中心，同时规范数据录入的方式和手段，加大基层部队数据收集基础设施建设[30]。通过这两方面着手，相信不久的将来，会有一批好的部队大数据成果展示出来。

1.2.2 国外研究现状

国外最早将数据分析方法应用于警务数据的例子出现在上世纪 90 年代中期的美国，2013 年美国面世了一种叫做 PredPol 的警情预测系统[31]，用于指导和帮助民警开展各项工作，预测系统是由一群数据科学专家、社会心理学家、政府官员和刑侦人员组成的团队，耗时 5 年研发而成的，该系统主要是把收集的数据进行分析，主要是对某个地区在未来一段时间可能发生犯罪事件进行预测[32]。截止到 2015 年有约 60 多个美国警局在使用这一系统，在一些警力不足的地区，该系统对于缓解警员压力还是十分有效的[33]，能在一定程度上帮助人员合理调配，制止犯罪事件发生，提高警察执行力。但随着警情预测系统的发展，也暴露出一些问题。一名名叫蒂姆·伯奇的奥克兰市警局调研部门负责人，用 4 个多月的时间对系统作了更深入的研究[34]。发现该系统在分析犯罪人员数据时，得出的结论相对来说少数族裔的人数较多[35]，伯奇认为，如果频繁根据系统数据来制订巡逻计划，本身就是一种种族歧视，会造成少数族裔群体的分布更加不均[36]。另外，除了涉嫌种族歧视，使用警情预测系统还可能会导致犯罪率不降反升。警情预测系统只是一个单纯基于历史数据的算法，因此它所能预测的犯罪行为也都是已知的，很难帮助警方深入掌握某些未知的犯罪模式[37]。更糟糕的是，这种预测算法可能会多次将警察带回已经处理过犯罪嫌疑人的同一地点，造成警力浪费甚至警民关系恶化。

曾经，美国洛杉矶警方将犯罪数据用于未来对地震区域余震预测的系统模型中，结果发现了某一区域未来的犯罪概率这一现象，于是警局利用这一结论，合理配置警力，大大降低了犯罪率[38]。2010 年美国警方还曾在某个专门从事跨境业务的金融公司植入一种监控程序，用于监测交易量和交易行为异常的情况，对

其进行实时报警，有效预防了诈骗等犯罪活动地发生。国外的这些研究和应用案例体现出现在我们公安系统提出的“主动警务”、“预测警务”的概念，同时也更加清晰地向我们展示了利用数据进行分析时存在的一些实际问题。

1.3 研究内容介绍

当前，随着新时代科技的迅速发展以及大数据行业的崛起，武警警务系统信息化改革也刻不容缓，在海量的武警数据中发现潜在信息，综合各种算法全方位运用数据，发现隐藏规律。同时，由于涉密、未公开等因素的影响，武警警务数据一直没有很好地被发掘利用，以往对于潜在价值较高的信息，诸如武警各类临时勤务的出动时间、地点、事件起因、规模等数据，更多的是根据经验和传统的数据统计方法，但是这种方法只适用于小范围的数据，海量的数据会耗费更多的时间，结果与实际相差较远且效率低下。随着大数据和机器学习的快速发展，为海量数据的分析提供了可能。本文根据武警警务数据的特点选择合适的实验平台，首先介绍了数据的来源，随后对数据进行初步处理、清洗，使其适合建模，再利用机器学习的算法模型对数据进行分析，得出一些有价值的判断和预测结果，本课题具体研究内容如下：

（1）武警警务数据获取渠道多样，结构各异，属性不一，涵盖了交通、商业、银行、安防、铁路等各个领域。这样就存在差别大、碎片化和不完整的数据并含有很多无用的脏数据。所以本文先对搜集的数据进行预处理，经过删除、概化、整理并按照设计好的标准化数据储存模型进行储存，然后预测未来一段时间内发生特定事件的概率，进而高效准确发现隐藏信息，提高数据资源利用率。

（2）基于武警警务数据特点，构建数据预测模型。将机器学习的方法应用于武警警务数据分析，使用 LSTM 神经网络模型，对处理后的数据挖掘深层特征。把数据划分为训练集和测试集，训练集作为 LSTM 神经网络模型的输入训练模型，在此过程中，根据数据特点选取合适的评价指标，并用 Adam 算法进行优化。本文的目的是运用警务数据训练 LSTM 神经网络模型，预测未来一段时间内，城市发生群体性闹事、暴力恐怖、个人极端等事件的数量。能够合理利用

资源，降低犯罪率。

1.4 论文组织结构和安排

本文主要对基于机器学习的武警警务数据分析模型进行研究，全文共分为四个章节，每个章节的内容安排如下：

第一章为绪论部分。主要介绍武警警务数据的发展和背景、近几年国内外的研究的情况以及本文的组织和安排。

第二章介绍武警警务数据的预处理和特征值的提取方法。在大量数据中，寻找合适的建模参数，能够建立精确的预测模型。面对海量的原始数据，需要先处理数据中的缺失值，然后进行数据的概化整理以及用归一化算法处理不同的属性，最后根据原始数据的特征，删除不必要的属性列。

第三章先对基于机器学习的数据预测模型中涉及到的相关技术进行详细地说明，并分析 LSTM 模型的相关技术和在本文中实现过程。接着介绍程序的整体框架和实验环境，简要说明了模型的设计需求，并给出实验流程。同时介绍数据预测模型的构建过程，用 MAE 作为评价指标。预测系统部分数据进行导入，通过这些数据产生模型，最终通过导入当年数据，预测未来发生群体性闹事、暴力恐怖、个人极端等事件的数量，并将结果以图表显现出来，实现数据情报预测模型的建立。

第四章为总结与展望。主要对前 3 章所做工作的梳理和概括。对研究的不足之处进行分析，并提出了模型优化的现实途径和几种新的模型应用场景，为今后进一步利用武警警务数据提供了借鉴思路。

第二章 数据处理

2.1 武警警务数据来源

本文主要是对武警警务数据进行分析，原始数据有的是从湖南省各大警务信息系统中授权获得，有的是在武警湖南总队担负各类任务的基层单位收集得到。如个人极端事件的数据是从担负省委省政府、市委市政府警卫勤务的武警中队和地方公安机关获取；群体性事件的数据是从长沙支队执勤十二中队、省委省政府执勤中队、长沙支队执勤十五中队等单位获取；恶性暴力事件数据是从湖南省部分县、市两级看守所获取。本文只从中选取了文本和表格数据，并把这些收集到的数据制成 Excel 表格。武警警务数据数据类型繁多，数据结构复杂，数据种类多样，但是经过数据的清洗和归一化，根据这些数据预测未来一段时间内，某个城市发生犯罪事件的概率是可行的，本实验选取的武警警务数据中的事件总体上可分为三类：杀人事件、集体事件和集访事件（数据都是经过严格脱密的数据，不涉及任何敏感信息），该数据集中记录了从 2008 年到 2018 年近 10 年间，湖南省各个城市（长沙、株洲、湘潭、邵阳、岳阳、常德、益阳、娄底、怀化、张家界等）的治安案件和群体性事件数据 325440 条，其中包括时间、地点、事件描述、人群规模、警力人数、持续时间、警力编成、案件人姓名、案件人职业、案件人民族、案件人特长、案件人年龄、案件人文化程度、案件人面貌、案件人籍贯、案件人单位、案件人罪名、案件人刑期和案件人前科这 19 列属性值，下面列出部分原始数据的截图，原始数据如表 2-1 所示。

在表 2-1 中，是把从湖南省各大警务信息系统数据网获取的数据制成的 Excel 表格，在这个说明表中，仅截取了原始数据的一部分属性，可以清楚的看到原始数据中含有许多缺失值，对于[案件人]和[案件人特长]这两个属性值，绝大部分的值都是缺失的，而[案件人职业]这个属性仅有少部分值是缺失的。还有[警力编成]、[案件人职业]和[案件人民族]这三个属性中有混合值，如[警力编成]有的是

由武警、公安和解放军这三个值参杂在一起，[案件人职业]则由退役军人和军人两个组合项，这些问题都对数据的读取产生一定的干扰性，针对以上许多不规范的信息，为了获取更加有效的数据，需要对原始数据进行预处理。

表 2-1 原始数据说明表

编号	时间	地点	事件描述	人群规模	警力人数	持续时间	警力编成	案件人姓名	案件人职业	案件人民族	案件人特长
1	2017/5/15	衡阳市太阳厂	群体性事件	81	105	1	武警、公安		工人		
2	2017/12/23	娄底市新化县	持枪杀人事件	1	5	1	武警	李一恒	个体	汉族	
3	2014/7/17	常德市	抗洪抢险	303	303	10	武警、公安、解放军		军人、退役军人		
4	2014/7/15	益阳安化	抗洪抢险	340	340	14	武警、公安、解放军		军人	维吾尔族、汉族	
5	2017/7/4	益阳郝山区	抗洪抢险	337	337	14	武警、公安、解放军			汉族	
6	2017/7/2	湘潭市雨湖区	抗洪抢险	450	450	15	武警、公安、		军人、退役军人	汉族	
7	2009/10/9	衡阳市太阳厂	涉毒涉黑事件	13	35	7	武警		无业		拳击
8	2008/1/5	全湖南省	抗险救灾	1050	1050	7	武警、公安、解放军			汉族	

2.2 武警警务数据预处理

如前文所述，实际工作中获取到的原始数据并不规范统一，常常伴随着缺失值、重复值、单一值等情况。这些数据不能直接使用，需要对原始数据进行预处理[23]。数据预处理没有具体的标准，由于数据的属性和类型不同而采用不同的处理方法。数据预处理的常用流程为：去除唯一属性、处理缺失值、属性编码、数据标准化正则化、主成分分析。

在数据处理之前，一般要先对数据进行整体分析，对影响分析结果的属性进行预选取。首先选取的属性意思明确，数值唯一不存在二义性；接着可以选择那

些存在多个数值但是含义明确的数据以及可能与预测结果有所关联的字段。对原始数据初步选取后，需要进行下一步的操作，包括填充或删除数据中的缺失值，去除产生乱码的数据，对数据类型的整理以及不同属性之间数据的加权操作等。本文具体的数据预处理方式有以下几种。

2.2.1 处理缺失值

在训练一个模型之前需要对数据进行预处理，因为用模型解决问题的最终效果取决于数据的质量和数据中蕴含有用信息的数量。在训练样本数据的实际模型中，原始数据可能会由于某些原因，存在一个或多个值的缺失。数据产生空缺值的原因有很多，也许是没有录入，也许是数据保存时产生了乱码，人为删除时导致的，也许本来数据就是缺失的。在表格中缺失值通常是以空值的形式或者是NA(Not A Number)存在的。如果直接忽视这些缺失值可能会造成算法无法处理这些空白项，触发异常情况。如果将包含缺失值的数据简单地全部删除也会造成数据的浪费，而且有时候数据量相对较少，删除包含缺失值的数据之后数据就更少，这将会大大地降低训练出模型的泛化能力。对数据的缺失值处理有以下三种方法：1.对于数据较多的情况，可以视情做一些删除处理。2.在不影响分析结果的基础上，可以适当地对缺失值进行填充。3.如果一行数据有多个缺失值，可以根据情况，对某些属性值进行删除，对另一些属性值进行填充。

比如下图 2-1 是一份罪犯的档案资料，可以看出，这份档案资料缺失罪犯的特长、别化名、累惯犯等信息，当这份数据录入系统中时，将缺失[特长]、[别化名]和[累惯犯]等属性值，如下表 2-2 所示。

如果这列属性存在的值大部分都是一样的，可以采用众数法，用出现最多的值填补空白值；如果这列属性中的值较为集中，可以采用均值法，把平均数当做空白值；中位数法也可以填补缺失值，相对于其他方法此方法适用于属性值特别分散的情况，在此种情况下，如果采用平均法填充会偏离实际值较多，进而产生较大的误差。针对本文中数据的空白值，根据数据特点，本文运用了以下 3 种缺失值处理方法。

图 2-1 罪犯原始档案图

表 2-2 录入罪犯数据信息

案件人员姓名	案件人员职业	案件人员民族	案件人员特长	案件人员年龄	案件人员文化程度	案件人员政治面貌
李孟军	农民	汉		28	小学	群众

（1）删除法

删除法是一种处理缺失值常用的方法，对属性过多的数据删除有很好的效果。在本文处理的海量原始数据中，如果某个属性存在大量的缺失值，可以直接删去此列属性。本文中原始数据获取的渠道不一，数据源多样（有从省委省政府警卫中队获取的数据，也有从武警 XX 总队机动支队获取的数据，还有从各州市公安局获取的数据），且结构各异（有 word、excel、pdf 等格式的数据），造成整体数据的不规范性。特别是如果大量数据都缺失某个属性值，就需要对包含缺失值多的属性采取列删除操作。表 2-3 为部分原始数据属性值 1。

下表 2-3 是原始数据的一部分属性值，列属性 [案件人员姓名]和[案件人员特长]两项均存在大量空白值，由于这两个属性的数据值缺失过多，并不会对整

体的预测结果产生影响，因此采用删除法去除这两列的属性。在图中，第二行缺失了[案件人员姓名]、[案件人员民族]和[案件人员特长]这三个属性所对应的值，但是以列来看，[案件人员职业]和[案件人员民族]这两个属性仅缺失部分属性值，初步判断这两个属性可能会对实验结果产生一定的影响，所以不能直接删除这两个属性。采用删除法处理数据后的结果如下表 2-4 所示：

表 2-3 原始数据属性值样本

编号	持续时间	警力编成	案件人员姓名	案件人员职业	案件人员民族	案件人员特长
1	1	武警、公安		工人		
2	1	武警	李一恒		汉族	
3	10	武警、公安、解放军		军人	汉族	
4	14	武警、公安、解放军			汉族	
5	14	武警、公安		军人		
6	15	武警		军人	汉族	
7	7	武警、公安			汉族	拳击
8	7	武警、公安、解放军、地方人员		军人	汉族	
9	15	武警、公安、解放军、地方人员		军人	汉族	

表 2-4 删除数据后的属性表

编号	持续时间	警力编成	案件人员职业	案件人员民族
1	1	武警、公安	工人	
2	1	武警		汉族
3	10	武警、公安、解放军	军人	汉族
4	14	武警、公安、解放军		汉族
5	14	武警、公安	军人	
6	15	武警	军人	汉族
7	7	武警、公安		汉族
8	7	武警、公安、解放军、地方人员	军人	汉族
9	15	武警、公安、解放军、地方人员	军人	汉族

（2）填充法

虽然删除法非常方便，在本文处理的数据中，时常会出现某列属性存在少量缺失值的情况，且该缺失值与预测结果具有很强的关联性，有时不仅要保留这个空缺值，还要选择一些方法对空缺值进行填充。表 2-4 演示的，是数据缺失较多且对预测结果影响不大的属性值进行删除操作，如果数据中仅存在部分的缺失值，

为了使这些缺失值对预测结果产生较小的影响，需要对这些缺失数据进行填充。如表 2-5 为部分数据的属性值 2。

如下表 2-5 所示，[案件人员职业]、[案件人员民族]、[案件人员年龄]和[案件人员政治面貌]这三个属性中都存在着少量的缺失值，所以可以使用填充法对缺少的值进行处理。本文的填充方法有众数法和平均法两种：

表 2-5 原始数据属性值 2

编号	案件人员职业	案件人员民族	案件人员年龄	案件人员文化程度	案件人员政治面貌
1	工人		20-40	小学、初中	群众、团员
2		汉族	26	初中	群众、团员
3	军人		18-42	初中	党员、团员
4	军人	汉族、维吾尔族	18-43	初中、高中、大学、硕士	党员、团员
5		汉族	18-41	初中、高中、大学、硕士	党员、团员
6	军人	汉族	17-39	初中、高中、大学、硕士	
7			15-46	初中	党员、团员
8	军人	汉族		初中、高中、大学、硕士	党员、团员
9	军人	汉族	20-45	初中、高中、大学、硕士	党员、团员
10	无业	汉族	23、30	小学、初中	
11	无业	汉族	17-19		党员
12	无业	汉族	15	初中	

众数法填充就是选择同一个类型的属性值中出现次数最多的那一个值作为关健值，用这一个值填充其余的缺失值。如上表 2-5 中，[案件人员职业]、[案件人员政治面貌]和[案件人员民族]这三个属性有少量缺失值，在[案件人员职业]这一属性中，出现次数最多的属性值是军人，所以可以对空白的地方填充军人；而在[案件人员政治面貌]这个属性中，出现次数最多的属性值是党员、团员，所以可以对空白的地方填充党员、团员这两个值；[案件人员民族]中，出现次数最多的是汉族，所以可以在空白值中填充汉族。经过众数法填充后的结果如下表 2-6 所示。

与此同时，从表 2-6 可知，[案件人年龄]这个属性中也存在着少量的缺失值，

此时应对缺失值进行填充。可是这个属性又和[案件人员职业]和[案件人员民族]有所区别。因为这一列属性中存在着年龄的范围，如第一行显示的年龄是 20-40 岁，并没有给定具体的数值，所以如果不处理此种类型的属性，可能对武警警务数据的预测过程产生干扰，继而影响预测结果的准确性。同样，在处理的过程中，也不可能通过众数法填充缺失值，这样会存在着明显的偏差，因为你并不知道在这个年龄范围内众数是多少，如第三行案件人的年龄是 18-42 岁，但是并不知道 18 岁这个年龄出现了几次。

针对此种情况，本文使用平均法对这种类型的属性进行填充。平均法是通过特征数据的平均指标，反映事物目前所处的位置和发展水平。再对不同时期、不同类型单位的平均指标进行对比，说明事物的发展趋势和变化规律。具体运用到本文中，首先对只给定年龄范围的属性值，算出平均值。如对 20-40 岁这个范围的年龄值，通过平均法计算出的平均年龄是(20+40)/2=30，那么就把 30 岁作为这个范围的年龄对应的取值，据此把各个年龄范围的值一一确定。在表的第 8 行中[案件人员年龄]这一项为空，需要先对此属性列进行统计，如果此列中存在的缺失值较多且大多数人员的年龄都是相同的值，那么就可以采用众数法填充空白值，否则运用平均法填充空白。当每个范围的属性值都有确定的值时，把[案件人员年龄]这一列求和，就可以再次使用平均法计算整体年龄的平均值，并用这个值填充存在的缺失值。经过平均法填充后的结果如表 2-7 所示：

（3）删填结合法

前文介绍了删除法和填充法处理缺失值，但本文中原始数据获取的渠道不一，数据源多样（有从省委省政府警卫中队获取的数据，也有从武警 XX 总队机动支队获取的数据，还有从各州市公安局获取的数据），且结构各异（有 word、excel、pdf 等格式的数据）导致文中一部分数据并不能用单一的删除法或填充法来处理。所以本文中还用了删除法和填补法相结合的方式对数据的缺失值进行处理。具体的做法是对于原始数据中存在大量的缺失值，如果某些属性只存在少量数值则可

表 2-6 众数填充后的数据表

编号	案件人员职业	案件人员民族	案件人员年龄	案件人员文化程度	案件人员政治面貌
1	工人	汉族	20-40	小学、初中	群众、团员
2	军人	汉族	26	初中	群众、团员
3	军人	汉族	18-42	初中	党员、团员
4	军人	汉族、维吾尔族	18-43	初中、高中、大学、硕士	党员、团员
5	军人	汉族	18-41	初中、高中、大学、硕士	党员、团员
6	军人	汉族	17-39	初中、高中、大学、硕士	党员、团员
7	军人	汉族	15-46	初中	党员、团员
8	军人	汉族		初中、高中、大学、硕士	党员、团员
9	军人	汉族	20-45	初中、高中、大学、硕士	党员、团员
10	无业	汉族	23、30	小学、初中	党员、团员
11	无业	汉族	17-19		党员
12	无业	汉族	15	初中	党员、团员

表 2-7 平均法填充数据表

编号	案件人员职业	案件人员民族	案件人员年龄	案件人员文化程度	案件人员政治面貌
1	工人	汉族	30	小学、初中	群众、团员
2	军人	汉族	26	初中	群众、团员
3	军人	汉族	30	初中	党员、团员
4	军人	汉族、维吾尔族	31	初中、高中、大学、硕士	党员、团员
5	军人	汉族	30	初中、高中、大学、硕士	党员、团员
6	军人	汉族	28	初中、高中、大学、硕士	党员、团员
7	军人	汉族	31	初中	党员、团员
8	军人	汉族	28	初中、高中、大学、硕士	党员、团员
9	军人	汉族	33	初中、高中、大学、硕士	党员、团员
10	无业	汉族	27	小学、初中	党员、团员
11	无业	汉族	18	小学、初中	党员
12	无业	汉族	15	初中	党员、团员

采用删除法，因为这些属性值太少对预测结果没有影响。而那些存在大量属性值

且与预测结果具有关联性的属性可直接采用填补法。下面列出了包含缺失值的部分原始数据属性值 3，数据格式如表 2-8 所示。

在表 2-8 中，由于原始数据的属性过多，全部展现出来并不现实。所以只列出了[地点]、[持续时间]、[警力编成]、[案件人员姓名]、[案件人员职业]、[案件人员民族]、[案件人员特长]和[案件人员年龄]这几个属性。从表 2-7 可知，部分属性存在着缺失值，[案件人员姓名]这个属性和[案件人员职业]、[案件人员民族]和[案件人员年龄]这三个属性通过比较后发现，[案件人员姓名]存在着大量的缺失值，在列出的 9 行数据中，仅仅第 2 行存在这个属性值，经过总体的统计与分析，发现这个属性的空缺值占这列数据量总和的 95%，由于缺失值过多，这个属性的存在与否对整个预测结果产生的影响微乎其微，可以忽略不计，经此分析，可以采用删除法删除这列属性。[案件人员职业]这个属性，仅从表中列出的部分数据可以知道，它虽然也有缺失值，但只有第 2 行、第 4 行和第 7 行存在值的缺失，在整体的统计中也只有少部分有缺失，那这行属性可能会对预测的结果产生一定的影响，因此这列属性就可以采取填充法。以第 1 行来说，这行数据存在三个缺失值，对应的属性分别是[案件人员姓名]、[案件人员民族]和[案件人员特长]，从列来看，[案件人员姓名]和[案件人员特长]这两个属性中存在大量缺失值，可以采用删除法；而[案件人员民族]这个属性中存在着少量缺失值，采用删除法肯定是不合适的，因此可以运用填充法对缺失值进行填补。针对这行数据而言，可以根据实际情况采用删填结合的方式处理武警警务数据中的缺失值。再如第 2 行，存在两个缺失值，对应的属性分别是[案件人员特长]和[案件人员职业]，从列项可知，[案件人员姓名]存在大量缺失值，虽然第二行这个属性有确定的值，但是整体上[案件人员姓名]存在大量缺失值。从上分析可知，在整体统计后，[案件人员民族]和[案件人员年龄]这两个属性有少量缺失值，可以运用填充法填补数据，[案件人员特长]有大量的缺失值，可以直接删去这一属性。运用删填结合方式处理后的数据效果图如下表 2-9 所示：

表 2-8 原始数据属性值 3

编号	地点	持续时间	警力编成	案件人员姓名	案件人员职业	案件人员民族	案件人员特长	案件人员年龄
1	衡阳市太阳厂	1	武警、公安		工人			20-40
2	娄底市新化县	1	武警	李一恒		汉族		26
3	常德市	10	武警、公安、解放军		军人	汉族		18-42
4	益阳安化	14	武警、公安、解放军			汉族		18-43
5	益阳郝山区	14	武警、公安		军人			18-41
6	湘潭市雨湖区	15	武警		军人	汉族		17-39
7	衡阳市太阳厂	7	武警、公安			汉族	拳击	15-46
8	全湖南省	7	武警、公安、解放军、地方人员		军人	汉族		
9	岳阳岳阳楼区	15	武警、公安、解放军、地方人员		军人	汉族		20-45

表 2-9 删填结合法处理数据表

编号	地点	持续时间	警力编成	案件人员职业	案件人员民族	案件人员年龄
1	衡阳市太阳厂	1	武警、公安	工人	汉族	30
2	娄底市新化县	1	武警	军人	汉族	26
3	常德市	10	武警、公安、解放军	军人	汉族	30
4	益阳安化	14	武警、公安、解放军	军人	汉族	31
5	益阳郝山区	14	武警、公安	军人	汉族	30
6	湘潭市雨湖区	15	武警	军人	汉族	28
7	衡阳市太阳厂	7	武警、公安	军人	汉族	31
8	全湖南省	7	武警、公安、解放军、地方人员	军人	汉族	30
9	岳阳岳阳楼区	15	武警、公安、解放军、地方人员	军人	汉族	33

原始数据通过以上的处理方式后的主要属性，如下表 2-10 所示。

表 2-10 实验数据说明表

武警事件相关属性列表	时间	相关事件人员信息	职业
	地点		
	事件描述		民族
	人群规模		
	警力人数		年龄
	持续时间		
	警力编成		

在表 2-9 中，由于原始数据中属性值过多，很多数据出现大量缺失值，而并没有对数据进行离散化，而是根据实际情况选取主要属性，经过筛选后，得到的主要的属性值如上表 2-10 所示。由于某些事件参与人数过多，并不能很详细的描述相关事件的人员信息，所以一般会选取一个特定范围来描述。

2.2.2 数据的概化

数据的概化就是将具体数据抽象化，以一种更加模糊的形式表现出来[37]。这样可以把原先具体的数据上升到另一个层面进行大致的概括，可以更加直观的了解整体。在本文中，数据的划分与处理有一定的难度，如果划分的太过于细致，就会失去数据的分类依据，所以需要将一些数据进行更高层次的泛化，即数据的概化，以此来达到实验分类的目的。本次收集的数据中[事件描述]这一属性包含恶性杀人事件、群体性事件、个人极端杀人事件、抢险救援、涉毒涉黑案件、法警集访、涉黑事件，就可以应用数据概化的方式，将它们总体分为杀人事件、群体性事件、集访事件，这样可以提升数据的一致性，便于后期提取有关方面的特征进行预测。但是计算机在读取数据的过程中，并不能直接处理输入的中文，所以为了更加具有可操作性，本文把杀人事件命名为 a,群体性事件命名为 b,集访事件命名为 c，这样就可以达到数据概化的目的，并且增强实际可操作性，经过概化处理后的结果如下表 2-11 所示：

表 2-11 概化方法处理数据表

编码	时间	地点	事件描述	人群规模	警力人数	持续时间（天）
1	2017/5/15	衡阳市太阳广场	a	81	105	1
2	2017/12/23	娄底市新化县	b	1	5	1
3	2014/7/17	常德市	c	303	303	10
4	2014/7/15	益阳安化	c	340	340	14
5	2017/7/4	益阳郝山区	c	337	337	14
6	2017/7/2	湘潭市雨湖区	c	450	450	15
7	2009/10/9	衡阳市太阳广场	c	13	35	7
8	2008/1/5	全湖南省	c	1050	1050	7
9	2015/6/2	岳阳岳阳楼区	C	570	570	15
10	2017/11/29	衡阳市耒阳	b	2	28	1
11	2012/2/28	衡阳市耒阳	b	3	36	1

2.2.3 数据的整理

对原始数据整理后，数据间的相互对比可以反映出数据的特征，比如前几年事件的数量和今年事件的数量对照，可以发现事件的发展趋势。从细节方面来看，可以观察到每个城市发生的事件数量是否与时间成正相关或负相关。通过将收集到的数据分析和对比后，往往能预知事情的大概走向，沿着这个确定的方向，预

能避免因选择数据方法失误，造成不理想的预测结果。

下表 2-12 是本文原始的部分采集维度，包含［案件人员姓名］,［案件人员职业］,［案件人员民族］,［案件人员特长］,［安检人员年龄］,［案件人员文化程度］,［案件人员政治面貌］等 7 个维度。在表中可以看到[案件人员特长]和[案件人员姓名]这两个维度存缺失值，对比原始数据表可知，这两个维度中存在的值占总体数据的几十万分之一，由此这两列属性应该删除，否则影响整体的预测结果。而[案件人员职业]和[案件人员民族]虽然也存在这缺失值，但是对整体而言，这两列的缺失值占的比例较小，不能直接删除，可采用填充法补足空白值。

表 2-12 原始数据采集维度

案件人员姓名	案件人员职业	案件人员民族	案件人员特长	案件人员年龄	案件人员文化程度	案件人员政治面貌
	工人			20-40	小学、初中	群众、团员
李一恒		汉族		26	初中	群众、团员
	军人			18-42	初中	党员、团员
	军人	汉族、维吾尔族		18-43	初中、高中、大学、硕士	党员、团员
		汉族		18-41	初中、高中、大学、硕士	党员、团员
	军人	汉族		17-39	初中、高中、大学、硕士	
				15-46	初中	党员、团员
	军人	汉族			初中、高中、大学、硕士	党员、团员
	军人	汉族		20-45	初中、高中、大学、硕士	党员、团员
	无业	汉族		23、30	小学、初中	
	无业	汉族		17-19		党员
	无业	汉族		15	初中	

本次实验中，采集的武警警务数据时间跨度从 2008 年到 2018 年，因此本文的预测模型是对未来某天或某个月本文涉及的每个城市的事件数量进行预测，帮助武警情报部门把握社会面整体趋势，提高行动知道效能。在训练模型时，需要将采集的数据进行数据分割，划分为测试集和训练集，数据的分割方式一般有两种：

1.随着科技的进步，近年来收集的武警警务数据更加的全面，大部分数据主要集中在 2015 年到 2018 年 ，所以将 2008 到 2016 年的数据进行合并作为训练集，2017 年和 2018 年的数据作为预测部分，这样划分既保证了训练集数据量的充足可靠，又保证了 2017 年和 2018 年的数据预测对武警情报的灵敏性。如果只选择 2008 年到 2015 年的数据作为训练集，由于本文中收集的武警警务数据更多的集中在 2016-2018 年，那么训练集数据量会有所不足，不能很好的训练出预测模型，所以才要根据原始数据的特点进行上述划分。除此之外这种划分方式也方便后期插入更多数据。效果如表 2-13 所示：

表 2-13 数据采集处理

时间	地点	事件描述	人群规模	警力人数	持续时间（天）	警力编成	案件人员姓名	案件人员职业
2017/12/23	娄底市新化县	持枪杀人事件	1	5	1	武警		个体
2017/12/17	湖南娄底	故意杀人					陈建湘	
2017/11/29	衡阳市耒阳	持枪暴力事件	2	28	1	武警、公安		无业
2017/11/28	湖南怀化	故意杀人						
2017/7/4	益阳郝山区	抗洪抢险	337	337	14	武警、公安、解放军		军人
2017/7/2	湘潭市雨湖区	抗洪抢险	450	450	15	武警、公安、解放军		军人
2017/5/30	长沙市	群体性事件						
2017/5/15	衡阳市太阳广场	群体性事件	81	105	1	武警、公安		军人

2.在训练样本时，要保证数据的充足性，只有在大量数据的基础上训练出的模型才更具有说明说服力。但是训练的数据不可能包含所有的样例，如果预测的结果非常完美的涵盖了训练集的所有点，这说明训练出的模型很可能处于过拟合状态，会导致结果产生偏差。因此需要对处理过的武警警务数据进行划分。一般来说，会使用留出法划分数据。即先把数据集划分成互不相交的两部分，一部分做训练集，一部分做测试集，类似分层抽样一样，需要保持数据大概一致，常规情况下，训练集数据需要占到总体数据的 70%，这样才能保证训练出模型的适用

性，为了保证抽取结果更具有随机性，需要对数据进行多次随机划分，然后对多次划分的结果求平均值。

2.2.4 数据归一化

目前数据的归一化方法有很多种，不同的方法对模型的评价结果也有所不同。在原始数据中会存在不同的数据单位，各个属性之间无法直接运算，为了转化成无量纲数值，需要归一化处理数据。

多指标评价体系中，通常具有不同的量纲和数量级。当各指标间的水平相差很大时，相互之间直接进行比较和运算，就会产生误差。因此，为了保证结果的可靠性，需要对原始指标数据进行归一化处理。数据的归一化处理，是把数据的结果都映射在[0,1]之间，这种方法不仅可以提升模型的收敛速度，还可以提升模型的精度。

（1）连续型数据

本文采用 min-max 标准化（Min-Max Normalization）算法对连续型数据进行归一化处理，是对原始数据的线性变换，使结果值映射到[0 - 1]之间[39]。转换函数如下：

$$x^* = \frac{x-min}{max-min} \qquad (2\text{-}1)$$

其中 x 为当前值，max 是最大值，min 是最小值。

如表 2-14 所示：

表 2-14 原始数据属性值 4

编码	时间	地点	事件描述	人群规模	警力人数	持续时间（天）
1	2017/5/15	衡阳市太阳广场	群体性事件	81	105	1
2	2017/12/23	娄底市新化县	持枪杀人事件	1	5	1
3	2014/7/17	常德市	抗洪抢险	303	303	10
4	2014/7/15	益阳安化	抗洪抢险	340	340	14
5	2017/7/4	益阳郝山区	抗洪抢险	337	337	14
6	2017/7/2	湘潭市雨湖区	抗洪抢险	450	450	15
7	2009/10/9	衡阳市太阳广场	涉毒涉黑事件	13	35	7
8	2008/1/5	全湖南省	抢险救灾	1050	1050	7
9	2015/6/2	岳阳岳阳楼区	抢险救灾	570	570	15

由上表仅列出的部分数据可知，在[人群规模]这列属性中，最大值为 1050，最小值为 1，在[警力人数]这列中最大值是 1050，最小值是 5，,在[持续时间]这个属性中，最大值是 15，最小值是 1，由此可得，这三个属性中的值都太过离散，特别是[人群规模] 和[警力人数]这两列的属性值之间差距过大，如果直接把这些值作为变量输入到模型中，预测的值离散程度不仅很大，而且会存在着严重的误差，为了避免这种情况，需要降低数据的离散程度，用归一化的方法处理数据，本文以[持续时间]这一列数据为例，进行归一化处理，结果如表 2-15 所示：

表 2-15 min-max 标准化

持续时间（天）	1	1	10	14	14	15	7	13	15
min-max 标准化	0	0	0.643	0.929	0.929	1	0.429	0.857	1

从上表中，可以观察到经过 min-max 标准化处理后，结果都映射到[0 - 1]之间，缩小了数据之间的离散程度，降低了过离散化数据对武警警务数据预测结果的干扰，从而达到预测结果上下浮动较小，预测精确的目的。在表 2-15 中，[人群规模]、[警力人数]和[持续时间]，这三者的单位并不一致，按照理论来说，三者之间是没有任何可比性的，自然也就不能进行统一运算，但是这三个属性经过归一化处理后，将其转化为无量纲的纯数值，那么这三个属性值之间就可以进行比较和加权运算。

（2）离散型数据

在机器学习任务中，大多数数据特征是离散值。为了使不同的数据类型之间都能进行比较和加权，可以采用独热编码（One-Hot Encoding）[40]。在本文中，获取的原始数据包括湖南省的各个地区，由于地区过多，不能一一列举，文中选取湖南省的 7 个城市，使用 one-hot 编码后，结果如下表 2-16 所示：

上图是对湖南省中选取的 7 个城市为例，使用 one-hot 编码，但是每个城市还包含很多县和区。为了使地方特征更加鲜明，预测结果更加精确，还可以对城市进一步的划分，本文以衡阳市为例，进行更加细致的划分后，由于变量过多，并不适合用 one-hot 编码。

表 2-16 one-hot 编码

长沙	衡阳	益阳	株洲	湘潭	邵阳	岳阳
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	0	0	1	0	0	0
0	0	0	0	1	0	0
0	0	0	0	0	1	0
0	0	0	0	0	0	1

所以可以使用类似邮编的方式对县和区编码，结果如下表 2-17 所示：

表 2-17 区域编码

珠晖区	雁峰区	石鼓区	南岳区	蒸湘区	衡阳县	横山县	祁东县	衡南县
42001	42002	42003	42004	42005	42006	42007	42008	42009

通过对每个城市管辖下的的县和区进一步编码区分，能够更加明确地了解事件的发生地点，对武警及地方公安机关开展工作有着指导意义。也对相关人员调配及整体管控有着更清晰的认知。

2.3 武警警务数据特征提取

当前期的数据预处理过程结束后，就进入到数据处理的下一个阶段，数据特征的提取。操作的依据是数据是否发散，如果这列属性不发散，也就意味着这列属性都有一个共同的值，那么这列属性的存在与否对整体的预测结果并没有影响，就可以采用删除法删除这列属性。比如［警力编成］这个数据特征中属性值包含武警和公安，因此[警力编成]这个属性是二维的，但是所有的数据统计中都有武警人员这个属性，即这个属性无法区别警力编成的身份，所以这个属性应该舍弃，如果继续使用，只是耗费运行时间，对预测结果也没有什么影响。由于数据是否发散是人为判别的，有时可能删除了本应存在的属性进而造成误差，所以应该进一步改进。本文采用特征提取数据的结果如下表 2-18 所示：

在特征提取过程中还要注意属性和警务数据预测模型是否具有相关性。如果两者的相关性很高，就需要优先处理或是填充缺失值。如果两者之间几乎没有相关性，那么此列属性的存在与否不会对也测结果产生任何不良影响，应该删除此列。

表 2-18 数据维度处理

时间	地点	事件描述	人群规模	警力人数	持续时间（天）	警力编成
2018/10/5	衡阳市衡东县	恶性杀人事件	1	4	1	武警
2018/9/20	衡阳市耒阳市	恶性杀人事件	1	5	1	武警
2018/9/15	湖南怀化	群体性事件	20		1	武警、公安
2018/9/13	衡阳市衡东县	个人极端驾车撞人事件	1	1	3	公安
2018/8/25	衡阳市耒阳市	群体性事件	40	80	1	武警、公安
2018/6/25	湖南湘潭	群体性事件	36	77	1	武警、公安
2018/4/13	湖南衡阳	个人极端杀人事件	1	5	5	武警
2017/12/23	娄底市新化县	持枪杀人事件	1	5	1	武警

如在表 2-18 中，[警力编成]这个属性和本文预测的结果毫无关联，因此本文经过数据预处理后产生类似的属性值，就可以直接丢弃。

2.4 小结

本章主要介绍武警警务数据的来源、数据的预处理和特征值的提取方法。本文收集的原始武警警务数据数据类型繁多，数据结构复杂，数据种类多样，为了保证预测结果的可靠性，需要对数据进行预处理和特征提取。在大量数据中，寻找合适的建模参数，能够建立精确的预测模型。在数据预处理中，主要包括数据的缺失值处理、数据的概化、数据的整理和数据的归一化。面对海量的原始数据，需要先处理数据中的缺失值，然后把相同数据概化整理以及用归一化算法处理不同的属性。在特征提取过程中需要判断属性和警务数据预测模型是否具有相关性。以此来进行数据处理，删除不必要的属性值或填充属性的缺失值。

第三章 武警警务数据预测模型的实现与测试

3.1 LSTM 模型

3.1.1 LSTM 模型的实现

在数据的处理工作准备完成后，就可以构建武警警务数据预测模型。在构建模型之前，需要搭建实验环境。实验环境如表 3-1 所示：

表 3-1 实验环境

操作系统	Windows10 操作系统
开发语言	Python 语言
开发平台	pycharm
深度学习库	Keras 神经网络库
模型	LSTM 神经网络

搭建好实验环境，就可以构造 LSTM 神经网络模型，构造模型步骤如下：

（1）首先划分训练集和测试集。这两者的划分可以依据原始数据特点，也可以随机划分，一般比例为 7:3。为了加速训练模型，本文根据数据的特征使用 2008-2016 年数据作为训练集训练模型，然后使用剩下 2 年得数据进行评估。模型的输入变量和输出变量分别由训练集和测试集提供，最终需要转换数据格式。

（2）LSTM 模型中，初始设置时，假设隐藏层有 50 个神经元，输出层 1 个神经元，输入变量是一个时间步（t-1）的特征，损失函数采用 Mean Absolute Error(MAE)，优化算法采用 Adam，初始化模型采用 128 个 epochs 并且每个 batch 的大小为 100。

（3）最后，在 fit()函数中设置参数，记录训练集和测试集的损失，并在完成训练和测试后绘制损失图。接下里就可以对模型效果进行评估。通过以上处理

之后，再结合 RMSE（均方根误差）来描述实验数据是否具有好的精确度。

在本文中，LSTM 神经网络模型的构建主要采用几个主要的函数方法，这些方法的代码及意思如下详解：

dataset = read_csv('data.csv', header=0, index_col=0); // 首先需要通过 read_csv()这个方法来加载数据集 data，并以文件的第一行为索引读取文件中的数据。

scaler = MinMaxScaler(feature_range=(0, 1)); //用 MinMaxScaler()离差标准化方法对输入的数据进行归一化处理，使其结果映射在[0,1]之间，具体解释详见 2.2.4 节的归一化处理。

reframed =series_to_supervised(scaled, 1, 1); //使用 series_to_supervised()方法将时间序列问题转换为监督学习问题，第一个参数表示输入的数据序列，第二个参数是输入的滞后步数，但第三个参数是输出的移动步数，默认为 1。

train = values[:n_train_hours, :]; //把数据分割为训练集。

train_X = train_X.reshape((train_X.shape[0], 1, train_X.shape[1]));//将数据集重构为符合 LSTM 要求的数据格式,即 [样本，时间步，特征]。

model.add(LSTM(50,input_shape=(train_X.shape[1],train_X.shape[2])));//设计 LSTM 神经网络，隐藏层神经元的个数为 50，并输入二阶张量。

history = model.fit(train_X, train_y, epochs=50, batch_size=72, validation_data=(test_X, test_y), verbose=2, shuffle=False); //训练 LSTM 神经网络模型，训练次数为 50 次，每个输入张量的批量大小为 72。

pyplot.plot(history.history['loss'], label='train') ; //绘制训练集的损失函数曲线图。

rmse = sqrt(mean_squared_error(inv_y, inv_yhat)); //使用均方根误差函数计算损失函数来衡量模型的精确度。详解见 3.2.3LSTM 损失函数。

3.1.2 LSTM 模型结构

Longshort term memory，是一种时间递归神经网络，是一种特殊的 RNNs，

能够学习长期依赖关系，适合于处理和预测时间序列中间隔和延迟相对较长的重要事件。在自然语言处理、语言识别等一系列的应用上都取得了很好的效果。其结构图如 3-1 所示

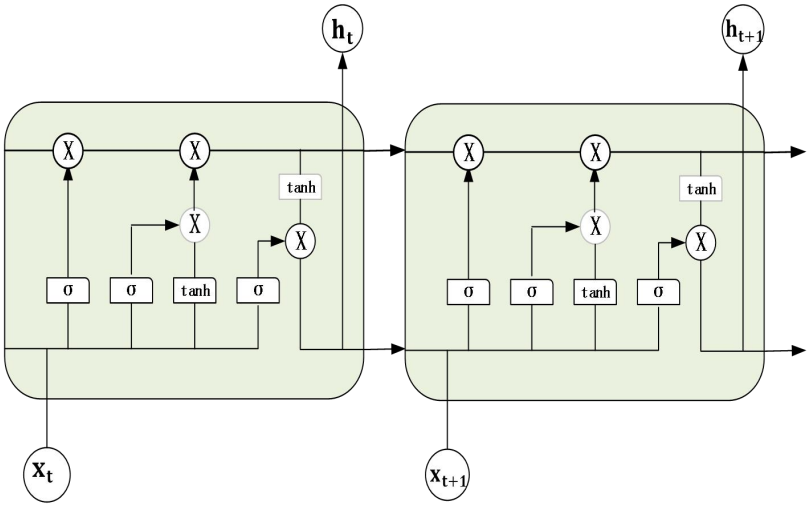

图 3-1 LSTM 结构图

从上图中可以看出，在向后传播的每个 t 时刻都和 RNN 一样有一个隐藏状态 h_{t-1}，LSTM 结构比 RNN 结构多了另一个隐藏状态，如下图中上面的长横线。这个隐藏状态称为细胞状态(Cell State)，记为C_t。如下图 3-2 所示：

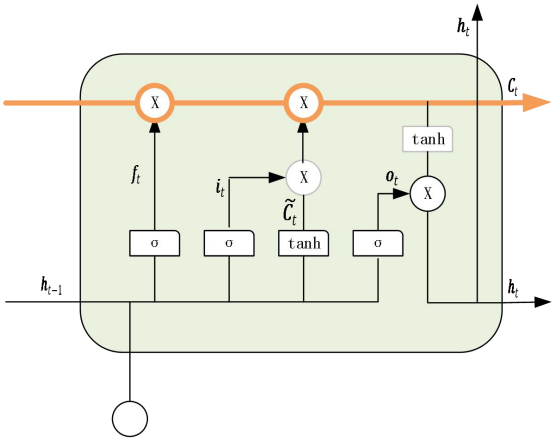

图 3-2 LSTM 的细胞状态图

除了细胞状态，LSTM 图中还有很多门控结构(Gate)。LSTM 在在每个序列索引位置 t 的门一般包括遗忘门，输入门和输出门三种。

（1）遗忘门

在 LSTM 神经网络中的第一步就是决定要对什么样的信息进行丢弃。这个决定的过程被称为在遗忘门层完成。首先当有信息输入时，这个门会读取本次需

要输入的值以及上一层传过来的值，进行处理后，会输出一个 0 到 1 之间的数字，1 表示“完全保留”，0 则表示“完全舍弃”。这主要是为了判断是否需要把这个值记下来，如果输出的值越大，表示上一次输出的值就越重要，应该保留，反之亦然。LSTM 神经网络的遗忘门状态图如 3-3 所示;

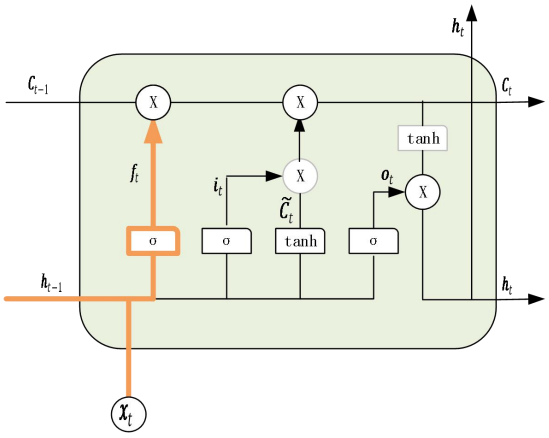

图 3-3 遗忘门状态图

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \qquad (3\text{-}1)$$

在公式 3-1 中，W_f是遗忘门的权重矩阵，b_f是遗忘门的偏移量，σ是激活函数 sigmoid，h_{t-1}是上一层神经元的输出，x_t是当前时间信息的输入。LSTM 层的工作方式是通过接收 3 维（N，W，F）的数字阵列，其中 N 是训练序列的数目，W 是序列长度，即时间步，F 是每个序列的特征数。本文是将处理后的数据集分成训练集和测试集。然后又可将训练集和测试集分别作为输入和输出变量。再用数据训练 LSTM 模型之前，需要做如下工作：

首先，需要设置每次神经元读取数据的大小，即 batchsize，假如初始化为 128，然后设置神经元需要训练的次数 epochs,初始化为 50，也就是说训练集训练的次数为 50 次；最后需要设置 LSTM 的三个指标：样本数为 235440 条数据，时间窗口为 1 天，即预测某个城市明天会发生某种特定类型事件的概率。特征 F 为 3，即[时间]、[地点]和[事件]这三个属性。初始时，假如说第一批数据经过第一层神经元后输出一个值，即h_{t-1}，然后进入第二层时，在当前时刻又来了一批新的样本作为输入变量，即x_t，这两个值通过激活函数 sigmoid，得到遗忘门的输出f_t。由于 sigmoid 的值一般在[0,1]之间，所以f_t是一个[0,1]之间的概率值，

表明遗忘上一层细胞的概率，如果为 0，则表示上一层输出的值h_{t-1}不需要被记忆。在本文中，试图根据以前各个城市发生的事件，预测未来某天某个城市发生特定事件的概率。实际上，如果一个地区已经连续多次发生了某种特定类型事件，那经过下一层神经元时计算出被遗忘的概率f_t就特别小。

（2）输入门

输入门的作用就是决定要输入多少个信息，这个过程需要两个具体的步骤：首先，sigmoid 函数层需要判断输入的哪些信息需要被更新；一个 tanh 层根据输入的信息生成一个向量，作为用来更新的内容。然后，需要将这两部分联合起来，对整个部分进行更新。LSTM 的输入门状态图如 3-4 所示：

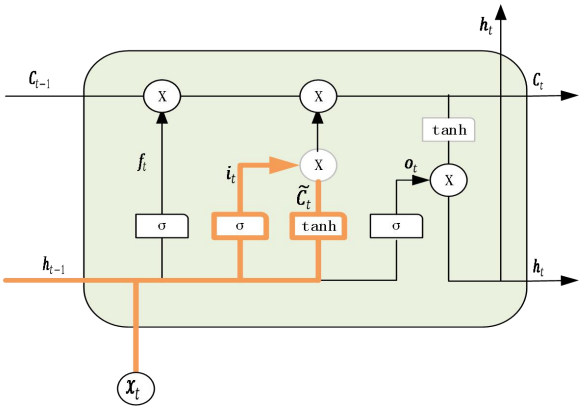

图 3-4 输入门状态图

$$i_t = \sigma(W_t \cdot [h_{t-1}, x_t] + b_i) \qquad (3\text{-}2)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \qquad (3\text{-}3)$$

从上图及公式可以看出，输入门由两部分组成，第一部分是运用激活函数 sigmoid 输出的值i_t，第二部分是通过激活函数 tanh 输出的值$\tilde{C}_t$，i_t主要是决定哪些值需要更新，$\tilde{C}_t$是生成新的后选值，通过 tanh 层生成新的候选值，可能会添加到新细胞中，可以把这两部分结合起来更新细胞状态。也就是说，在 t 时刻又输入了一个新的信息向量，通过激活函数 sigmoid 后看上一层输出的h_{t-1}和输入的向量x_t组合的张量中哪些值需要记忆，而这两个向量组合后经过 tanh 激活函数会生成一个新的后选值。在本文中，t 时刻时又输入了一批训练集，此时这个训练集需要和上一层输出的信息组合成一个新的向量，由于预测未来一天某个城市发生事件的概率，那假如这个城市是长沙，那通过 sigmoid 函数可以去掉不属

于长沙发生事件的信息。另一方面输入的x_t中，可能提取到长沙发生了事件这个消息，然后通过 tanh 函数把信息提取出来作为候选值。细胞更新状态图如下图 3-5 所示：

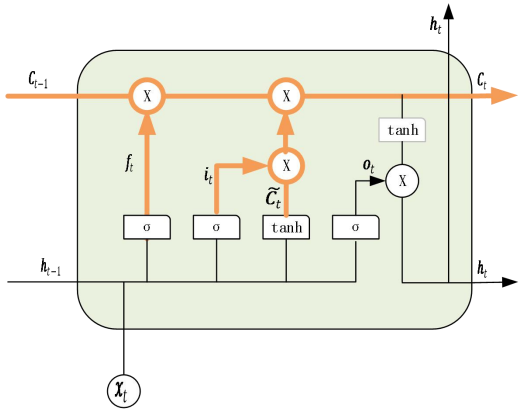

图 3-5 细胞更新状态图

$$\tilde{C}_t = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \qquad (3\text{-}4)$$

在上图及公式中，现在需要对旧的细胞状态进行更新。$i_t * \tilde{C}_t$是对细胞状态进行更新，$f_t * C_{t-1}$是丢弃掉不需要的信息，然后两者相加，获取候选值C_t。即通过在 t 时刻的输入和上一层信息的输出把上一层C_{t-1}的细胞状态更新为C_t。

（3）输出门

首先运行一个 sigmoid 函数层进行确定是要将哪个部分进行输出。这个值经过 tanh 激活函数后，得到一个-1 到 1 之间的值，将该值和o_t相乘，最后只会输出确定输出的部分[41]。LSTM 的输出门状态图如 3-6 所示：

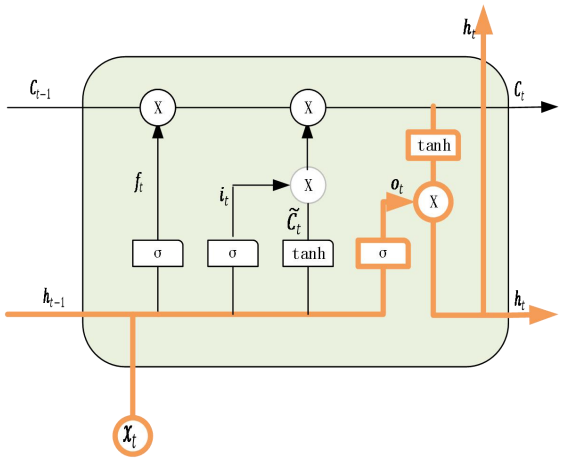

图 3-6 输出门状态图

$$o_t = \sigma(W_o[h_{t-1},x_t] + b_o) \qquad (3\text{-}5)$$

$$h_t = o_t * \tanh(C_t) \qquad (3\text{-}6)$$

在上图及上式中，首先，本文使用 Sigmod 层决定哪一部分的神经元状态需要被输出；然后让神经元状态经过 tanh（让输出值变为-1~1 之间）层并且乘上 Sigmod 门限的输出，只输出与预测结果相关的参数。在本文中，如果给定了一个县，要输出这个县相关的信息，那么可以从包含这个县的上一级行政单位的信息中获取。

3.1.3 LSTM 损失函数

在构建模型时，为了使预测结更和真实值之间的关系更加清晰，一般用损失函数衡量。损失函数（loss function）或代价函数（cost function）是将随机事件或其有关随机变量的取值映射为非负实数以表示该随机事件的“风险”或“损失”的函数。在应用中，损失函数通常作为学习准则与优化问题相联系，即通过最小化损失函数求解和评估模型。

（1）均方误差（Mean Squared Error）

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y^i - f(x^i))^2 \qquad (3\text{-}7)$$

均方误差是指预估结果和实际结果之差平方的期望值; MSE 可以评价数据的变化程度，MSE 的值越小，说明预测模型描述实验数据具有更好的精确度[42]。（ i 表示第 i 个样本，N 表示样本总数）

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y^i - f(x^i))^2} \qquad (3\text{-}8)$$

均方根误差是均方误差的算术平方根，能够直观观测预测值与实际值的离散程度。通常用来作为回归算法的性能指标。

（3）平均绝对误差（Mean Absolute Error）

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y^i - f(x^i)| \qquad (3\text{-}9)$$

平均绝对误差是绝对误差的平均值，平均绝对误差能更好地反映预测值误差的实际情况。通常用来作为回归算法的性能指标。

本文武警警务数据预测模型的整体框架如图 3-7 所示：

3.1.4 训练 LSTM 神经网络总体框架

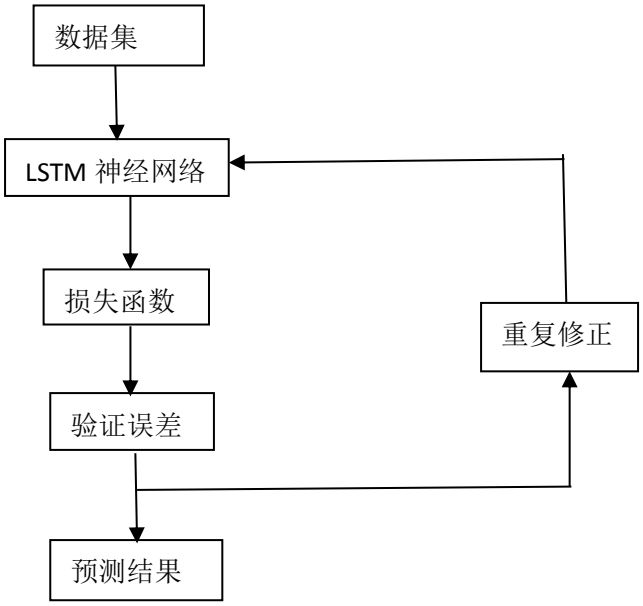

图 3-7 整体框架图

3.2 实验设计

3.2.1 参数设置

在本文中，由于原始警务数据繁多，杂乱无章，所以需要对数据进行合适的处理。本文在处理数据的过程中，针对数据的缺失值、数据特征、数据的分割等采用了不同的方法。具体方法如下表 3-2 所示：

表 3-2 相关参数设置表

方法	缺失值	数据特征	数据分割	预测时间
1	删除法	删除部分属性	按数据特点划分训练集和测试集	按天预测
2	填充法	保留全部属性	随机划分训练集和测试集	按月预测
3	删填结合法			

上表表明，在预测模型有四个环节（包括缺失值处理、数据特征、数据分割和预测时间）需要选取相关参数。不同参数组合可以建立不同的测试方法。具体的方法如下表 3-3 所示：

表 3-3 模型参数组合表

测试种类	组合方法
测试 1	按天预测、删除法、保留全部属性、训练集：2008-2016 年数据，测试集：2017-2018 年数据
测试 2	按天预测、填充法、保留全部属性、训练集：2008-2016 年数据，测试集：2017-2018 年数据
测试 3	按天预测、删减结合法、保留全部属性、训练集：2008-2016 年数据，测试集：2017-2018 年数据
测试 4	按天预测、删除法、删除部分属性、训练集：2008-2016 年数据，测试集：2017-2018 年数据
测试 5	按天预测、删减结合法、删除部分属性、训练集：2008-2016 年数据，测试集：2017-2018 年数据
测试 6	按天预测、删减结合法、删除部分属性、按 7:3 的比例随机划分训练集和测试集
测试 7	按天预测、删减结合法、保留全部属性、按 7:3 的比例随机划分训练集和测试集
测试 8	按天预测、填充法、删除部分属性、按 7:3 的比例随机划分训练集和测试集
测试 9	按月预测、删除法、保留全部属性、训练集：2008-2016 年数据，测试集：2017-2018 年数据
测试 10	按月预测、填充法、保留全部属性、训练集：2008-2016 年数据，测试集：2017-2018 年数据
测试 11	按月预测、删减结合法、保留全部属性、训练集：2008-2016 年数据，测试集：2017-2018 年数据
测试 12	按月预测、删除法、删除部分属性、训练集：2008-2016 年数据，测试集：2017-2018 年数据
测试 13	按月预测、删减结合法、删除部分属性、训练集：2008-2016 年数据，测试集：2017-2018 年数据
测试 14	按月预测、删减结合法、删除部分属性、按 7:3 的比例随机划分训练集和测试集
测试 15	按月预测、删减结合法、保留全部属性、按 7:3 的比例随机划分训练集和测试集
测试 16	按月预测、填充法、删除部分属性、按 7:3 的比例随机划分训练集和测试集
测试 17	按天预测、删除法、保留全部属性、按 7:3 的比例随机划分训练集和测试集
测试 18	按天预测、删除法、删除部分属性、按 7:3 的比例随机划分训练集和测试集
测试 19	按天预测、填充法、保留全部属性、按 7:3 的比例随机划分训练集和测试集
测试 20	按月预测、删除法、删除部分属性、按 7:3 的比例随机划分训练集和测试集（后四个测试参数不再赘述）

3.2.2 评价指标

创建好模型后，需要对其进行分析来判断预测能力的好坏。合适的评价指标算法能够更加容易的验证模型的性能，本文用平均绝对误差（MAE）来计算回归模型的损失函数。MAE 是目标值和预测值之差的绝对值之和，其计算结果较为精确，公式简单，便于理解且不考虑方向，能够对模型更好的评判。本文运用 MAE 来分别计算每一次递归中训练集和测试集的损失率，两者的损失率越小，

则表明数据的精确度越好。之后通过各个参数结果之间的比较，确定最终的警务数据预测模型，然后把本年度数据导入模型中进行数据预测。在第三章中介绍了LSTM 模型的几个损失函数，本文由于数据量过多，为了能够更好地反映出预测值产生误差的实际情况，选择了平均绝对误差作为损失函数。原始数据经过处理后，根据处理方法的不同，每次训练模型时，选择不同的参数类型。在公式 $\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y^i - f(x^i)|$中，N 表示样本个数，$y^i$表示第 i 个样本的预测值，$f(x^i)$表示第 i 个样本的真实值。每循环一次就会得到一个预测值，然后求这个预测值和真实值之间的损失率，把所有的损失率求和并取平均值，求出平均损失率。如果平均损失率越小表示武警警务数据预测得越准，反之亦然。

3.3 实验结果

在本次的武警警务数据预测模型中，数据量多而杂，因此为了使预测结果更加准确，减少失误率，本文先对数据进行归一化，为了使时间序列适配机器学习，需要先将时间序列数据转换为监督学习数据。在构造 LSTM 模型后，为了度量神经网络输出的预测值与实际值之间的差距，本文使用回归算法中的 MAE 算法作为损失函数。在第二章中，对数据的处理有以下几种：

第一种：对原始数据中缺失值的处理，有删除法、填充法和删填结合法，在填充法中，使用了众数填充和平均数填充两种。

第二种：在数据特征提取中，根据特征是否发散，人为判断这列属性是否舍弃。那对于属性的选择可分为舍弃其中几列属性和保留全部属性。

第三种：在实验中对数据集需要划分成训练集和测试集。根据划分方法的不同又可以分成两种情况，第一种是根据原始数据的特征进行划分，第二章中，把2008-2016 年的数据作为训练集，把 2017 年和 2018 年的数据作为测试集；第二种方法是随机划分，把原始的数据量以 7:3 的比例划分为训练集和测试集。

在本文中，依据各个环节所选取的不同参数，可构造 24 种测试种类。每 6 种类型构建一张表，表 3-4 如下所示：

表 3-4 实验参数组合 1

序号	参数 1	参数 2	参数 3	参数 4	预测结果
1	按天预测	删除法	保留全部属性	训练集：2008-2016 年数据，测试集：2017-2018 年数据	RMSE=29.533 训练集损失率：0.15 测试集损失率：0.55
2	按月预测	删除法	保留全部属性	训练集：2008-2016 年数据，测试集：2017-2018 年数据	RMSE=4.721 训练集损失率：0.08 测试集损失率：0.35
3	按天预测	填充法	保留全部属性	训练集：2008-2016 年数据，测试集：2017-2018 年数据	测试 2：RMSE=4.966 训练集损失率：0.05 测试集损失率：0.33
4	按月预测	填充法	保留全部属性	训练集：2008-2016 年数据，测试集：2017-2018 年数据	RMSE=4.691 训练集损失率：0.08 测试集损失率：0.34
5	按天预测	删减结合法	保留全部属性	训练集：2008-2016 年数据，测试集：2017-2018 年数据	RMSE=42.599 训练集损失率：0.2 测试集损失率：0.55
6	按月预测	删减结合法	保留全部属性	训练集：2008-2016 年数据，测试集：2017-2018 年数据	RMSE=2.672 训练集损失率：0.03 测试集损失率：0.05

在上表的 6 个实验中，首先设置所有的参数 3 的值为保留全部属性，参数 4 的训练集采用 2008-2016 年数据，测试集为 2017-2018 年数据，只比较参数 1 和参数 2 的不同。在 1 和 2 两个实验中，参数 2 都是采用删除法，以此比较参数 1 的变化。在两者的结果对比中，按天预测的 RMSE 要远高于按月预测的结果，这说明按月预测的精确度相对较高。实验 3 和实验 4，参数 2 采用填充法，在实验结果中，两者的 RMSE 结果相近，但测试集的损失率都远高于训练集的损失率，可能产生过拟合现象。实验 5 和实验 6，参数 2 采用删减结合法，在实验结果中，两者的 RMSE 相差较大，按月预测的精确度要远高于按天预测，但测试集的损失率都远高于训练集的损失率，可能产生过拟合现象。在实验 1、3 和 5 中，都是按天预测，只有参数 2 的设置不同，对数据的处理分别采用了删除法、填充法和删减结合法，得到的实验结果中，实验 1 的精确度最低，实验 3 次之，实验 5 的 RMSE=2.672，精确度最高，但是每个实验中训练集和测试集的损失率都相差较大，可能产生过拟合现象。在实验 2、4 和 6 中，都是按月预测，只有参数 2 的设置不同，对数据的处理分别采用了删除法、填充法和删减结合法，得到的实验

结果中，实验 2 的精确度最高，实验 2 和实验 4 的 RMSE 都在 4.7 左右，两者精确度次之；在训练集和测试集的损失率中，实验 6 的损失程度相对较小，且两者差距相对较小。总之，上述的实验中，在参数 3 为保留全部属性，参数 4 的训练集采用 2008-2016 年数据，测试集为 2017-2018 年数据下，参数 1 采用按月预测的效果会比按天预测的效果在 RMSE 方面有所降低；参数 2 采用删减结合法会比删除法和填充法在训练集和测试集的损失率方面有所降低。本文选取部分实验结果并进行分析：

（1）测试 5 实验结果

在测试 5 中，选取按天预测并用删填结合法处理缺失值，选取参数 3 中的保留全部属性，选取参数 4 中的根据原始数据的特征划分训练集和测试集。原始数据根据以上方法处理数据后，就可以通过训练 LSTM 模型预测后一天每个城市发生事件的数量。预测结果如下图 3-8 和 3-9 所示：

```
Epoch 43/50
 - 1s - loss: 0.0161 - val_loss: 0.0352
Epoch 44/50
 - 1s - loss: 0.0157 - val_loss: 0.0345
Epoch 45/50
 - 1s - loss: 0.0154 - val_loss: 0.0344
Epoch 46/50
 - 1s - loss: 0.0150 - val_loss: 0.0333
Epoch 47/50
 - 1s - loss: 0.0149 - val_loss: 0.0336
Epoch 48/50
 - 1s - loss: 0.0145 - val_loss: 0.0328
Epoch 49/50
 - 1s - loss: 0.0144 - val_loss: 0.0331
Epoch 50/50
 - 1s - loss: 0.0141 - val_loss: 0.0322
Test RMSE: 42.599
```

图 3-8 测试 5 实验结果图 1

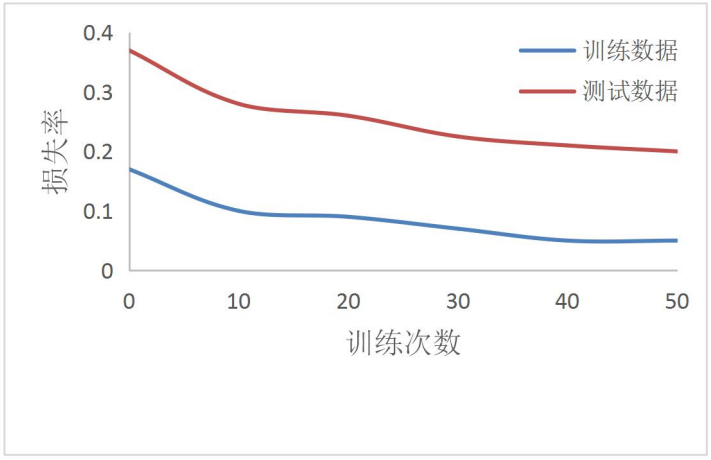

图 3-9 测试 5 实验结果图 2

从图 3-8 可知，数据集的每一次训练，其真实值和预测值之间的误差率都在 0.09 左右徘徊，与前两次试验的结果相比，减少了一定误差率。在图中最下面一行可以看到 RMSE 的值为 42.599，这种方法训练的模型和其余 5 种得失相比在精确度方面差距较大，有大幅度下降。 在图 3-9 中，随着训练次数的增多，训练集的这条曲线由刚开始的 0.4 左右的损失率逐渐降到 0.2 左右，测试集的这条曲线的损失率逐渐趋于 0.05，说明过拟合现象在逐渐减少，由以上结果分析可知，用此种参数组合处理数据并不能训练出有效的预测模型。

（2）测试 6 实验结果

在测试 6 中，选取按月预测且参数 2 中的删填结合法处理缺失值，选取类参数 3 中的保留全部属性，选取参数 4 中的根据原始数据的特征划分训练集和测试集。通过训练 LSTM 模型预测后一月每个城市发生事件的数量。具体内容如 3.3.2 参数设置中所示。预测结果如下图 3-10 和 3-11 所示：

```
Epoch 42/50
 - 0s - loss: 0.1187 - val_loss: 0.1945
Epoch 43/50
 - 0s - loss: 0.1206 - val_loss: 0.1979
Epoch 44/50
 - 0s - loss: 0.1194 - val_loss: 0.1933
Epoch 45/50
 - 0s - loss: 0.1225 - val_loss: 0.1921
Epoch 46/50
 - 0s - loss: 0.1194 - val_loss: 0.1935
Epoch 47/50
 - 0s - loss: 0.1191 - val_loss: 0.1932
Epoch 48/50
 - 0s - loss: 0.1188 - val_loss: 0.1930
Epoch 49/50
 - 0s - loss: 0.1193 - val_loss: 0.1952
Epoch 50/50
 - 0s - loss: 0.1185 - val_loss: 0.1923
Test RMSE: 2.673
```

图 3-10 测试 6 实验结果图 1

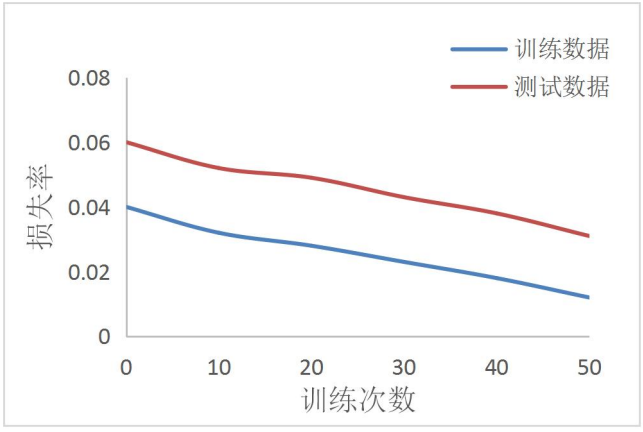

图 3-11 测试 7 实验结果图 2

从图 3-10 可知，数据集的每一次训练，其真实值和预测值之间的误差率都在 0.08 左右，与前几次试验的结果相比，减少了误差率。在图中最下面一行可以看到 RMSE 的值为 2.672，这种方法训练的模型和前两次相比在精确度方面有很大的提升。 在图 3-11 中，随着训练次数的增多，训练集的这条曲线由刚开始的 0.05 左右的损失率逐渐降到 0.01 左右，测试集的这条曲线的损失率逐渐 0.07 降到 0.02，但是仍存在过拟合现象。由以上结果分析可知，这种方法处理数据虽然也不能训练出预期的预测模型。但和前五次实验结果相比，在参数 3 和参数 4 恒定的情况下，按月预测并使用删填结合法处理缺失值在某种程度上对预测结果产生了很大的影响。

在本文中，测试 7-12 的实验结果如下表 3-5 所示：

在表 3-5 的 6 个实验中，首先设置所有的参数 3 的值为保留全部属性，参数 4 的数据集按 7:3 的比例随机划分训练集和测试集，只比较参数 1 和参数 2 的不同。在 7 和 8 两个实验中，参数 2 都是采用删除法，以此比较参数 1 的变化。在两者的结果对比中，按天预测和按月预测的 RMSE 相差不大，精确度都较低，训练集和测试集相差较多，这说明这两种参数组合方式不能训练出有效的预测模型。实验 9 和实验 10，参数 2 采用填充法，在实验结果中，两者的 RMSE 结果相近，但测试集的损失率都远高于训练集的损失率，可能产生过拟合现象。实验 11 和实验 12，参数 2 采用删减结合法，在实验结果中，两者的 RMSE 在 3.6 左右，相差较小并且精确度较高，但测试集的损失率都高于训练集的损失率，可能产生过拟合现象。在实验 7、9 和 11 中，都是按天预测，只有参数 2 的设置不同，对数据的处理分别采用了删除法、填充法和删减结合法，得到的实验结果中，实验 7 的精确度最低，实验 9 次之，实验 11 的 RMSE=3.656，精确度最高，但是每个实验中训练集和测试集的损失率都相差较大，并不能训练出预测模型。在实验 8、10 和 12 中，都是按月预测，只有参数 2 的设置不同，对数据的处理分别采用了删除法、填充法和删减结合法，得到的实验结果中，实验 12 的精确度最高，实验 8 和实验 10 的 RMSE 都很高，与实验 12 相比，两者的精确度都很低；在训练集和测试集的损失率中，实验 12 的损失程度相对较小，且两者差距相对较小。

总之，上述的实验中，在参数 3 为保留全部属性，参数 4 的数据集按 7:3 的比例随机划分训练集和测试集下，这 6 种参数组合方式都不能训练出有效的预测模型。本文选取部分实验结果并进行分析：

表 3-5 实验参数组合 2

序号	参数 1	参数 2	参数 3	参数 4	预测结果
7	按天预测	删除法	保留全部属性	按 7:3 的比例随机划分训练集和测试集	RMSE=40.367 训练集损失率：0.17 测试集损失率：0.46
8	按月预测	删除法	保留全部属性	按 7:3 的比例随机划分训练集和测试集	RMSE=38.579 训练集损失率：0.21 测试集损失率：0.48
9	按天预测	填充法	保留全部属性	按 7:3 的比例随机划分训练集和测试集	RMSE=31.754 训练集损失率：0.22 测试集损失率：0.53
10	按月预测	填充法	保留全部属性	按 7:3 的比例随机划分训练集和测试集	RMSE=32.115 训练集损失率：0.28 测试集损失率：0.57
11	按天预测	删减结合法	保留全部属性	按 7:3 的比例随机划分训练集和测试集	RMSE=3.656 训练集损失率：0.10 测试集损失率：0.20
12	按月预测	删减结合法	保留全部属性	按 7:3 的比例随机划分训练集和测试集	RMSE=3.638 训练集损失率：0.13 测试集损失率：0.21

（3）测试 11 实验结果

在测试 11 中，按天预测选取参数 2 中的删减结合法处理缺失值，参数 3 中的保留全部属性，选取参数 4 中随机划分划分训练集和测试集。通过训练 LSTM 模型预测后一天每个城市发生事件的数量。预测结果如下图 3-20 和 3-21 所示：

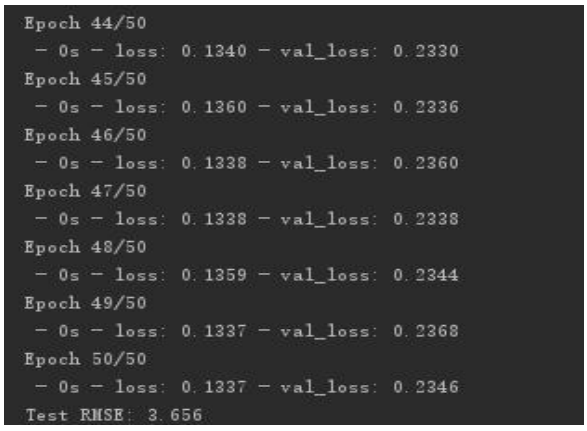

图 3-12 测试 11 实验结果图 1

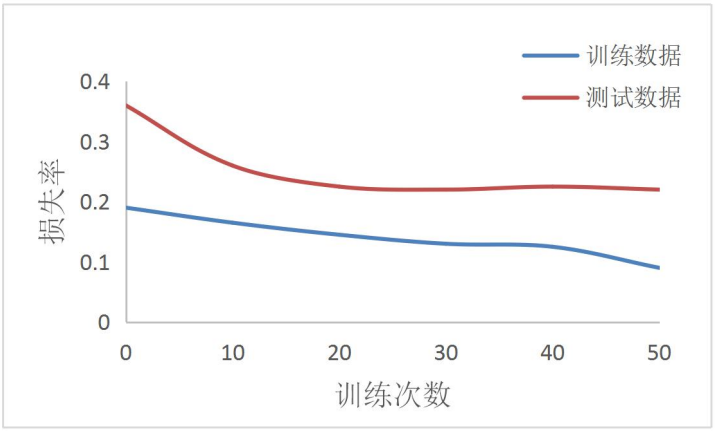

图 3-13 测试 11 实验结果图 2

从图 3-12 可知，数据集的每一次训练，其真实值和预测值之间的误差率都在 0.1 左右，与测试 7-10 的实验结果相比，误差率有小幅度下降。在图中最下面一行可以看到 RMSE 的值为 3.656，这种方法训练的模型和前几次相比在精确度方面有所提升。 在图 3-13 中，随着训练次数的增多，训练集和测试集这两条曲线仍存在很大的损失率，而且产生过拟合现象。由此分析可知，这种参数的组合方式，对整体的训练结果并没有产生很大的影响。所以对于本文的警务数据而言，这种数据处理方法对训练 LSTM 模型并没有产生实质性的影响。

（4）测试 12 实验结果

选取参数 1 中的按月预测，参数 2 中的删减结合法处理缺失值，参数 3 中的保留全部属性，以及参数 4 中随机划分划分训练集和测试集。通过训练 LSTM 模型预测后一月每个城市发生事件的数量。具体内容如 3.3.2 参数设置中所示。预测结果如下图 3-14 和 3-15 所示：

```
Epoch 43/50
 - 0s - loss: 0.1208 - val_loss: 0.1984
Epoch 44/50
 - 0s - loss: 0.1212 - val_loss: 0.1983
Epoch 45/50
 - 0s - loss: 0.1213 - val_loss: 0.1983
Epoch 46/50
 - 0s - loss: 0.1211 - val_loss: 0.1986
Epoch 47/50
 - 0s - loss: 0.1210 - val_loss: 0.1988
Epoch 48/50
 - 0s - loss: 0.1211 - val_loss: 0.1989
Epoch 49/50
 - 0s - loss: 0.1213 - val_loss: 0.1989
Epoch 50/50
 - 0s - loss: 0.1213 - val_loss: 0.1989
Test RMSE: 3.638
```

图 3-14 测试 12 实验结果图 1

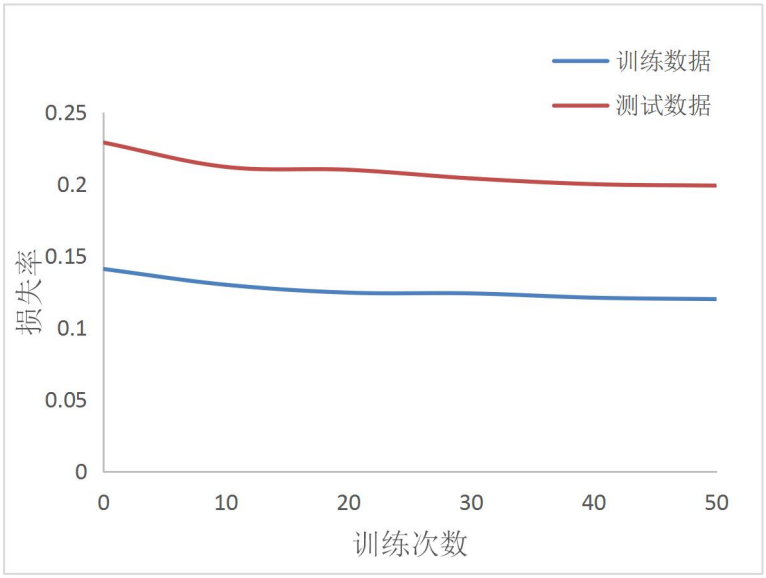

图 3-15 测试 12 实验结果图 2

从图 3-36 可知，数据集的每一次训练，其真实值和预测值之间的误差率都在 0.07 左右，与测试 11 的实验结果相比，误差率有所降低。在图中最下面一行可以看到 RMSE 的值为 3.638，这种方法训练的模型和第五次相比在精确度方面下降幅度不是很大。 在图 3-37 中，随着训练次数的增多，训练集和测试集这两条曲线之间仍存在很大的误差率，而且产生过拟合现象。由此分析可知，和测试 6 的实验数据处理的方法相比，在前三种参数一致的情况下，在把第四种参数由根据原始数据的特征划分训练集和测试集换成按照 7:3 随机划分数据时，数据的精确度以及测试集和训练集的误差率都有所提升，因此，数据集划分的方法不同对实验的预测也会有一定的影响。

在本文中，测试 13-18 的实验结果如下表 3-6 所示：

在表 3-6 的 6 个实验中，首先设置所有的参数 3 的值为删除部分属性，参数 4 的数据集按 7:3 的比例随机划分训练集和测试集，只比较参数 1 和参数 2 的不同。在 13 和 14 两个实验中，参数 2 都是采用删除法，以此比较参数 1 的变化。在两者的结果对比中，按天预测和按月预测的 RMSE 相差不大，精确度都较低，训练集和测试集相差较多，这说明这两种参数组合方式不能训练出有效的预测模型。实验 15 和实验 16，参数 2 采用填充法，在实验结果中，按月预测的精确度高于按天预测，但测试集的损失率都远高于训练集的损失率，可能产生过拟合现象。实验 17 和实验 18，参数 2 采用删减结合法，在实验结果中，两者的 RMSE

表 3-6 实验参数组合 3

序号	参数 1	参数 2	参数 3	参数 4	预测结果
13	按天预测	删除法	删除部分属性	按 7:3 的比例随机划分训练集和测试集	RMSE=28.427 训练集损失率：0.17 测试集损失率：0.44
14	按月预测	删除法	删除部分属性	按 7:3 的比例随机划分训练集和测试集	RMSE=27.665 训练集损失率：0.19 测试集损失率：0.35
15	按天预测	填充法	删除部分属性	按 7:3 的比例随机划分训练集和测试集	RMSE=35.003 训练集损失率：0.03 测试集损失率：0.05
16	按月预测	填充法	删除部分属性	按 7:3 的比例随机划分训练集和测试集	RMSE=31.724 训练集损失率：0.035 测试集损失率：0.045
17	按天预测	删减结合法	删除部分属性	按 7:3 的比例随机划分训练集和测试集	RMSE=3.420 训练集损失率：0.12 测试集损失率：0.24
18	按月预测	删减结合法	删除部分属性	按 7:3 的比例随机划分训练集和测试集	RMSE=3.339 训练集损失率：0.18 测试集损失率：0.21

在 3.4 左右，相差较小并且精确度较高，但测试集的损失率都高于训练集的损失率，可能产生过拟合现象。在实验 13、15 和 17 中，都是按天预测，只有参数 2 的设置不同，对数据的处理分别采用了删除法、填充法和删减结合法，得到的实验结果中，实验 15 的精确度最低，实验 13 次之，实验 17 的 RMSE=3.420，精确度最高，但是每个实验中训练集和测试集的损失率都相差较大，并不能训练出有效的预测模型。在实验 14、16 和 18 中，都是按月预测，只有参数 2 的设置不同，对数据的处理分别采用了删除法、填充法和删减结合法，得到的实验结果中，实验 18 的精确度最高，实验 14 和实验 16 的 RMSE 都很高，与实验 18 相比，两者的精确度都很低；在训练集和测试集的损失率中，实验 16 的损失程度相对较小，且两者差距相对较小。总之，上述的实验中，在参数 3 为保留全部属性，参数 4 的数据集按 7:3 的比例随机划分训练集和测试集下，这 6 种参数组合方式都不能训练出有效的预测模型。本文选取部分实验结果并进行分析：

（5）测试 17 实验结果

在测试 17 中，选取参数 1 中的按天预测，参数 2 中删减结合法处理缺失值，参数 3 中的根据数据是否发散，人为删除部分属性列；选取参数 4 中的对原始数据 7:3 划分为训练集和测试集。通过训练 LSTM 模型预测后一天每个城市发生事件的数量。具体内容如 3-6 参数设置中所示。预测结果如下图 3-16 和 3-17 所示：

```
Epoch 43/50
 - 0s - loss: 0.1094 - val_loss: 0.2257
Epoch 44/50
 - 0s - loss: 0.1092 - val_loss: 0.2242
Epoch 45/50
 - 0s - loss: 0.1111 - val_loss: 0.2248
Epoch 46/50
 - 0s - loss: 0.1095 - val_loss: 0.2272
Epoch 47/50
 - 0s - loss: 0.1085 - val_loss: 0.2250
Epoch 48/50
 - 0s - loss: 0.1090 - val_loss: 0.2263
Epoch 49/50
 - 0s - loss: 0.1086 - val_loss: 0.2261
Epoch 50/50
 - 0s - loss: 0.1089 - val_loss: 0.2228
Test RMSE: 3.420
```

图 3-16 测试 17 实验结果图 1

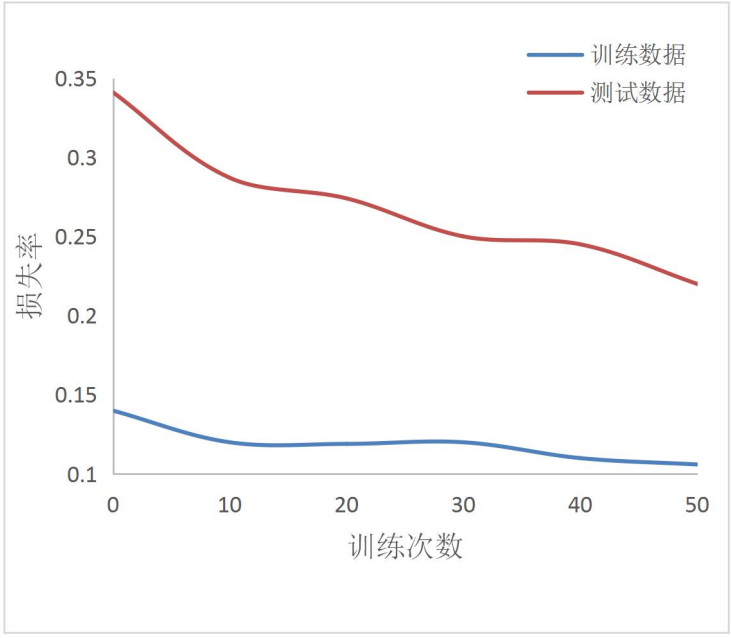

图 3-17 测试 17 实验结果图 2

从图 3-16 可知，数据集的每一次训练，其真实值和预测值之间的误差率都

在 0.1 左右，与前几次的实验结果相比，误差率还是比较高。在图中最下面一行可以看到 RMSE 的值为 3.420，这种方法训练的模型和测试 13-16 相比在精确度方面都有很大的下降。 在图 3-17 中，随着训练次数的增多，训练集和测试集这两条曲线仍存在很大的损失率，而且产生过拟合现象。由此分析可知，和测试 11 实验数据处理的方法相比，参数 3 的不同，模型的精确度产生一定的影响。所以针对本文警务数据而言，在其他参数保持恒定的情况下，参数 3 的设置可能会对实验预测模型产生作用。

（6）测试 18 实验结果

选取参数 1 中的按月预测，参数 2 中的删减结合法处理缺失值，选取参数 3 中的根据数据是否发散，人为删除部分属性列；选取参数 4 中的对原始数据随机划分为训练集和测试集。通过训练 LSTM 模型预测后一月每个城市发生事件的数量。具体内容如 3.3.2 参数设置中所示。预测结果如下图 3-18 和 3-19 所示：

从图 3-18 可知，数据集的每一次训练，其真实值和预测值之间的误差率都在 0.03 左右，与测试 17 的实验结果相比，损失率有所降低。在图中最下面一行可以看到 RMSE 的值为 3.339，这种方法训练的模型和前几次相比在精确度方面

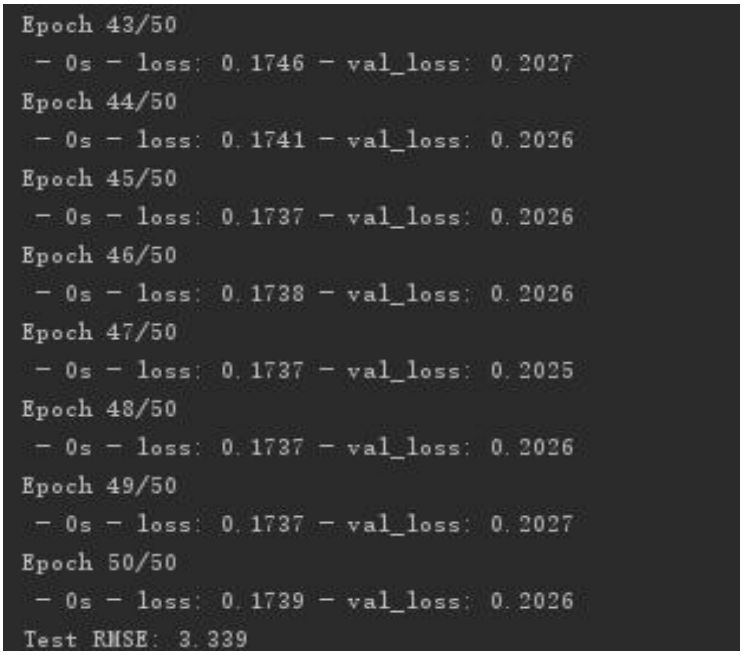

图 3-18 测试 18 实验结果图 1

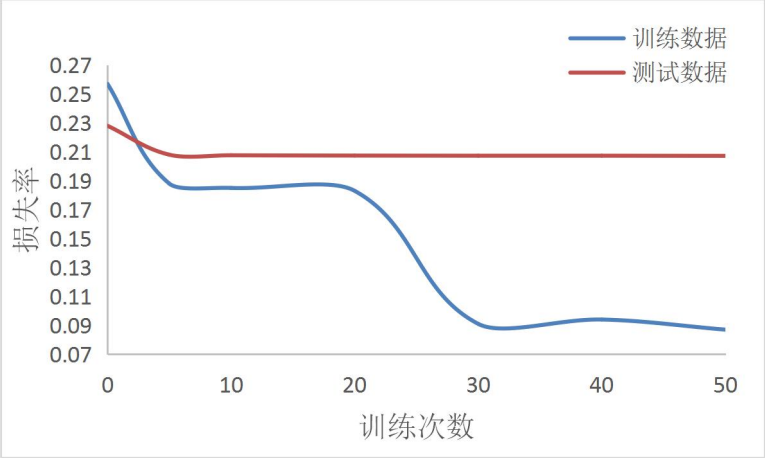

图 3-19 测试 18 实验结果图 2

都有很大的下降。在图 3-19 中，随着训练次数的增多，训练集和测试集这两条曲线之间仍存在很大的误差率，而且产生过拟合现象。由此分析可知，和测试 12 实验数据处理的方法相比，在其他参数一致的情况下，把参数 3 由保留全部属性改为删除部分属性，会对整体的训练结果产生一定影响。所以针对本文警务数据而言，应该人为的删除一些不必要的属性。

在本文中，测试 19-24 的实验结果如下表 3-7 所示：

表 3-7 实验参数组合 4

序号	参数 1	参数 2	参数 3	参数 4	预测结果
19	按天预测	删除法	删除部分属性	训练集：2008-2016 年数据，测试集：2017-2018 年数据	RMSE=29.533 训练集损失率：0.07 测试集损失率：0.33
20	按月预测	删除法	删除部分属性	训练集：2008-2016 年数据，测试集：2017-2018 年数据	RMSE=26.397 训练集损失率：0.08 测试集损失率：0.24
21	按天预测	填充法	删除部分属性	训练集：2008-2016 年数据，测试集：2017-2018 年数据	RMSE=24.397 训练集损失率：0.19 测试集损失率：0.60
22	按月预测	填充法	删除部分属性	训练集：2008-2016 年数据，测试集：2017-2018 年数据	RMSE=22.397 训练集损失率：0.12 测试集损失率：0.75
23	按天预测	删减结合法	删除部分属性	训练集：2008-2016 年数据，测试集：2017-2018 年数据	RMSE=4.568 训练集损失率：0.015 测试集损失率：0.014
24	按月预测	删减结合法	删除部分属性	训练集：2008-2016 年数据，测试集：2017-2018 年数据	RMSE=3.848 训练集损失率：0.010 测试集损失率：0.01

在上表的 6 个实验中，首先设置所有的参数 3 的值为删除部分属性，参数 4 的值仍像第一个表中一样采用相同的数据集，只比较参数 1 和参数 2 的不同。在 19 和 20 两个实验中，参数 2 都是采用删除法，以此比较参数 1 的变化。在结果对比中，按天预测的 RMSE 与按月预测的结果相差不大，精确度都相对较低，测试集的损失率相对较高，说明这两种组合方法可能会产生过拟合现象，并不能训练出有效的预测模型。实验 21 和实验 22，参数 2 采用填充法，在实验结果中，两者的 RMSE 结果相近，但测试集的损失率都远高于训练集的损失率，这种参数组合方式也仍然存在过拟合现象。实验 23 和实验 24，参数 2 采用删减结合法，在实验结果中，两者的 RMSE 相差较小，按月预测的精确度要稍高于按天预测，测试集的损失率和训练集的损失率近似，所以这两种参数组合方式都能够很好地训练出预测模型。但 24 的组合方式在精确度和损失率方面相对实验 23 都表现较好，更能最大程度的训练出合适的预测模型。在实验 19、21 和 23 中，都是按天预测，只有参数 2 的设置不同，对数据的处理分别采用了删除法、填充法和删减结合法，得到的实验结果中，实验 19 的精确度最低，实验 21 次之，实验 23 的 RMSE=4.568，精确度最高，但是实验 19 和 21 中训练集和测试集的损失率都相差较大，可能产生过拟合现象。在实验 20、22 和 24 中，都是按月预测，只有参数 2 的设置不同，对数据的处理分别采用了删除法、填充法和删减结合法，得到的实验结果中，实验 24 的精确度远高于实验 20 和 22；在训练集和测试集的损失率中，实验 24 的损失程度都趋近于 0.01。总之，在上述的实验中，参数 1 采用按月预测，参数 2 采用删减结合法，参数 3 采用删除部分属性，参数 4 的训练集采用 2008-2016 年数据，测试集采用 2017-2018 年数据，这种组合方式可以建立更合适的 LSTM 神经网络预测模型。本文选取部分实验结果并进行分析：

（7）测试 23 实验结果

在测试 23 中，选取参数 1 中的按天预测，参数 2 中的删减结合法处理缺失值，选取参数 3 中的根据数据是否发散，人为删除部分属性列；选取参数 4 中的根据原始数据的特征划分训练集和测试集。通过训练 LSTM 模型预测后一天每个城市发生事件的数量。具体内容如 3-7 参数设置中所示。预测结果如下图 3-20

和 3-21 所示：

```
 - 0s - loss: 0.0591 - val_loss: 0.3121
Epoch 43/50
 - 0s - loss: 0.0595 - val_loss: 0.3118
Epoch 44/50
 - 0s - loss: 0.0601 - val_loss: 0.3175
Epoch 45/50
 - 0s - loss: 0.0609 - val_loss: 0.3154
Epoch 46/50
 - 0s - loss: 0.0605 - val_loss: 0.3130
Epoch 47/50
 - 0s - loss: 0.0595 - val_loss: 0.3128
Epoch 48/50
 - 0s - loss: 0.0601 - val_loss: 0.3119
Epoch 49/50
 - 0s - loss: 0.0596 - val_loss: 0.3103
Epoch 50/50
 - 0s - loss: 0.0594 - val_loss: 0.3149
Test RMSE: 4.568
```

图 3-20 测试 23 实验结果图 1

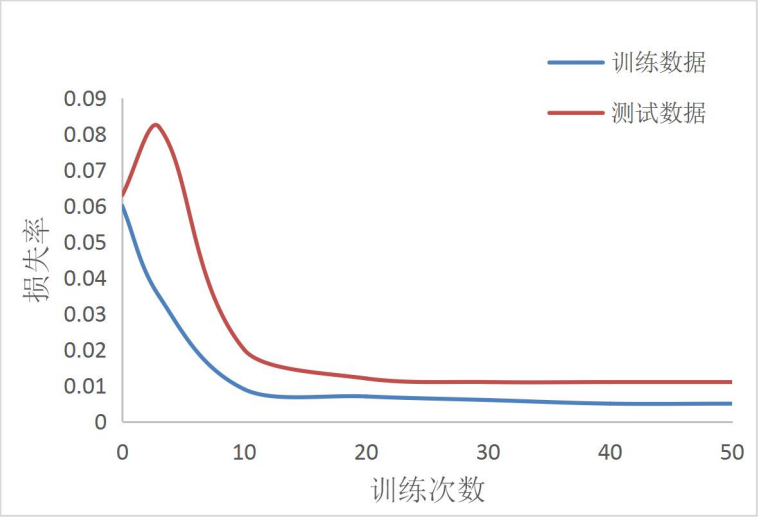

图 3-21 测试 23 实验结果图 2

从图 3-20 可知，数据集的每一次训练，其真实值和预测值之间的误差率都在 0.002 左右，与前 22 次的实验结果相比，误差率都有极为明显的降低。在图中最下面一行可以看到 RMSE 的值为 4.568，这种方法训练的模型和大多数实验相比在精确度方面都有很大的改进。 在图 3-23 中，随着训练次数的增多，训练集和测试集这两条曲线都逐渐趋于平缓，两者的损失率都在 0.015 左右，和之前的实验结果相比误差率几乎可以忽略不计。由此分析可知，这种方法处理数据能够较好的训练出预测模型。即运用这种数据处理方法可以较为合理的预测未来一天某个城市发生的事件量。

（8）测试 13 实验结果

选取参数 1 中的按月预测，参数 2 中的删减结合法处理缺失值，选取参数 3

中的根据数据是否发散，人为删除部分属性列；选取参数 4 中的根据原始数据的特征划分训练集和测试集。通过训练 LSTM 模型预测后一月每个城市发生事件的数量。具体内容如 3.3.2 参数设置中所示。预测结果如下图 3-22 和 3-23 所示：

从图 3-22 可知，数据集的每一次训练，其真实值和预测值之间的误差率都在 0.001 左右，与前 23 次的实验结果相比，误差率都有极为明显的降低。在图中最下面一行可以看到 RMSE 的值为 3.848，这种方法训练的模型和之前实验相比在精确度方面都有很大的提升。 在图 3-23 中，随着训练次数的增多，训练集

```
Epoch 43/50
 - 0s - loss: 0.0812 - val_loss: 0.2223
Epoch 44/50
 - 0s - loss: 0.0813 - val_loss: 0.2212
Epoch 45/50
 - 0s - loss: 0.0807 - val_loss: 0.2210
Epoch 46/50
 - 0s - loss: 0.0806 - val_loss: 0.2212
Epoch 47/50
 - 0s - loss: 0.0806 - val_loss: 0.2212
Epoch 48/50
 - 0s - loss: 0.0806 - val_loss: 0.2209
Epoch 49/50
 - 0s - loss: 0.0806 - val_loss: 0.2211
Epoch 50/50
 - 0s - loss: 0.0807 - val_loss: 0.2219
Test RMSE: 3.848
```

图 3-22 测试 24 实验结果图 1

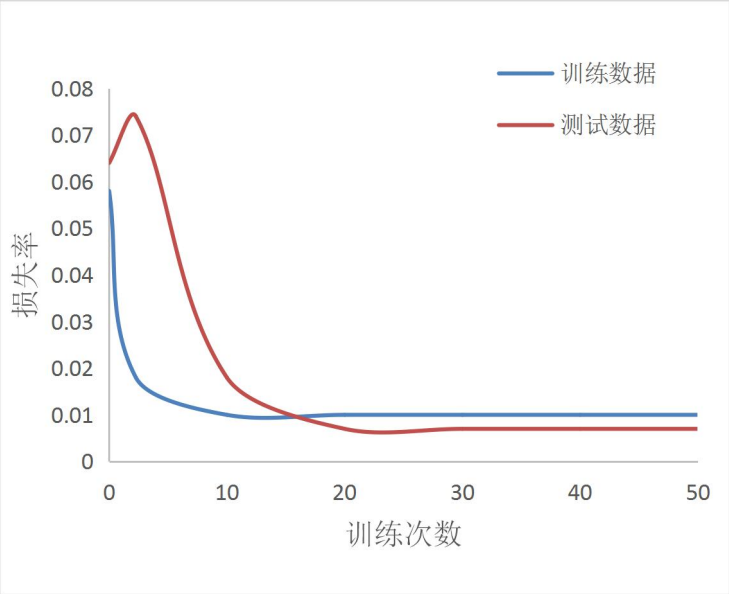

图 3-23 测试 24 实验结果图 2

和测试集这两条曲线都逐渐趋于平缓，两者的损失率都在 0.015 左右，和前几次实验结果相比误差率几乎可以忽略不计。由此分析可知，这种方法处理数据能够较好的训练出预测模型。即运用这种数据处理方法可以较为合理的预测未来一月某个城市发生的事件量。

在以上 24 个测试中，实验 23 和 24 都可以预测出合适的实验模型。上图 3-20 和 3-21 是以天为单位建立的武警警务数据预测模型，图 3-22 和 3-23 是以月为单位建立的武警警务数据预测模型，两者的后三种参数都保持一致。从上述实验结果中可以看到图 3-20 中预测值和真实值这之间的误差率为 0.002，且模型的精确度为 4.568，而图 3-22 中预测值和真实值之间的误差率为 0.001，模型精确度为 3.848。两者相比可知：以月为单位建立的 LSTM 神经网络预测模型不论是在误差值还是精确度方面都比以天为单位建立的模型效果要好。

整体而言，两种测试的预测结果数据还是有着较好准确性和较小误差值的。近年来，随着大数据、云计算等相关技术的发展，我国各地武警情报部门也在不断地探索运用大数据技术指导和服务于部队经常性的执勤、处突、反恐维稳、抢险救援等任务。虽然本文建立的模型可以从大方向上预测结果，但是为了提高预测模型的准确度，还可以进一步的完善和细化。

本章运用 LSTM 神经网络模型对武警警务数据进行分析，得出能指导部队实战的结果，辅助武警情报部门进行科学合理决策，提高兵力部署效能。比如运用模型预测出可能出现重大情况的重点区域，加强在该区域的兵力部署，增派巡逻人员，提高见警率，形成强大震慑效应，有效降低犯罪率；本文在实验过程中发现衡阳市的个人犯罪率相比其他城市较高，那么在下一年各类演习地点可以选在衡阳地区，同时增加衡阳支队的兵力和装备配置。为了进一步检测基于机器学习的武警警务数据预测模型的有效性，本文中又把处理后的数据重新进行分割，将 2016-2018 年的数据作为训练集，然后逐步加入 2015 年及以前的数据，发现模型的损失率越来越高，预测效果越来越差。通过分析发现，2016 年习近平主席提出了科技强国战略，该战略的提出使此后几年数据量急速增多。因此用这几年的数据训练出的模型拟合度比较高。这为下一步模型的进一步优化提供了遵循

和思路。

3.4 小结

本章先对基于机器学习的数据预测模型中涉及到的相关技术进行详细地说明，并分析 LSTM 模型的相关技术和在本文中实现过程。接着介绍程序的整体框架和实验环境，简要说明了模型的设计需求，同时介绍数据预测模型的构建过程，并给出实验流程。在实验设计部分，预测系统部分数据进行导入，通过这些数据产生模型，最终通过导入当年数据，预测未来发生事件的数量，并将结果以图表显现出来，实现数据情报预测模型的建立。在数据导入的过程中，由于原始数据预处理方法的不同，设置了 24 种实验，并不断调整每种实验的参数。最后通过模型对比，验证本文武警警务数据预测模型具有有效性。

第四章 总结与展望

本文通过对原始武警警务数据的分析与研究，设计了 LSTM 神经网络预测模型，解决了武警警务数据长期未被充分利用，分析效率低下的实际问题，并通过现有的数据预测未来某一天或某个月城市发生重大事件的数量。本章节对全文进行了总结，并分析了预测模型存在的不足，提出了后期工作的展望。

4.1 论文工作的总结

本文把原始武警警务数据处理后，运用机器学习的相关算法和理论知识，建立 LSTM 神经网络预测模型，帮助分析近几年湖南发生的事件内在联系，认清现实风险问题，提升各地武警部队决策的精确性和科学性，进而有效提升了部队战斗力。

本文先采集了 2008 年到 2018 年湖南省的武警警务数据信息。通过科学的统计方法获取了大量的武警警务数据，掌握这些数据，可以使管理者在很多方面建立信息优势，并根据这些数据建立合适的信息预测模型，能够进一步加强武警部队的执行力以及掌控力，节约大量的人力和物力，提升执行效率。但是由于原始的警务数据获取渠道众多，每个部门对数据的记录方式和统计情况不一致，往往导致许多信息的缺失以及数据更加驳杂。所以需要对原始警务数据进行一系列的处理。由于传统的武警警务数据的预测模型往往耗费更多的时间并具有较低的精确度，因此，本文中将警务数据和新兴领域相结合，以机器学习的算法和理论为依据，建立了可靠的武警警务数据预测模型，可以对城市发生某种事件的关联度进行合理的预测，帮助武警能够预知事件发生的地点和时间，提早做好防范工作，降低犯罪率，减少人们的伤亡。主要包含以下几个方面：

（1）进行大量数据采集工作。本文搜集了湖南各地区武警情报信息系统近几年的数据，由于互联网迅猛地发展，并广泛用于服务部队各个领域，每年产生了大量数据，其中伴随着大量无用的信息。在信息记录过程中，由于每个城市没有统一的记录规则，导致许多武警警务数据缺失信息；有时还要边搜集数据边清

洗，然后再进行存储，很多时候都是手工和程序相结合的方式进行筛选，删除了大量缺失数据和特征值不发散的属性，补上了部分缺失的信息，并对数据进行合理的转换，这为警务数据信息的处理打下坚实基础。

（2）对搜集整理的数据进行预处理。在建立 LSTM 神经网络预测模型的过程中，先对各个地区的武警警务数据进行采集，运用机器学习的相关理论和算法，如：平均绝对值误差、均方根误差等多个算法。然后依据不同的参数设置，进行多次训练，最终建立了武警警务数据预测模型，帮助武警部队对未来某个时间段该城市可能发生的事件量进行预测，为人力物力的的分配提供合适的方法，把握事件发生动态，为其营造和平氛围，降低伤亡提供合理依据，达到资源利用最大化，创造和谐社会。

（3）在不同参数条件下，对模型进行了 24 次测试，并对结果进行比较分析。在对原始警务数据进行预处理后，使用 LSTM 神经网络模型的长短期记忆功能，分析数据的时间序列特征。通过实验证明了本文提出的基于机器学习的武警警务数据预测模型具有一定的实用性。本文充分考虑了武警警务数据的特点，使用先验知识处理数据，采用半监督方法，有效提高模型预测的准确率。本文以作者实际工作中的武警警务数据为对象，通过对原始数据对象实施不同的处理操作，设置不同的参数训练预测模型。并对每种实验结果进行比较分析，最终得到一种较为有效的 LSTM 神经网络预测模型。

4.2 对未来工作的展望

本文对武警警务数据的预测还存在许多不足，需要进一步研究。同时基于机器学习的理论和方法在武警警务数据处理领域还可以做许多有趣的研究，比如通过对训练成绩数据和部队人员身高体重数据的分析来评判训练方法的有效性，通过了解某地某一时间段内发生群体性事件的概率来调整全省武警机动兵力配置等。当然，完善模型的出发点应向提升部队战斗力聚焦，必须以实际工作需要为牵引。在此我结合下步工作重点方向，谈几点具体的想法：

（1）本文建构的模型还存在很多方面不足。一是运用武警警务数据训练模

型时属性较少，可以在后期收集数据的过程中加入更多属性，比如人口这一项，每个城市发生事件量的多少也与这个城市人口的数量有关联，人越多越密集，越可能发生犯罪事件；二是本文建立的武警警务数据模型预测的变量比较单一，可以在后期的工作中，加入更多属性进行多变量预测。

（2）本文中的数据预测模型，对预测未来事件发生的准确性仍然有很大的提升空间，可以进一步优化算法，提升模型预测的精确度。

（3）针对预测内容相对比较单一，预测范围比较大，下一步将探究如何将本文提出的预测模型的方法进行更细致地划分并对多变量进行预测，以此进一步提高决策信息的可靠性、可用性。

致谢

岁月不居，时节如流，短暂而又充实的三年研究生生涯就这样过去了，我在导师罗迅老师地指导下完成了自己的硕士毕业论文，从选题开题到框架结构、从纲要拟定到实验探索、从初稿写作到盲审修改，都是在罗老师的悉心指导下完成的，非常感谢恩师一步一动、一章一节指导，才有今天收获的喜悦。同时，在预答辩和盲审前刘院长和各位老师也热心给我提出了修改意见，我都悉心领会，认真修改，一并表示感谢！

其实不管是论文的写作还是在平时的生活中，我要感谢的人实在太多，我要感谢和我一起踢足球的伙伴们，感谢数统院易兴对我的帮助，感谢郭晓艳学妹的雪中送炭，还有领导的包容理解，让我能在繁忙的基层工作中抽出时间完成学业，最后我想感谢我的母亲，她默默付出，时常牵挂着我的学习生活，他们都是我向上向善的动力，我这三年的成长历程都伴随着他们的帮助与关怀，我也很欣慰能用这样的方式表达我对他们的感激之情，希望他们都能活成自己心中理想的模样，能够健康快乐。

参考文献

[1] 张于喆,李红宇.互联网新变革与制造业发展[J].中国国情国力,2013(12):39-40.

[2] 闫洲. 基于深度学习的多社交网络中虚拟身份关联技术研究[D].国防科学技术大学,2017.

[3] 王庆娟,祁智宏,徐楠.基于大数据的武警部队财务信息化创新[J].新会计,2017(10):42-43.

[4] 崇卫之. 数据预处理机制的研究与系统构建[D].南京邮电大学,2018.

[5] 李广丽,朱涛,刘斌,殷依,邱蝶蝶,张红斌.面向大数据的数字图书馆多媒体信息检索系统优化研究[J].情报科学,2019,37(02):115-119.

[6] 彭国清. 武警军事数字化评估系统的设计与实现[D].电子科技大学,2014.

[7] 王晓峰.武警部队数据网络安全风险分析[J].计算机与网络,2009,35(12):46-47.

[8] 王学潮.大数据时代人工智能在计算机网络技术中的应用[J/OL].电子技术与软件工程,2019(05):6[2019-03-20].http://kns.cnki.net/kcms/detail/10.1108.TP.20190320.1144.012.html.

[9] 赵海涛,赵毅.大数据时代计算机信息安全防范措施[J/OL].电子技术与软件工程,2019(05):200[2019-03-21].http://kns.cnki.net/kcms/detail/10.1108.TP.20190320.1145.294.html.

[10] 丁岩,鲍焱,胡晓.基于多媒体信息的双向 LSTM 情感分析方法[J].计算机与现代化,2019(02):88-92.

[11] 刘向荣,农忠海,陈雅.公安大数据应用研究的几点思考[J].数字通信界,2016(11):38-41.

[12] 彭知辉.大数据:让情报主导警务成为现实[J].情报杂志,2015,34(05):1-6+16.

[13] 安晖.美国大数据维稳镜鉴[J].人民论坛,2014(12):61-63.

[14] 冯冠筹.大数据时代背景下实施预测警务探究[J].公安研究,2013(12):10-15.

[15] 陈雅,夏元松,农忠海.大数据技术在涉毒案件侦破中的应用[J].电脑知识与技术,2017,13(05):1-2.

[16] 王晓通.机器学习在警务综合系统中的研究和应用[D].电子科技大学,2017.

[17] 刘德伦.公安机关加强“四项建设”的意义与路径[J].四川警察学院学报,2015,27(03):69-73.

[18] 杨宇.大数据技术在人工智能中的应用研究[J].中国新通信,2018,20(18):123.

[19] 刘德松.大数据技术在网络安全管理中的应用分析[J].通讯世界,2018(06):110-111.

[20] 刘希,王萌.大数据和社区警务[J].法制博览,2019(01):232.

[21] George Casella and Roger L.Berger.Statistical Inference,second edition.The Wadsworth Group,2002. Andrew Gelman et al.Bayesian Data Analysis,Third edition.CRC,2014.

[22] Sathyadevan, S.,Devan, M.S.,Surya Gangadharan, S.. Crime analysis and prediction using data mining[P]. Networks & Soft Computing (ICNSC), 2014 First International Conference on,2014.

[23] Featherstone, C.. The relevance of social media as it applies in South Africa to crime prediction[P]. ,2013.

[24] Tayebi, M.A.,Ester, M.,Glasser, U.,Brantingham, P.L.. CRIMETRACER: Activity space based crime location prediction[P]. Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on,2014.

[25] 高帅. 基于机器学习的扫视路径估计方法研究[D].西安电子科技大学,2018.

[26] 李敏蓉,张明,佟志伟.和谐警民关系评价指标体系研究[J].北京人民警察学院学报,2011(03):57-64.

[27] 陈绍骏.对县级公安机关构建和谐警民关系评价指标体系的思考[J].四川警察学院学报,2010,22(01):97-103.

[28] 陈刚.大数据时代犯罪新趋势及侦查新思路[J].理论探索,2018(05):109-114.

[29] 陈甜甜,钟鑫.基于大数据的预测警务在美国的发展现状[J].中国安防,2018(06):106-112.

[30] 李博.多指标综合评价方法应用中存在的问题与对策[J].沈阳工程学院学报(社会科学版),2010,6(03):349-351+363.

[31] 胡宇辉.数据挖掘技术在软件工程中的应用[J/OL].电子技术与软件工程,2019(05):187[2019-03-21].http://kns.cnki.net/kcms/detail/10.1108.TP.20190320.1145.274.html.

[32] 张卫华,靳翠翠.多指标综合评价方法及方法选优研究[J].统计与咨询,2007(01):32-33.

[33] 陈帮鹏.“大数据”时代的计算机信息处理技术探讨[J].科技风,2019(08):95.

[34] 高广尚.大数据环境下市场营销专业建设探讨[J].中国市场,2019(06):129-130.

[35] 张玺君,袁占亭,张红,高玮军,张恩展.交通轨迹大数据预处理方法研究[J/OL].计算机工程:1-6[2019-03-21].https://doi.org/10.19678/j.issn.1000-3428.0049450.

[36] 李广丽,朱涛,刘斌,殷依,邱蝶蝶,张红斌.面向大数据的数字图书馆多媒体信息检索系统优化研究[J].情报科学,2019,37(02):115-119.

[37] 陈建锋.WEB 挖掘数据预处理方法分析与实现[J].安徽职业技术学院学报,2018,17(04):5-7+11.

[38] 陈洁. 数据挖掘分类算法的改进研究[D].南京邮电大学,2018.莫济谦. 基于深度学习的法律问题层叠分类研究[D].湖南大学,2018.

[39] 田杰,周晓娟,吕建新.数据挖掘中聚类算法比较及在武警网络中的应用研究[J].现代电子技术,2008(08):115-117.

[40] 周连明. 云计算在武警部队信息化建设中的应用研究[D].天津大学,2014.

[41] 罗向龙,李丹阳,杨彧,张生瑞.基于 KNN-LSTM 的短时交通流预测[J].北京工业大学学报,2018,44(12):1521-1527.

[42] 窦珊,张广宇,熊智华.基于 LSTM 时间序列重建的生产装置异常检测[J].化工学报,2019,70(02):481-486.

湖南师范大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：周聪　　2019年 8月 15日

湖南师范大学学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，研究生在校攻读学位期间论文工作的知识产权单位属湖南师范大学。同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权湖南师范大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

1、保密□，在_____年解密后适用本授权书。

2、不保密☑。

（请在以上相应方框内打"√"）

作者签名：周聪　　日期：2019年 8月 15日

导师签名：罗迅　　日期：2019年 8月 15日"