**中图分类号：TP31**

**论文编号：**



学术硕士学位论文

**交通事故预测的机器学习方法**

**研究与应用**

作者姓名 任红雷

学科专业 软 件 工 程

指导教师 宋友 教授

培养院系 软件学院

**Research and Application on the Machine Learning Method of Traffic Accident Prediction**

A Dissertation Submitted for the Degree of Master

**Candidate：Honglei Ren**

**Supervisor：Prof. Song You**

School of Software Engineering

Beihang University, Beijing, China

**中图分类号：TP31**

**论文编号：**

硕 士 学 位 论 文

交通事故预测的机器学习方法

研究与应用

作者姓名 任红雷 申请学位级别 学术硕士

指导教师姓名 宋友 职 称 教授

学科专业 软件工程 研究方向 软件工程与管理

学习时间自 2015年 09月 01日 起至 2018年03月31日止

论文提交日期 2018年 01月 10日 论文答辩日期 2018年 01月10日

学位授予单位 北京航空航天大学 学位授予日期 年 月 日

关于学位论文的独创性声明

本人郑重声明：所呈交的论文是本人在指导教师指导下独立进行研究工作所取得的成果，论文中有关资料和数据是实事求是的。尽我所知，除文中已经加以标注和致谢外，本论文不包含其他人已经发表或撰写的研究成果，也不包含本人或他人为获得北京航空航天大学或其它教育机构的学位或学历证书而使用过的材料。与我一同工作的同志对研究所做的任何贡献均已在论文中作出了明确的说明。

若有不实之处，本人愿意承担相关法律责任。

学位论文作者签名： 日期： 年 月 日

学位论文使用授权书

本人完全同意北京航空航天大学有权使用本学位论文（包括但不限于其印刷版和电子版），使用方式包括但不限于：保留学位论文，按规定向国家有关部门（机构）送交学位论文，以学术交流为目的赠送和交换学位论文，允许学位论文被查阅、借阅和复印，将学位论文的全部或部分内容编入有关数据库进行检索，采用影印、缩印或其他复制手段保存学位论文。

保密学位论文在解密后的使用授权同上。

学位论文作者签名： 日期： 年 月 日

指导教师签名： 日期： 年 月 日

**摘 要**

随着现代交通的发展和城市化趋势，交通领域的问题愈发凸显，在一些大城市如：北京，上海，广州等，交通事故越来越多，据统计2016年中国交通事故约50万起，因交通事故而死亡的人数约25万人。此外，国务院在第十二个五年规划纲要中明确提出要将交通事故致死率减少36%。这些证明了交通安全问题已变成一个不可忽视的问题，为了减少交通事故以及相关的生命和财产损失，对交通事故准确的实时预测研究很有必要。

同时，随着大数据时代的到来，传统的交通研究也开始从基于统计规律的研究转变为基于交通大数据的研究。越来越多的异质类型的数据如交通流探测器产生的：浮动车GPS数据、线圈车流量数据，地铁、公交的刷卡数据，基于互联网和社交媒体产生的LBS数据。这些不同的数据源每天产生着大量的数据，为研究和和解决交通问题提供了无限的可能。而近年来兴起的机器学习和深度学习技术，也为我们分析和解决交通问题提供了新思路和有效的手段。因此，采用基于城市交通大数据的方法，对实时城市交通事故预测的机器学习方法进行研究很有必要。本课题研究在分析当前需求与问题的基础上，主要研究内容包括：

1. 设计一种具有较高准确率和召回率的交通事故预测的机器学习方法

利用相关性分析找出了交通事故时空模式，以及相关影响因素。并在此基础上，设计了一种能够较精准的预测城市区域交通事故的机器学习方法。

1. 基于预测性能对影响交通事故的因素进行了分析和排序

单个的交通事故较为随机，通过统计方法分析出影响交通事故的重要因素，对改善城市交通管理策略、以及进行更好的事故预测至关重要。基于格兰杰因果系数，通过比较不同特征和因素组合的情况下的预测性能，对影响交通事故的主要因素进行了重要性排序。

1. 设计异质交通大数据的存储策略和架构

根据不同类型的交通数据（出租车轨迹数据、交通事故数据、天气、空气质量等）的特点和数据量，设计了较为通用和高效的存储策略和架构。

1. 原型系统的开发与验证

通过建立原型系统，使用前述所提出的交通事故预测方法、框架、存储方案，基于 B/S(浏览器/服务器)架构，设计了一种较为通用的、具有良好交互性的，实时交通事故可视化、及事故预测平台，方便交通管理者和决策者进行实时的交通决策。本课题从系统功能和运行效率两方面验证工作的有效性，为一般性的交通事故预测平台提供参考依据。

本课题通过设计具有较高准确率的交通事故预测方法，以及相应的设计出在线的实时事故预测可视化平台。为城市的交通事故避免和决策的提供了方法和策略的支持和方法参考。

关键词：交通事故预测，机器学习方法，异质大数据，影响因素分析

**Abstract**

With the development of modern traffic and the trend of urbanization, the problems in the transportation field have become increasingly prominent. In some big cities such as Beijing, Shanghai and Guangzhou, there are more traffic accidents. According to statistics, about 500,000 traffic accidents in China in 2016 About 250,000 people died as a result of traffic accidents. In addition, the State Council explicitly proposed in the 12th Five-Year Plan to reduce the fatality rate of traffic accidents by 36%. These prove that the problem of traffic safety has become a problem that can not be ignored. In order to reduce the traffic accidents and the related loss of life and property, it is necessary to study the real-time prediction of traffic accidents accurately.

At the same time, with the advent of the era of big data, traditional traffic research has also begun to shift from research based on statistical laws to research based on big data of traffic. More and more heterogeneous types of data are generated by traffic detectors: floating car GPS data, coil car flow data, subway, bus swipe data, LBS data generated from the Internet and social media. These different data sources generate a large amount of data every day, providing endless possibilities for studying and solving traffic problems. In recent years, the rise of machine learning and deep learning technology, but also provide us with new ideas and effective means to analyze and solve traffic problems. Therefore, it is necessary to study the machine learning method of real-time urban traffic accident forecasting based on the method of urban traffic big data. Based on the analysis of current needs and problems, the main research contents include:

1) Design a machine learning method for traffic accident prediction with high accuracy and recall

Using the correlation analysis to find out the traffic accident spatio-temporal pattern, as well as related influencing factors. Based on this, a machine learning method that can predict traffic accidents in urban areas more precisely is designed.

2) Based on the predictive performance of the factors that affect traffic accidents were analyzed and sorted

Individual traffic accidents are more random. Analyzing the important factors influencing traffic accidents through statistical methods is very important to improve the urban traffic management strategy and make better accident prediction. Based on the Granger causality coefficient, the main factors influencing traffic accidents are sorted by importance by comparing the predictive performance of different features and combinations of factors.

3) Design heterogeneous transportation big data storage strategy

According to the characteristics and data volume of different types of traffic data (taxi track data, traffic accident data, weather, air quality, etc.), a more general and efficient data storage strategy and architecture is designed.

4) Prototype system development and verification.

Through the establishment of a prototype system, the traffic accident forecasting method, the framework and the storage scheme proposed in the foregoing are used to design a generic and good interactive visualization system for real-time traffic accidents based on the B / S (browser / server) architecture , And accident prediction platform to facilitate traffic managers and policy makers to make real-time traffic decisions. This subject verifies the validity of the work from the aspects of system function and operation efficiency, and provides reference for the general traffic accident forecasting platform.

The subject of this paper is to design a traffic accident prediction method with high accuracy and to design a visualization platform of real-time accident prediction online. For the city's traffic accident prevention and decision-making methods and strategies to provide support and methods for reference.

**Key words**: Traffic Accident Prediction, Machine Learning Method, Heterogeneous and Big data, Analysis of Influencing Factor

**目 录**

[第一章 绪论 1](#_Toc503020809)

[1.1 研究背景 1](#_Toc503020810)

[1.1.1 研究背景 1](#_Toc503020811)

[1.1.2 研究意义 1](#_Toc503020812)

[1.2 国内外研究现状 2](#_Toc503020813)

[1.2.1 交通事故严重性的影响因素分析 2](#_Toc503020814)

[1.2.2 交通事故热点区域发现 3](#_Toc503020815)

[1.2.3 交通事故起因识别 4](#_Toc503020816)

[1.2.4 实时交通事故预测 4](#_Toc503020817)

[1.3 研究内容 4](#_Toc503020818)

[1.3.1 设计较高准确率的交通事故预测的机器学习方法 4](#_Toc503020819)

[1.3.2 基于预测性能对影响交通事故的因素进行了分析和排序 5](#_Toc503020820)

[1.3.3 设计异质交通大数据的存储策略和架构 5](#_Toc503020821)

[1.3.4 原型系统的开发与验证 5](#_Toc503020822)

[1.4 论文组织结构 5](#_Toc503020823)

[第二章 交通事故预测的机器学习方法 7](#_Toc503020824)

[2.1 整体设计思路和架构 7](#_Toc503020825)

[2.2 异质大数据 7](#_Toc503020826)

[2.3 数据预处理过程 7](#_Toc503020827)

[2.4 相关性分析 7](#_Toc503020828)

[2.5 机器学习模型 7](#_Toc503020829)

[2.6 参数选取及训练过程 7](#_Toc503020830)

[2.7 预测效果及性能对比 7](#_Toc503020831)

[2.8 本章小结 7](#_Toc503020832)

[第三章 交通事故影响因素的重要性排序 8](#_Toc503020833)

[3.1 格兰杰因果分析的方法介绍 8](#_Toc503020834)

[3.2 事故影响因素的重要性排序 8](#_Toc503020835)

[3.3 本章小结 8](#_Toc503020836)

[第四章 异质交通大数据的存储策略与架构设计 9](#_Toc503020837)

[4.1 异质大数据类型与格式 9](#_Toc503020838)

[4.2 整体存储架构设计 9](#_Toc503020839)

[4.3 本章小结 9](#_Toc503020840)

[第五章 原型系统的实现与应用 10](#_Toc503020841)

[5.1 应用场景与功能介绍 10](#_Toc503020842)

[5.2 系统的设计与实现 10](#_Toc503020843)

[5.3 运行效率分析 10](#_Toc503020844)

[5.4 本章小结 10](#_Toc503020845)

[第六章 总结与展望 11](#_Toc503020846)

[6.1 工作内容总结 11](#_Toc503020847)

[6.2 工作局限性 11](#_Toc503020848)

[6.3 未来展望 11](#_Toc503020849)

[参考文献 12](#_Toc503020850)

[攻读硕士学位期间取得的学术成果 13](#_Toc503020851)

[致谢 14](#_Toc503020852)

**图 目 录**

**表 目 录**

1. 绪论
   1. 研究背景
      1. 研究背景

在现代社会,城市化的快速发展带来了车辆的增长和繁荣。然而，这却带来了很多亟待政府解决的问题，如交通拥堵。在大数据和深度学习的帮助下，尽管高性能的实时交通流预测已经使得人们可以通过选择相对不拥堵的路线，来避免交通堵塞。而另外一个重要的问题，交通事故，尽管已经造成了巨大的经济损失和人员伤亡，却一直没能被很好的理解和解决。根据世界卫生组织在2015年发布的《全球道路安全现状报告》，每年大约有125万人死于交通事故。因此，对于我们的安全来讲，交通事故预防的研究显得重要而且急迫。

同时随着大数据时代的到来，传统的交通研究也开始从基于统计规律的研究转变为基于交通大数据的研究。越来越多的异质类型的数据如交通流探测器产生的：浮动车GPS数据、线圈车流量数据，地铁、公交的刷卡数据，基于互联网和社交媒体产生的LBS数据。这些不同的数据源每天产生着大量的数据，为研究和和解决交通问题提供了无限的可能。而近年来兴起的机器学习和深度学习技术，也为我们分析和解决交通问题提供了新思路和有效的手段。

在预防交通事故的解决方案中，建立高效而准确的交通事故系统预测无疑是很有希望的一种。因为通过预测系统给出的危险区域信息能够被方便地传递给驾驶员，并且他们可以通过选择不太危险的区域来避免交通事故。然而想要达到准确的事故预测却十分困难，因为许多相关因素都会影响交通事故。举例而言，像下雨和下雪这样的坏天气条件能够限制道路的可见能力和道路通行能力从而导致更多的事故。另外，交通事故的发生率在一天中的不同时间段是变化的。尽管许多研究都关注于识别与交通事故相关联的关键因素，但是他们并不能动态和定量地给出交通事故风险的信息。

* + 1. 研究意义

从学术研究领域看，本研究课题属于城市计算领域下的智能交通子领域。由于交通事故造成了大量的生命和财产损失，而据世界卫生组织统计，中国每年有超过25万人死于交通事故危害。对交通事故的实时预测研究，将可以使得交通管理者们实时调整安全资源的配置、实时发布道路安全信息、使得驾驶员和行人可以选择安全性更高的道路和区域出行，避免高风险的事故区域，从而改善道路的安全性，避免生命和财产损失。

本课题在一方面可以帮助交警部门科学地和有效地安排他们的警力，另一方面可以很容易地应用于现实世界，并通过以选择更安全的区域的方式，来帮助人们尽可能地避免交通事故。

* 1. 国内外研究现状

国内外研究领域在交通事故预测发面，大致经历了四个阶段，即从基于统计方法的交通事故严重性影响因素分析，到依旧基于统计方法的事故起因分析，再到通过聚类手段进行事故热点区域发现，到通过机器学习方法直接预测事故。目前最新的成果是2016年Chen et al，基于日本的移动人群密度进行事故预测的深度降噪编码机(Stacked denoising Autoencoder, SdAE)方法。尽管目前已有不少工作直接或者间接地进行交通事故预测，但它们目前仍存在很多问题。

首先，基于统计方法的事故严重性影响因素分析和事故起因分析，只能给出若干与事故有关的指标，比如5分钟内平均速度变化等等，这些一方面只能进行事故识别，另外一方面最多能给大家一些事故避免的建议，而都无法准确地告诉人们如何具体的避免事故。这一问题尽管在事故热点区域的研究开始后才得以缓解，比如这些方法能够告诉人们哪些地方容易发生事故，请小心或者绕行等等。但实际上事故发生的规律是动态变化的，某些区域如菜市场等可能只有早高峰才会比较容易发生事故，而大型娱乐商圈可能只有晚上事故频繁。

尽管最新的工作——基于深度的降噪编码机方法，利用动态的人群密度来进行事故预测，但实际上该方法太过粗糙，与事故相关的主要因素远不止人群移动特征一种，如交通流情况对车辆间的交通事故更有意义。另外，事故本身的周期性、空间分布规律、天气影响等，该模型以及现有的其他工作都没有考虑到。因此，通过探索交通事故的规律出发，如事故的时空相关性、空间分布规律等，进一步建立考虑因素较为丰富的机器学习模型显得十分必要。基于此，本课题尝试并建立了一种新的机器学习方法，综合分析和考虑这些因素。

下面各小节分别详细介绍这三个阶段对应的主要研究成果。

* + 1. 交通事故严重性的影响因素分析

在目前，国内外已经在道路交通事故严重程度影响因素分析领域开展了广泛研究，其中以经典统计模型的相关应用最为广泛。其中Chang采用了逻辑斯蒂模型对2000年台湾单一车辆道路事故对比了摩托车驾驶员和非摩托车驾驶员的死亡率的影响因子，发现摩托车驾驶员的死亡率是非摩托车驾驶员的3倍多，区别两者的主要因素在于安全带是否使用、行驶速度，与受伤严重性相关的因素包括道路类型、车辆类型。Malyshkina 采用了多项式逻辑斯蒂回归模型对正规设计的道路和设计的例外道路分别进行了事故严重性和事故频率的估计，对比结果显示例外设计的道路并在事故严重性和频率的结果上与正规设计的道路并没有统计的显著差异。Huang 采用了贝叶斯层次二项式逻辑斯蒂回归模型来识别对信号灯路口事故中人员和车辆分别的损伤严重程度的影响因素，结果显示时间、路口的光线照明情况对人员的伤亡严重性有较大影响。而车辆的大小、重量等对车辆在事故中的受损程度有直接影响。国内学者李世民等利用累积Logit模型研究了交叉口交通事故严重程度与转弯车辆、用地性质等因素的关系。马壮林等分别利用累积Logit模型和灰色模型研究了道路设计因素对隧道交通事故严重程度的影响。侯树展等利用主成分分析法研究了交通流特征和高速公路事故严重程度之间的联系。

绝大多数的统计回归模型都依赖于其自身的假设和预定义的变量间的潜在依赖关系。一旦这些假设和依赖关系被违反，这些模型将产生错误的事故严重性估计。考虑到经典统计模型的缺陷和不足，近年来有学者开始尝试将非参数、人工智能方法应用于交通事故严重程度影响因素分析。Mussone等提出了对传统模型的替代性的方法——人工神经网络法，该模型对米兰不同情况下交通路口发生事故的严重性做了计算和分析，结果显示路口的复杂性对高交通事故死亡率有着决定性作用。Chang 提出了分类回归树模型，该模型不需要目标变量和预测变量间的潜在依赖关系，通过对台北交通事故数据的分析，该模型表明对交通事故严重性影响最大的因素是交通工具的类型，行人、骑摩托车、自行车的驾驶者往往是在交通事故中受伤最严重的群体。Juan de Oña 采用了贝叶斯网络模型来对西班牙1536起乡间的交通事故严重程度进行分类，贝叶斯网络模型也无需依赖前提假设，并且能够利用图形化的表征对有着相互组件依赖的辅助系统进行刻画，其采用了18个变量构建了三个不同的网络来对三种不同的事故严重性程度（轻微受伤、严重受伤、死亡）进行分类。其结果显示对于死亡或者严重受伤类型的交通事故，有很大影响的因子有事故类型、司机年龄、道路光线情况、事故受伤人数。研究表明, 非参数模型在实证研究中具有更优的统计拟合度和泛化能力。

* + 1. 交通事故热点区域发现

目前绝大多数识别交通事故聚类的方法主要分为两类：第一类是对一条道路简单输出其是否有聚类的假设检验结果。其代表性的方法有：k-方法[21, 22]和最近邻方法[23, 24]。第二类是可以在一段道路或者路网中识别出精确地聚类区域或者位置，例如核密度估计方法(KDE)[25]或者危险指标方法[26]等。

关于道路交通事故热点，有关文献并没有统一而明确的定义。Hauer 在论文中指出部分学者根据交通事故发生概率按位置给出热点排名，而另外一些学者根据事故发生的频率，每公里道路上的历史事故数量[53]。传统的道路事故热点分析主要集中在道路片段或者特定连接位置[54]，而区域范围的热点以及该区域事故引起的风险分散被人忽略。自20世纪70年代开始，统计模型就被应用于交通事故分析，而早期模型的主要缺陷在于假设事故的发生服从正态分布[55-57]。随后一些学者为了改正这项缺陷，而采用泊松对数线性回归来对事故在空间和时间的随机性进行刻画[58]。而很多其他的学者例如，Hauer和Persaud[59], Miaou[60], Shanker[61], Maher 和 Summersgill[62], Abdel-Aty和Radwan[63]，都采用负二项回归模型。以上提到的模型都考虑了在给定时间段内的报告的事故发生数量，而对于该时段内事故的位置特征则被当作常量。

随后的研究主要基于核密度估计(KDE)的方法来对事故热点区域进行聚类，Sabel 首先采用了核密度估计聚类分析技术来自动的识别交通事故的热点区域，该研究结合了蒙特卡洛模拟技术，能够快速的识别统计显著的聚类[25]。其他学者也陆续采用核密度估计方法提出了改进的方法或者应用[27]。Plug应用空间、时间以及时间-空间结合的技术对西澳大利亚不同尺度的交通事故模式进行了分析，采用蜘蛛图来识别按天和按星期的时间尺度上的交通事故模式，用核密度估计的方法来探究事故的空间结构模式，采用Comap来识别事故的时空间的相互影响[64]。Xie 和Yan之后考虑到KDE长久以来一直被用作事故热点的可视化工具，因为缺乏对其统计推断结果的定量评估，因此结合了NetKDE（一种被证明在网络结构的事故分析中常用的工具）[65]，与100米范围内的局部莫兰指数(一种度量空间相关性的指标)[66]。

* + 1. 交通事故起因识别

在早期交通安全研究方面，许多研究都致力于识别导致交通事故的关键条件或特定的交通模式。例如，Oh提出了破坏性的交通流是导致撞车的导火索(Ohetal.2001)。基于线圈检测器的数据和撞车事故数据，他们发现交通事故发生前的5分钟速度标准差是撞车事故的一个有效指示器。Abdel-Aty则将交通模式的空间变化考虑在内，并且发现事故发生地点上游5分钟的平均占有率和下游车速的5分钟变异系数是交通事故的标志(Abdel-Atyetal.2004)。Golob则通过研究发现平均车流量，流量变化，速度中位数和速度变化是撞车事故的良好标志(GolobandRecker2004)。尽管已有各种各样的交通事故指示因子提出，但是因为很多因素都与交通事故有着复杂的关联，因此它们不能满足精准预测交通事故的需要。

* + 1. 实时交通事故预测

随着机器学习的发展，已有许多的研究开始关注实时交通事故预测。Lv基于欧氏距离选择特征变量并且利用k近邻方法来预测交通事故(Lv,Tang,andZhao2009)。Hossain提出了基于贝叶斯网络的方法来对高速公路的车辆事故进行实时预测(HossainandMuromachi2012)。Park收集了首尔高速公路的交通事故大数据并且构建了一个基于k-均值聚类分析和逻辑斯蒂回归方法的预测工作流程(Park,Kim,andHa2016)。最近，Chen使用日本的行人移动数据并且构建了堆叠降噪编码机来实时的推测交通事故风险(Chenetal.2016)。

然而他们都忽视了交通事故的时间特征，并且也没有意识到其他与交通事故相关联因素的重要性，比如交通流，天气条件和空气质量等。这种忽视会带来预测准确率上的巨大差异，尤其是在城市和国家层面的事故预测中。

* + 1. 交通领域的深度学习研究

对于智能交通系统的深度学习研究，许多研究都聚焦于交通流预测上，如(Lv et al. 2015; Wu and Tan 2016)。在更长的时间尺度上，一些研究致力于预测大规模交通网络的拥堵演化(Ma et al. 2015b)。(Zhu et al 2017) 利用Graph-CNN对城市路网进行交通流预测，并获得了一定的预测性能改进。另一个有趣的应用利用深度增强学习来控制交通信号灯的时长 (Li, Lv, and Wang 2016)。

* 1. 研究内容
     1. 设计较高准确率的交通事故预测的机器学习方法

利用相关性分析找出了交通事故时空模式，以及相关影响因素。并在此基础上，设计了一种能够较精准的预测城市区域交通事故的机器学习方法。

* + 1. 基于预测性能对影响交通事故的因素进行了分析和排序

单个的交通事故较为随机，通过统计方法分析出影响交通事故的重要因素，对改善城市交通管理策略、以及进行更好的事故预测至关重要。基于格兰杰因果系数，通过比较不同特征和因素组合的情况下的预测性能，对影响交通事故的主要因素进行了重要性排序。

* + 1. 设计异质交通大数据的存储策略和架构

根据不同类型的交通数据（出租车轨迹数据、交通事故数据、天气、空气质量等）的特点和数据量，设计了较为通用和高效的存储策略和架构。

* + 1. 原型系统的开发与验证

通过建立原型系统，使用前述所提出的交通事故预测方法、框架、存储方案，基于 B/S(浏览器/服务器)架构，设计了一种较为通用的、具有良好交互性的，实时交通事故可视化、及事故预测平台，方便交通管理者和决策者进行实时的交通决策。本课题从系统功能和运行效率两方面验证工作的有效性，为一般性的交通事故预测平台提供参考依据。

* 1. 论文组织结构

本文共分为六个部分，组织结构如下：

**第一章 绪论**。在该部分中阐述了研究背景，提出了本课题将集中解决的主要任务，并对现有情况的国内外现状进行了分析。

**第二章 交通事故预测的机器学习方法**。在该部分中，详细讲解了交通事故预测方法的整体设计思路和架构、异质数据来源、预处理过程、机器学习模型、参数选取、训练过程。之后对本课题提出的基于深度循环神经网络方法的预测性能，与其他方法包括传统机器学习方法与深度学习方法进行了对比。

**第三章 交通事故影响因素的重要性排序**。在该部分中，首先介绍了格兰杰因果分析的方法，即通过比较不同单独因素或者组合因素的预测准确率。之后，本课题采用该方法，对影响交通事故的若干因素进行了重要性排序。

**第四章 异质交通大数据的存储策略与架构设计**。该部分讲述了本课题如何根据不同类型的交通数据的特点和数据量，设计较为通用和高效的存储策略和存储架构。

**第五章 原型系统的实现与应用**。该部分讲解了开发的交通事故可视化与事故预测原型系统，对其设计、功能及效率进行了一定的分析。

**第六章 总结和展望**。本部分总结本课题的研究内容，研究工作的局限性，并对未来的工作方向进行了展望。

1. 交通事故预测的机器学习方法

交通事故预测的机器学习方法是本课题研究的核心。为了能够建立较高准确率和召回率的机器学习方法，以便于对区域实时的交通事故进行预测，本课题收集了北京1年的异质交通大数据，包括交通流、交通事故、天气、空气质量指数数据等。在对数据进行了相关性分析的基础上，本课题发现交通事故发生的频率具有明显的时间周期性；而空间分布上，则是否的不均匀；另外，每天的城市的总交通事故量和时间段、是否是周末、空气质量指数等明显相关。因此，本课题针对交通事故发生的规律及特点，设计了交通事故预测的整体框架，以及具体的机器学习模型。

因此，本章首先介绍机器学习方法的整体设计思路及架构。接着，对收集的异质大数据以及对其预处理过程进行介绍。之后，本课题对事故的时空相关性进行分析，并探索事故的关联因素以及其空间分布特性。之后，对构成交通事故预测的机器学习模型的网络结构、特征组成、训练过程、参数选择等进行介绍。最后，我们介绍本课题提出方法的预测效果，并与传统机器学习模型和深度降噪编码机模型进行预测准确率对比。

* 1. 整体设计思路和架构

与Chen采用的对交通事故风险预测不同(Chen et al. 2016)，我们尝试以一种直接的方式来对交通事故进行预测。这意味着对在一定区域r，一定时间段t内发生的事故量进行直接预测。因为直接进行预测一方面结果更加简单直观；另一方面，相对于事故风险预测而言，它对交通部门警力配置和事故避免也更有指导意义。为了达到这些目的，我们收集了大量的交通事故以及和事故相关的异质大数据，并构建了一个用来进行实时交通事故预测的深度学习模型。

本课题提出的交通事故预测方法的整体架构图如图 1所示。

首先，我们收集了与交通事故相关的异质大数据，包括交通事故、交通流、天气和PM2.5指数的数据。其次，为了便于计算机对数据进行处理，我们采用了数值化、离散化和归一化的方法对数据进行了预处理。之后我们构建了基于深度循环神经网络的模型来对交通事故进行预测。并且将预处理后的数据输入到了该模型中训练。训练好之后，我们将近期的交通事故、交通流、天气和空气质量的历史数据输入到训练好的模型中。此处的近期历史数据是指前若干小时、昨天的和上周的相关历史数据。通过输入近期历史数据至模型，我们可以从模型输出获得实时的交通事故预测地图。

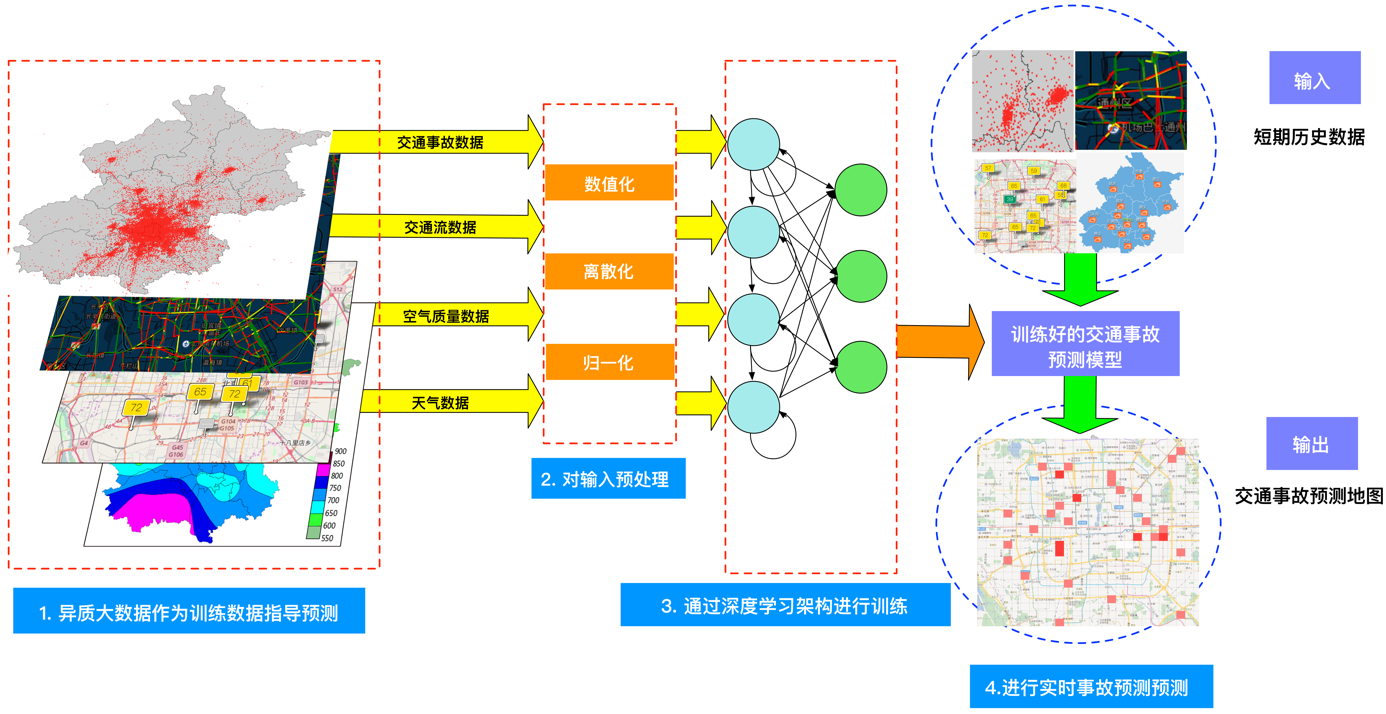


图 1 交通事故预测方法的整体架构图

* 1. 异质大数据

在本研究中，为了准确的预测交通事故，我们收集了与交通事故相关的交通流、天气、空气质量的异质大数据，它们具体如下:

交通事故数据：我们收集了北京从2016年1月-2017年2月一共将近100万条交通事故记录数，每条事故包括事故发生时间、发生地点描述、经度和纬度四个属性。

交通流数据：我们收集了2016年8月北京的出租车每隔5分钟上报一次的GPS位置和行驶平均速度数据。

天气数据：利用自己编写的网络爬虫，我们从历史天气网（http://lishi.tianqi.com）爬取了北京2016年1月-2017年2月每天的天气描述数据。

空气质量数据：同天气数据一样，利用爬虫，我们从天气后报网（http://www.tianqihoubao.com/aqi/），爬取了2016年1月-2017年2月北京每天的PM2.5指数（AQI）数据。

* 1. 数据预处理过程

在我们构建深度学习模型并预测交通事故之前，一个合适的数据结构对于构建准确的和高性能的模型至关重要。首先我们对原始数据进行了一定的预处理：例如数值化，离散化，标准化等等。

对于交通事故数据，我们首先将数据根据时空信息进行了离散化。对于时间维度，我们选择60分钟作为时间间隔。对于空间维度，我们将事故的位置映射到1000米\*1000米的网格中。在我们的实验中，对于北京的数据而言，每个网格的纬度跨度为∆dlat = 0.0084，经度 跨度为∆dlng = 0.012。原始事故数据如图 2所示，我们将交通事故数据进行了空间离散化，如图 3所示。

将数据按照空间和时间离散化后，我们将原本连续的交通事故时空信息转换成了一个带有区域索引和时间索引的离散矩阵形式，记为S。对于特定区域r和特定时间段t，发生的事故总量可以被表示为矩阵S的一个元素，即Sr,t。为了表述方便，下文我们将矩阵的元素Sr,t记为Sr(t)。

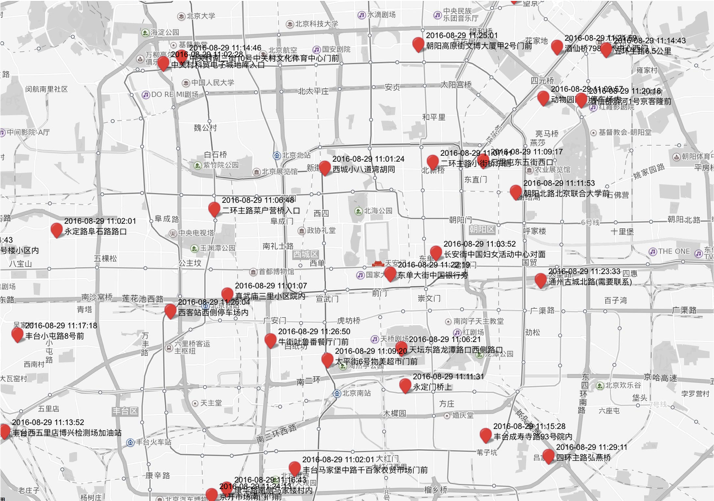


图 2 2016年8月29日11：00-11：30的交通事故原始数据可视化

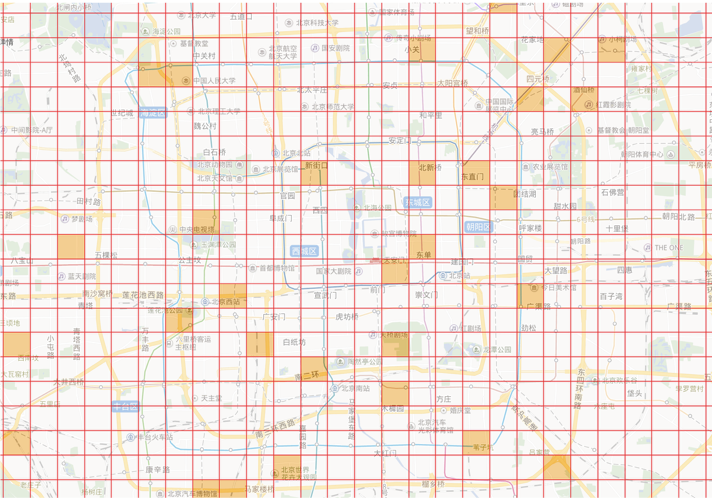


图 3 进行空间离散化之后的事故可视化

对于交通流数据，我们的原始数据是每隔5分钟出租车上报的GPS位置和行驶速度数据。首先我们按照与交通事故的位置数据同样的离散化方法将交通流数据的离散化。这样我们得到的交通流矩阵的每个元素为在特定区域r和特定时间段t内的上报的出租车GPS位置和行驶 速度集合。之后，我们对每个矩阵元素的行驶速度集合分别进行平均从而得到一个新的平均行驶速度矩阵V， 其中每个矩阵元素Vr (t)为在特定区域r和特定时间段t内经过的出租车的平均速度。

因为节假日的交通流状况和交通事故的模式和平时具 有很大的不同，除了与交通事故矩阵和交通流矩阵，我 们还构建了矩阵H来标注每个交通事故矩阵元素Sr(t)对应的时间t是否处于节日或者假日。我们是根据中国法 定放假的节日来定义节假日的。具体而言，包括：元旦，春节，清明节，五一劳动节，端午节，国庆节，中秋节。

另外，由于一天中不同的时间段的交通事故模式，相差很大，如高峰时间较正常时间事故要多很多，我们根据工作时间模式和中国人的生活方式，按照如表 1所示方式对事故矩阵元素Sr (t)所在的时间段t进行标注，形成了与交通事故矩阵对应的时间段矩阵G。

表 1 时间段划分表

|  |  |  |
| --- | --- | --- |
| 序号 | 时间段名称 | 时间区间 |
| 1 | 凌晨至黎明 | 00:00–06:59 |
| 2 | 早高峰 | 07:00–08:59 |
| 3 | 早上工作时间 | 09:00–11:59 |
| 4 | 中午休息时间 | 12:00–13:59 |
| 5 | 下午工作时间 | 14:00-16:59 |
| 6 | 晚高峰 | 17:00–19:59 |
| 7 | 夜间 | 20:00–23:59 |

对于天气数据，我们收集到的原始数据是每天的天气描述，但是这种文字描述很难输入到计算机中，并且对于建立预测模型是没有太大帮助的。因此，必须将文字转化成计算机容易输入和处理的数值形式。我们主观地 根据天气对交通影响的严重性，将天气描述转化成了表 2对应的天气严重性数值。

如果描述为混合天气，如小雨转晴等，其天气严重性为小雨和晴对应严重性的平均。在本例中，天气严重性为 (2 + 0)/2 = 1。对应交通事故矩阵，我们构建了与之每个元素对应的天气严重性矩阵W 。举例而言，元素Wr (t)代表区域r和时间t对应的天气严重性。

表 2 天气描述与其对应的天气严重性

|  |  |
| --- | --- |
| 天气描述 | 天气严重性 |
| 晴, 浮尘, 阴, 多云 | 0 |
| 霾，雾 | 1 |
| 小雨，阵雨 | 2 |
| 中雨, 小雪 | 3 |
| 大雨, 雨夹雪, 暴雨 | 4 |
| 中雪、大雪 | 5 |

对于空气质量数据，我们提供将其除以最大PM2.5指数(500)来对数据进行归一化。即归一化后的PM2.5指数=PM2.5指数/500.0。类似于天气严重性矩阵W，我们利用归一化的数据，构建了与交通事故矩阵对应的空气质量矩阵A。举例而言，元素Ar(t)代表在区域r和时间段t内的标准化空气质量指数。经过预处理后，我们得到了6个矩阵，它们是：交通事故矩阵S和与之元素对应的交通流矩阵V，节假日矩阵H，时间段矩阵G，天气严重性矩阵W，空气质量矩阵A。我们将这些矩阵叠加到一起，形成一个新的矩阵I，对于特定区域r和特定时间t，Ir(t)=[Sr(t), Vr(t), Hr(t), Gr(t), Wr(t), Ar(t)]T。深度学习模型的输入序列元素就是矩阵的每个元素Ir(t)。输出就是该区域r，在该时段t内，预测发生事故的数量。

* 1. 事故时空相关性
     1. 事故区域分布规律

为了探索事故之间是否存在时间或者空间上的相关性，首先我们先观察事故在空间区域上的分布情况。此处，我们观察不同的区域内，在2016年各自发生的总事故量的分布。图 4显示的是2016年北京5环内事故分布热图，红色程度越深，代表事故量越多。我们从图中可以看出，事故在空间区域上呈现出了非常不均衡的现象，即使是对于5环内的城区而言。而事故量最集中的区域是国贸地区、东四十条地区、中关村地区和西直门地区等。

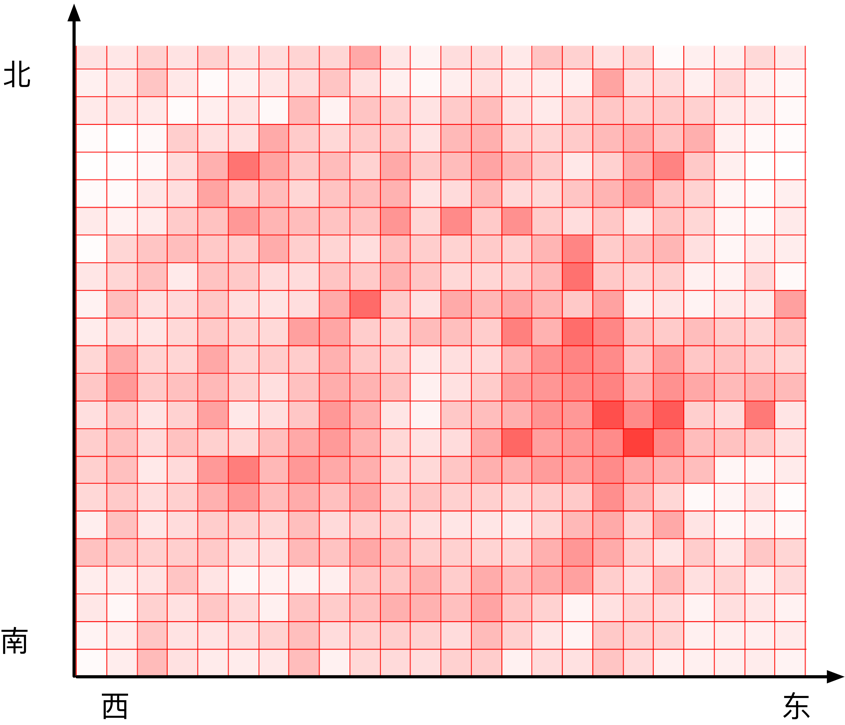


图 4 2016年北京5环内事故分布热图

* + 1. 事故的时间分布规律

为了探索事故的发生是否存在时间段上的规律性。我们给出了2016年北京每天交通事故量在不同时间段的数量分布情况（图 5），图中的时间段的划分方式参见表 1。从图中我们可以看出事故量在不同时间段上具有明显的规律。具体而言，呈现早晚双峰分布，这主要是由于早晚高峰人多、车流量增加造成的。其中以早上9：00—12：00，下午14：00—20：00的事故量最高；其次是中午12：00—14：00；再次是晚间20：00—23：59，事故在00：00到次日7：00的发生频率最低，这也是由于这时候的道路的车流量最少造成的。

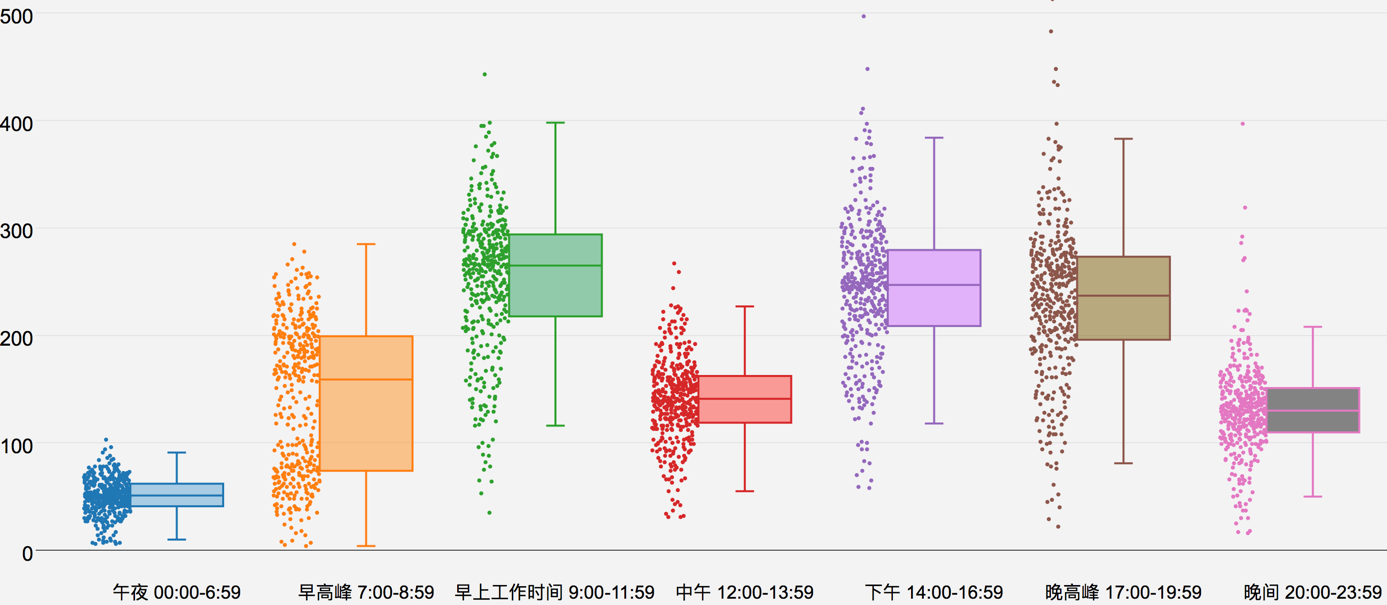


图 5 北京每天事故量在不同时间段的分布

为了探寻交通事故在长时间尺度（每周）上是否具有一定的周期性，我们画出了2016年8月28日（周日）—2016年9月10日（周六），两周内北京总事故量随着时间变化的柱形图（图 5），时间间隔为一小时。图中每个细柱代表一小时内北京发生的交通事故量。图左边给出了事故量的刻度，下方给出了具体日期和星期几。

从图 5中我们可以看出，每天的北京的总交通事故量大致呈双峰分布，峰值附近为早晚高峰代表的时间段；最低的时间为晚上22：00-次日7：00左右。而且通常早高峰的事故量要大于晚高峰的事故量；周六日的双峰现象更加不明显一些；整体上，事故量的分布呈现一定的周期性，周期为7天。举个例子，这周一的事故量分布与上周一有很大的相似性。而从最后两张子图中，即9月16日、9月17日的交通事故分布图中我们也可以看出，事故与昨天的分布也可能存在一定的关系，这在9月9日和9月10日的事故分布中也可以看出来。

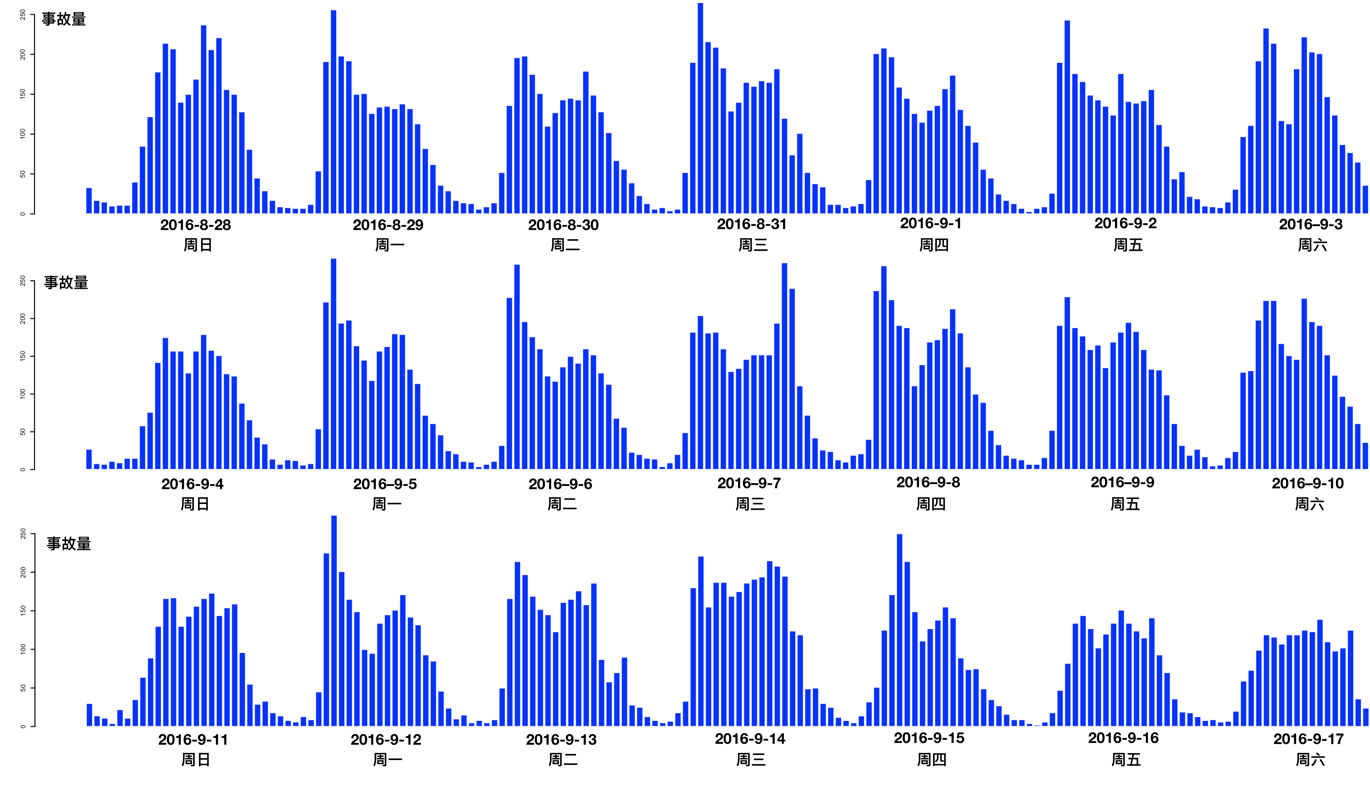


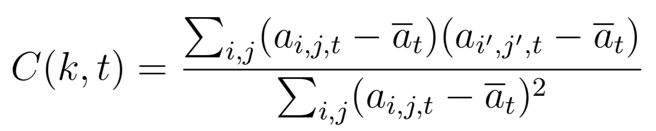
图 6 北京总事故量的时间分布图

* + 1. 事故间的时空相关性

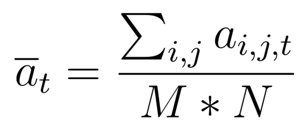
1. 时空相关性定义

根据2.4.1和2.4.2 的分析结果，我们可以看出事故发生的频率在区域上并非不均匀分布，有很大的区域偏好性。而在时间上，事故量也有一定的时间周期性。为了定量描述事故之间在时间和空间上的相关性，我们引入时空相关性概念。

事故的空间相关性定义如下：在时间t上，空间的曼哈顿距离为k的事故间的相关性记为:C(k, t)，其计算公式如下：



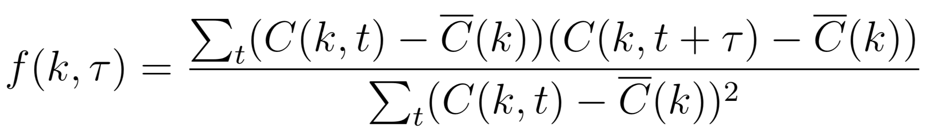
其中为当第网格位置在时间t的事故量,为在时间t上满足的网格， distance为曼哈顿距离；其中t时间段北京五环内平均事故量公式如下：



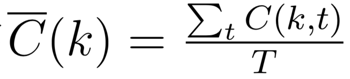
公式中的M=32，N=29，分别代表北京对五环内的经度和纬度划分的数量。

定义出了事故的空间相关性C(k, t)之后，我们便可以定义事故的时空相关性f (k, τ )。

事故的时空相关性：空间曼哈顿距离为k，并且时间相隔为τ的事故间的相关性记为：f (k, τ)。其计算公式如下：



其中平均空间相关性为：

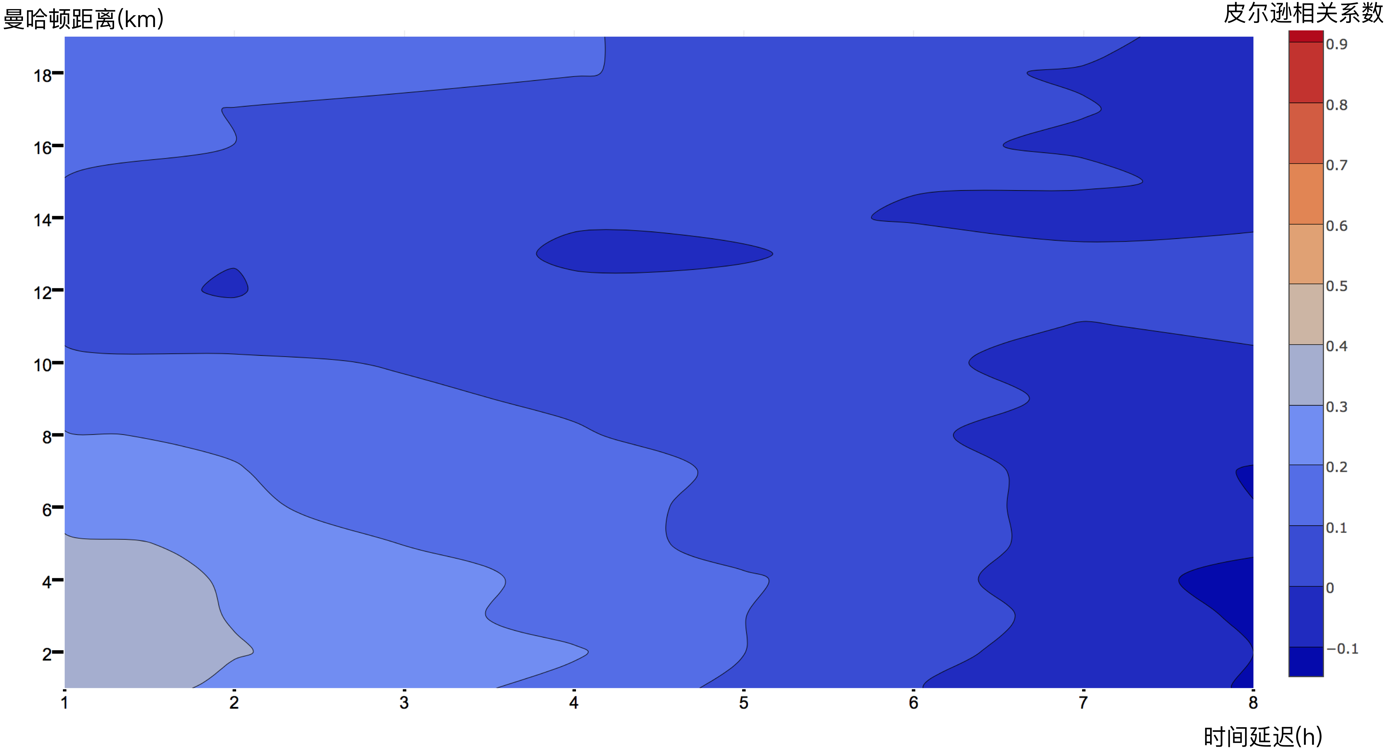


T为总的时间延迟数量。

1. 时空相关性分析结果

我们选取f (k, τ)中的最新时间间隔τmin=1h, 即在计算时空相关性时，最小计算相隔1小时及1小时以上的事故时空相关性。而k最小取1代表最小计算空间间隔为1千米(km)的事故时空相关性。

为时间延迟为8小时内 (1<= τ <= 8)，空间距离(1<= k <=20)的时空相关性的等高线图。



* 1. 事故关联因素分析

方法

* 1. 事故预测的机器学习模型
     1. 模型结构

在本节中，我们介绍我们基于LSTM的交通事故预测方法（Traffic Accident Risk Prediction Method based on LSTM，TAPML）。如图 4所示，该图描述了交通事故预测方法的深度模型。模型的第一层是时间特征输入层，它是一个向量T。它融合了短期和周期特征元素，包括交通事故数量、平均交通速度，天气严重性，归一化后的PM2.5指数，时间段和是否是节假日标注。模型的另外一个输入层是区域经纬度输入层，在该层，输入待预测区域r的中心经度和纬度。因为事故分布在区域上不均衡，这样可以提供区域位置信息。深度模型的隐藏层依次包括4个LSTM层和3个全连接层。模型的最后一层即输出层，它给出对于输入的数据特征，区域r，在时间段t内发生交通事故量的预测。

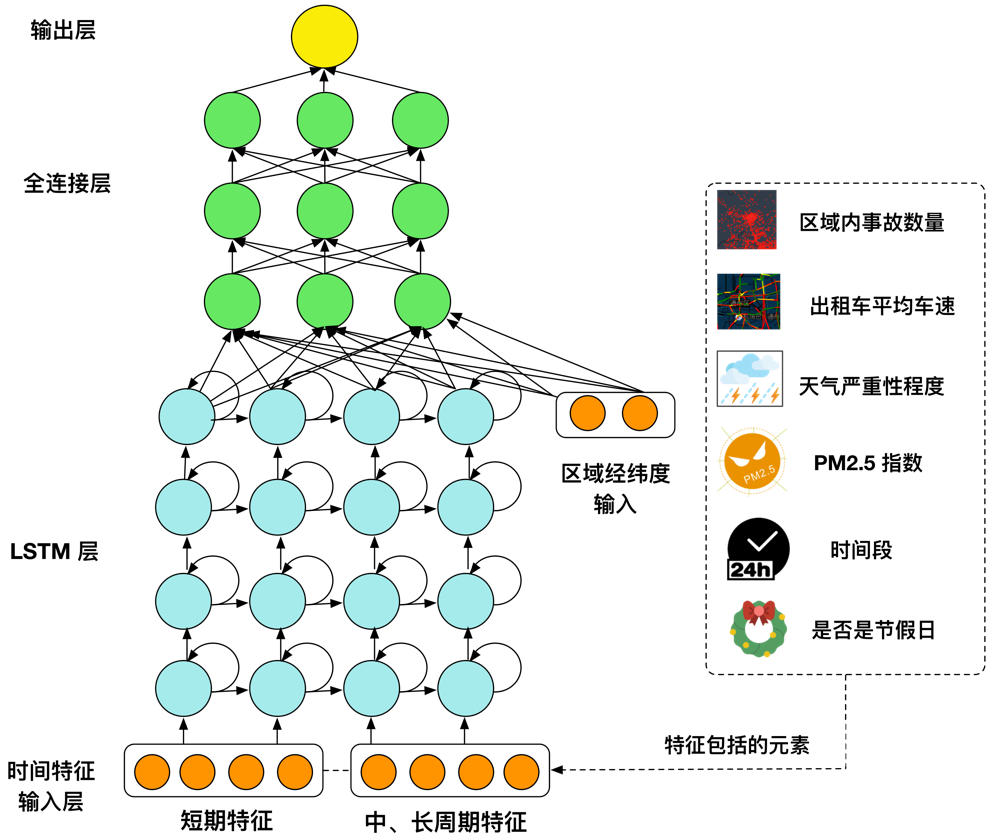


图 7 基于LSTM的交通事故预测方法的深度模型

我们之所以选择LSTM而不是传统RNN模型，主要是因为当时间跨度较长时，传统RNN模型的结果很差并且很难训练。这些传统RNN的缺点已经在有关交通流预测的研究中被证明了\cite{ma2015long}。而另一方面，带有显式记忆单元的LSTM能有效避免传统RNN的梯度消失和爆炸的问题。LSTM与传统的RNN结构类似，有一个输入层，一个或若干隐藏层以及一个输出层组成。LSTM的核心是隐藏层的记忆单元，其结构如图 5所示。一个记忆单元主要由四个主要部分组成：一个输入门，一个包含自循环连接的神经元，一个遗忘门和一个输出门。关于LSTM更细节的介绍可以参见\cite{hochreiter1997long}。

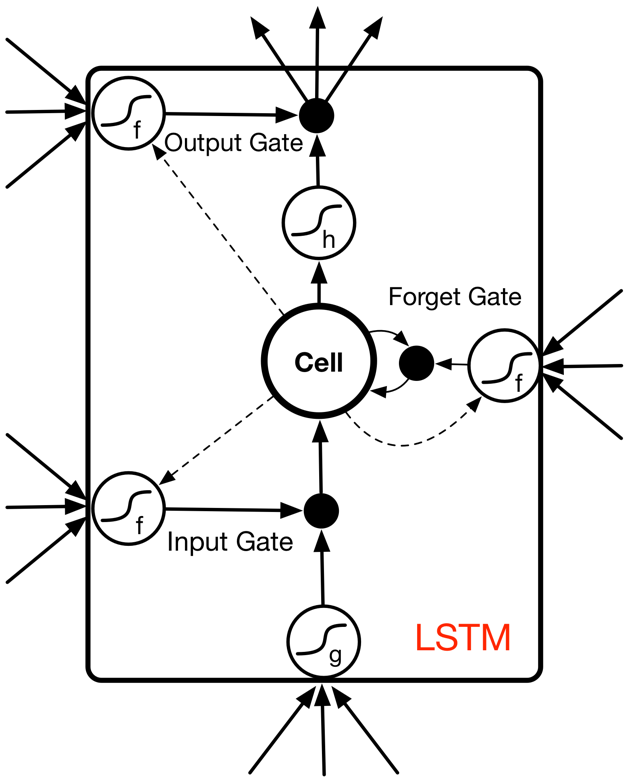


图 8 LSTM记忆单元的结构

* + 1. 短期时间特征

短期特征反映了预测时间前若干小时的短期时间内交通事故和交通流等相关情况。根据之前我们已经介绍过的矩阵，我们可以将输入的短期特征部分记为Ts = [Ir(t-n), Ir (t-n+1), ..., Ir(t-1) ]T。其中n是选取的时间窗长度，Ir(t)为矩阵I的元素，Ir(t) = [Sr(t), Vr(t), Hr(t), Gr(t), Wr(t), Ar(t)]T，代表区域r，在时间段t内的交通事故量、交通流、天气、空气质量、时间段、节假日标注信息。

* + 1. 周期时间特征

交通流和交通事故等交通现象经常在不同周期内的相同时间重复一定的模式。这个周期长度可能是一天或者一星期。因为在工作日，人们通常每天的上下班时间是相同的，出行规律也很少变化。另外，每周六和周日的交通模式都很相似，但与日常的工作日不大一样。这种每周或者每天重复的交通模式能够帮助我们利用历史的交通模式改进交通事故预测。在本节，我们提出了周期特征作为除了短期特征外，LSTM的另一部分输入。这种周期特征包括日周期特征和周周期特征两部分。假设我们要预测区域r，在时间段t内发生的交通事故量，则其日周期特征可以表示为: Td = [Ir(td - nd), Ir(td - nd +1), ..., Ir(td+ nd)]T，表示昨天与时间t相同的时间段td的前后nd个时间间隔内的交通相关情况。对于周周期特征，它可以表示为: Tw = [Ir(tw - nw), Ir(tw – nw +1), ..., Ir(tw + nw)]T，它表示上周与预测时间相同的时间段tw前后nw个时间间隔的交通相关情况。

* + 1. 深度模型

我们将短期特征和周期特征进行顺序地连接后输入深度模型，其可以表示为T=[ Ts, Td, Tw]T。我们的深度模型由4个LSTM层和3个全连接层组成。为了防止模型过拟合，在层与层之间，我们加入了丢弃率为0.5的Dropout层\cite{hinton2012improving}。模型隐藏层的激活函数为RELU（Rectified Linear Units），其数学形式为:max(0, x)。最后，因为模型的输出是二元的，因此我们选择了Sigmoid作为输出层的激活函数。

* 1. 预测效果及性能对比
  2. 本章小结

本章提出了本课题中多人协同环境下3D建模框架的整体概况。本章首先从应用场景出发，通过分析典型的应用场景，得出在新环境下对于多人协同3D建模系统的需求和系统需要解决的主要问题。随后针对这些问题，提出了框架的总览，分层次介绍了每个部分的主要功能和解决的问题。然后对本架构的应用模式进行了分析，并总结了本框架和研究的主要应用方式和拓展点，最后根据提出的协同框架，设计了一种基于零件库的协同建模系统工作流程[36]。

1. 交通事故影响因素的重要性排序

在本章中将讨论对协同3D场景数据的存储工作。基于现有的对于协同3D场景数据的传输研究，本章首先对面向协同环境的3D场景数据进行归纳，总结出在本框架下实现多人协同的数据类型和表示方式，而后针对这些数据，抽象提取出一种合适的抽象数据结构，用于协同环境下3D场景数据的核心描述，并设计相应的文件格式。最后提出一种联合存储方案，实现去冗余的场景数据存储。

* 1. 格兰杰因果分析的方法介绍
  2. 事故影响因素的重要性排序
  3. 本章小结

本章首先介绍了在协同3D建模过程中所需要的数据，为建立用于保存的抽象场景结构提出了具体需求。然后通过具体描述抽象场景结构，详细讲述了结构中各个部分如何满足这些需求并提供了足够的数据和功能支持。而后通过设计对应的文件格式来对数据结构进行实现。最后，通过上述基础提出了一种联合存储方案，详细描述了对于Web3D实时协同建模平台场景的存储和还原方法。

1. 异质交通大数据的存储策略与架构设计

目前已经有一些协同应用使用了HTML5。本章总结现有研究成果，对HTML5技术在协同建模中的应用方法进行统一论述，并与传统的实现方法进行对比。分别是基于WebGL的三维场景构建，基于WebSocket的协同通信，和基于WebStorage的数据存储。

* 1. 异质大数据类型与格式
  2. 整体存储架构设计
  3. 本章小结

本章主要阐述了如何应用HTML5技术解决协同建模平台中的问题，详细介绍了方法与原理，并与传统的技术进行了比较。利用WebGL特性能建立三维协同场景，利用WebSocket能实现实时协同通信，利用WebStorage可以进行客户端数据存储。可以看出使用HTML5技术能实现良好的性能。在下一章原型系统中我们也将使用这些技术。

1. 原型系统的实现与应用

基于上述的研究工作，本课题开发了一个实际工程项目，同时作为原型系统来验证框架设计、场景存储、HTML5应用等工作的有效性。本章首先将介绍该工程项目的应用场景与业务需求，然后介绍该项目的总体设计与详细实现，最后通过功能和效率两方面来检验论文的研究工作。

* 1. 应用场景与功能介绍
  2. 系统的设计与实现
  3. 运行效率分析
  4. 本章小结

本章通过分析功能需求，实际建立Web3D协同建模平台，并详细介绍了系统的技术框架、功能模块设计和细节设计。随后对实现的系统进行了功能性和效率两方面的考察，确定了验证的平台达到了预期的功能完备性和效率可行性，从一定程度上证明了本课题研究的可行性与有效性。最后通过和其他协同3D平台实现方式的比较，展示了本框架的优势和可扩展性。

1. 总结与展望
   1. 工作内容总结
   2. 工作局限性
   3. 未来展望

参考文献

111

攻读硕士学位期间取得的学术成果

学术论文：

[1]. You Song, **Honglei Ren**, and Jinzhi Lei. "Collaborations between CpG sites in DNA methylation." International Journal of Modern Physics B (2017): 1750243.

(**SCI** 第一学生作者)

[2]. **Ren, Honglei**, et al. "Secure smart home: A voiceprint and internet based authentication system for remote accessing." Computer Science & Education (ICCSE), 2016 11th International Conference on. IEEE, 2016. (**EI** 第一学生作者)

[3]. Siyu Yang, You Song, **Honglei Ren**, Xinxing Huang. "An automated student attendance tracking system based on voiceprint and location." Computer Science & Education (ICCSE), 2016 11th International Conference on. IEEE, 2016. (**EI** 第二学生作者)

[4]. 王世华,郭伟,**任红雷**,刘京欣,张腾化,宋友,雷锦誌.交通指挥指数(TCI)的研究与应用[J].中国科技信息,2017(17):74-76+79. (核心期刊 第一学生作者)

[5]. **Ren, Honglei**, et al. "A Deep Learning Approach to the Prediction of Short-term Traffic Accident Risk." arXiv preprint arXiv:1710.09543 (2017). (预印本 第一学生作者)

致谢

222