浙江大學

硕 士 学 位 论 文

论文题目： **基于深度学习的室内时空客流预测**

作 者 姓 名：李邦鹏

指 导 教 师：寿黎但教授

学科 (专业)：计算机科学与技术

所 在 学 院：计算机科学与技术学院

提 交 日 期：二〇一七年一月

A Dissertation Submitted to Zhejiang University
for the Degree of Master of Computer Science

TITLE: Predictions of Indoor Passenger Flow Based On Deep Learning

Author: Bangpeng Li

Supervisor: Prof. Should

Subject: Computer Science and Technology

College: College of Computer Science

Submitted Date: January, 2017

浙江大学研究生学位论文独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得**浙江大学**或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：　　签字日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解**浙江大学**有权保留并向国家有关部门或机构送交本论文的复印件和磁盘，允许论文被查阅和借阅。本人授权浙江大学可以将学位论文的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日

摘要

随着城市的快速发展，商场、会展中心、机场等大型建筑不断涌现。人们的日常活动越来越多的在室内空间完成。而智能手机及室内 WiFi 的普及，也使用户室内位置信息的获取变得更为容易。基于 WiFi 的室内定位系统不仅为用户提供了免费的网络服务，也为建筑的管理方提供了一个低成本的客流监测方案。准确的统计和预测室内的时空客流对商场等大型建筑实现智能运营合理安排人力物力有着重要意义。

本文首先回顾了基于 WiFi 的室内定位技术的研究进展，总结目前在实践过程中存在的一些问题：从原始的 WiFi 信号强度进行定位前时不能高效的识别室内和室外物体。针对这一情况，本文提出了一种基于机器学习的算法用于清洗室外数据噪声。在训练数据的构建上，我们采用众包的思想自动化的采集数据。在我们搭建的实验平台上，我们的方法获得了 98% 的准确率，相比传统的室内室外检测算法，我们的方法更简单，具有一定的研究意义。

此外，传统的客流预测往往通过时间序列的相关方法预测室外环境的宏观客流。预测的出发点是未来的客流分布与历史的客流分布存在相当的关联。然而对于室内客流而言，这种关联不仅体现在相同区域的历史关联，也体现在不同区域的历史关联。基于此本文提出了一种基于卷积神经网络和长短时神经网络的预测模型，将多个区域的客流时空序列作为模型输入，预测室内未来客流的时空分布。基于此方法，我们在国内某大型机场候机楼的室内时空客流数据上进行了测试，结果显示我们短期预测的均方误差在 5 个单位左右，相对于传统时间序列的回归方法，我们的方法准确度更高，鲁棒性更好。

关键词： 室内定位，移动预测，深度学习，卷积神经网络，数据挖掘，时空预测

Abstract

There are more and more large buildings during rapid urbanization, such as shopping mall, airport and convention center. People spent more time indoors. And the WiFi infrastructure can not only provide WLAN networks but also provide an approach to monitor indoor passenger flow. Accurate and timely passenger flow information is important for the successful of material and human resources management.

In this paper, we review the research progress of the passenger flow prediction technology, summarize some existed problems: ineffectively discriminate indoor movements from outdoor movements in the original WiFi signal strength records. To address the problem, this paper presents a novel approach based on random forest. we also present a method to build training data which inspired by crowdsourcing, and we show that our techniques get high accuracy on our platform.

Besides, existing passenger flow prediction methods, mostly time series methods, mainly focus on outdoors and are still based on the context relation with the past records. However, indoor passenger flow is not only related to the past records separately, but also to the past records of different regions. According to the hypothesis, we present a method integrated convolution neural network and long-short term memory and extract spatial and temporal feature tensor to predict passenger flow of different regions in large building. As a result, we verify the reliability of the method with the dataset from Guangzhou Baiyun International Airport, acquire good performance. The mean square error is about 5. Our method is much cleaner and shows better generalization ability in comparison with other methods.

Keywords: Indoor position, Deep learning, Convolutional Neural Networks, Data mining,

Feature learning

目录

图目录

表目录

第 1 章 绪论

1.1 引言

随着现代城市的发展和商业的日益繁荣，许多大型建筑例如购物中心，交通枢纽，医院，会展中心在城市中拔地而起，它们建筑结构复杂，内部设施丰富而完备，提供了购物、餐饮、休闲娱乐等多方面功能。一方面，这些大型建筑给人们的生活带来了极大的便利，能容纳周边大量人群的生活需要，另一方面，顾客对商场内部的环境及服务要求也日渐提高。同时作为商场、会展中心的等大型建筑的管理方而言，在不断满足顾客服务需求的基础上，充分合理安排利用资源，减少运营投入成为管理者和研究者所关注的重要问题。这个问题的解决首先依赖于对当前室内客流的时空分布有准确的认识，同时也需要对未来室内时空客流可能发生的变化有所准备。

1.2 研究背景与意义

大型建筑不同于室外空间，其室内环境往往对应了多个提供不同服务的区域，其设施和布局都会有所差别，例如商场中的中庭往往比较宽阔，而护肤品等专柜则布置的比较密集。因此大型建筑内部的结构往往比较复杂，展示给用户的是一个丰富的立体结构。随着智能手机及室内 WiFi 的普及，基于 WiFi 的室内定位系统逐渐普及，获取室内用户位置变得更为容易。虽然室内的空间的无线信号更容易被复杂的空间建筑结构所影响，定位精度相较于 GPS 会有所下降。但对于商场等大型建筑的运营方而言，室内 WiFi 系统不仅为用户提供可靠网络服务的同时也为其提供了一个统一，低成本的客流监测方案。其对室内客流的监测基本不会影响到用户的正常体验，并且在实现的过程中也很容易对用户的

位置数据做匿名化处理，从而有效的保护用户隐私。因此通过 WiFi 的室内定位系统获取室内客流的相关研究逐渐成为一个非常有现实意义的课题。

1.3 国内外研究情况

室内时空客流预测往往包含两个问题：如何通过室内部署的 WiFi 基础设施准确记录用户的移动终端数；如何根据室内不同区域客流的历史分布预测未来一段时间内这些区域的客流分布。

对于前者，问题产生的原因在于室内不仅有用户的手机等无线终端，也还存在一些固定的无线接入点设备（Access Point，AP）。另外由于室内无线信道传输的特殊性，室外路过的行人的智能手机若打开了 WiFi 开关，则其发射的无线信号也会被室内的 WIFI 探针所探测到，从而被误判为室内物体。此外一些新版本的手机客户端在探测室内 AP 时会发送带随机 MAC 地址的数据[1]，从而影响室内客流的统计。在这些问题中，室内室外物体判别是最复杂的问题，关于这个问题，国内外主要使用的方法是在用户客户端设备比如手机上实现。

现代用户的智能手机搭载了非常多的传感器，不仅有直接可见的触摸屏幕，摄像头，其内部还搭载了 GPS、光线传感器、距离传感器、电子罗盘、重力感应器、加速传感器和三轴陀螺仪等等。因此从用户手持的智能手机入手判别是否在室内环境是国内外主要的研究方向。微软亚洲研究院的 Zhou 等人提出的 IODector[2] 是其中的代表方案。Zhou 提出的方法是利用大部分手机都会搭载的光线传感器，磁感应器以及移动天线等获取光照强度的变化（室内弱，室外强），接收到的移动蜂窝信号 RSSI 的变化（室内接收到的移动蜂窝信号 RSSI 有一个陡坡的下降），以及磁场信号的变化（室内变化小，室外大）。然后建立状态转移方程，并最终通过维特比算法计算出当前最有可能处在的状态。Zhou 的实验结果显示 IODector 的准确率已经比较高。并且相比于其他方法，IODector 不需要对环境的先验信息有任何了解。因此可以推广应用到大部分的手机无线设备上。除此之外更直接的方法是直接使用 GPS 传感器，由于室内建筑的遮挡，智

能手机的 GPS 信号此时往往比较弱，Blunck 等人对此做了更深入的相关研究[3]，并运用一些简单的机器学习算法例如朴素贝叶斯和决策树就取得了不错的效果。

但在商场等室内场景中，管理方若要统计商场各个区域的用户数，要求用户安装一个用于检测用户处于室内室外环境的应用是不大现实的。在基于 WiFi 基础设施的定位系统中，我们能统一获取的是用户移动终端发射的无线信号强度和数据包数，因此仍旧通过对用户移动终端的信号强度和数据包数的变化进行建模是一种更可行的方案。

对于后面一个问题，即室内时空客流的预测研究，国内外主要的研究方向是室内移动物体的轨迹预测或者室内符号轨迹（Symbolic Tracking）预测[4]，另外一部分客流预测研究往往集中在室外客流的预测，主要方法是时间序列[5]、前馈神经网络[6] 和卡尔曼滤波[7] 等。

室内移动物体的轨迹预测是无线网络下的移动物体轨迹预测的子问题。其重要意义在于若能比较准确预测某个移动物体在无线网络中下一个需要连接的节点，则可优化无线路由选择，从而提高无线局域网的服务质量（Quality of Service, QoS）。这方面的研究以 Pratap[8] 为代表，他研究的对象是以校园内的无线网络为基础，假设移动用户的下一个位置和当前位置以及前一位置相关联，即二阶马尔科夫模型，模型的转移概率可以记为 $P(AP_{next}|AP_{current}, AP_{previous})$。算法的流程是一旦用户离开当前 AP 的覆盖范围，则服务立即计算用户下一个连接的 AP，对应的 AP 则需要做资源保留，因为此时很有可能会有新的 AP 接入。而符号轨迹的预测的目的往往在于分析室内区域用户感兴趣的点（Points of Interests，POI）。

时间序列的客流预测的研究对象一般是某个区域总的客流进行预测，考虑未来值与当前值的关联以及过去的关联或者季节等周期性，最常用的方法一般有线性回归、ARIMA、前馈神经网络、以及基于模糊集的时间序列分析[5][6][7]。这些研究在其各自的领域取得了一定的进展。但从应用场景来说对于商场等大型建筑内部，一个区域内未来的客流不仅与当前客流以及过去客流有关，还与周边客流甚至另外一个区域的一段时间之前的客流相关，也就是不同区域的时

空关联更为紧密。因此不能直接应用现有的模型对商场的客流量进行直接预测。另外对于商场的智能管理系统而言，也不需要对移动物体的轨迹进行预测，因为管理者更关心的是建筑内部某个时间点对应的顾客人数，至于顾客到底是谁，并不十分关心。关注建筑内不同区域的客流人数以及未来可能的客流时空分布对管理者更有意义，对于普通客户而言，位置数据隐私也能得到了很大限度的保护。

1.4 本文工作与贡献

为了完成上述研究目标，本文调研总结了国内外相关的研究成果和方法，在此基础上，结合当前机器学习和深度学习在数据挖掘领域的应用，提出了一些关于室内时空客流数据的清洗技术与预测方法。

用户在室内不同时刻的位置来源于日常携带的手机等无线设备发射的 WiFi 信号被室内的 WiFi 探针（Sniffer）所接收探知，从而在某个时间点可以得到来自多个 WiFi 探针的接收信号强度值 (received signal strength indicator, RSSI)。合并这些信号强度值，我们得到一个信号强度向量 $\vec{v} =< rss_1, rss_2, ..., rss_m >$。根据该信号强度向量与 WiFi 探针的相对位置关联以及 WiFi 探针室内的标准坐标，从而可以根据常见的室内定位算法得到用户此刻在室内的位置 $p_i =< x, y, z >$。该位置不仅包含了相对于建筑物水平面方向的信息，还包括了楼层位置信息。因此对于商场等大型建筑的管理者而言，这类有序的空间位置信息在时间和空间维度上进行聚合即可形成室内不同时间不同区域的客流信息，即室内时空客流。

在统计时空客流时，通常的做法是使用 MAC 地址标识一个用户。然而商场等大型建筑内部有固定 AP 设备，周边有行人的手机设备，这些设备发送的数据包中包含的 MAC 地址都会影响室内客流的统计精度。尤其是室外行人的手机设备，由于室内的无线传播环境比较复杂，如果直接通过观察信号强度向量中值的变化来建模设置确定性的规则会显得非常困难。

决策树是一种比较理想的从数据中提取出其内在规律的有监督学习算法。

它通过信息论里的信息熵以及信息增益来计算数据中变量的相关性。这种相关性不仅包含了线性相关也包含了非线性相关。通过计算变量变化区间能最大化类别的信息增益的点作为切分点来判别数据的类别。这个过程等价于 IODetector 通过观测数据变化设定启发式规则的过程。决策树通过该方法能获取比人工设定阈值更精确的结果。此外决策树不同高度的分支还可以很好的形成一系列组合规则，从而更好的判定结果。此外，决策树在集成学习领域是一种非常优良的基学习器，可以很好通过集成学习拓展成随机森林，AdaBoost，梯度增强决策树（Gradient Boosting Decision Tree）等泛化能力较强的方法。

基于上述思想，我们采用基于决策树的随机森林方法对室内室外物体进行判别。在训练数据采集上，我们采用了“众包”（crowdsourcing）的思想采集室内固定工作人员的 MAC 地址和工作时间等一系列方法，自动化的完成了该过程。相比于精心采集的数据，众包获得的数据训练结果在精度上只有略微下降，但在实现上更简单，并且可以周期性的自动采集。

在室内客流的时空预测上，室内某一区域的客流与它前多个时刻的客流有着非常密切的联系，同时室内空间相比于室外空间要小，因此空间之间的关联也非常紧密，常常会出现入口的客流变化在一段时间后就会影响到另外一个区域下一个时刻的客流。也就是说室内时空客流显示了比较复杂的非线性关系。而深度学习在复杂问题上有非常优异的非线性拟合能力。基于此，本文提出了基于深度神经网络模型预测不同区域的时序客流的方法，并通过对比试验验证该方法的科学性和有效性。

综上所述，本文主要的研究成果有以下几点：

提出了一种基于 WiFi 信号强度的高效室内室外物体判别方法。

通过对 WiFi 探针搜集的室内室外物体信号强度向量及数据包的研究，本文发现室外物体的 WiFi 信号强度及数据包数总体上相比室内物体要低的特征，只是由于室外空间比较大，而室内的 WiFi 探针数量较多。这些规律隐含在多个 WiFi 探针接收的信号强度变化中，直接寻找规律并不容易。因此通过集成学习的随机森林方法从数据中提取规律，从而实现室内室外物体的判别。

提出了一种新的基于循环神经网络和卷积神经网络的室内时空客流的预测模型。

根据同一区域前后之间与过去历史间相关的假设，构建了长短时神经网络，根据不同区域之间客流的相互影响，构建了卷积神经网络，合并该两个网络，将不同区域近期的客流和历史客流作为输入，同时预测多个区域未来某一时刻的客流。

对上述方法模型进行了实验验证。

基于上述方法模型，我们通过搭建室内无线定位系统对室内室外物体判别方法进行了实证研究，并对比了不同训练数据采集方法对最终判别结果的影响，通过实际部署一段时间后验证了该方法的稳定性。而对于室内时空客流的预测模型，我们在国内大型机场的室内时空客流数据集上进行了实验，对比了多种传统方法，验证了我们方法的有效性。

1.5 论文组织

本文共分六章，基于室内 WiFi 定位数据预测室内时空客流分布，组织结构如下：

第一章介绍了本文的研究背景、研究意义、研究目标，以及本文针对研究目标所做的工作和贡献。

第二章对基于 WiFi 定位的室内时空客流处理和预测的相关研究分不同方向进行了综述。

第三章详细描述了室内室外物体判别方法的整体描述，包括数据预处理，分类器的选择，训练数据的采集和特征的构建。

第四章详细描述了室内时空客流预测的场景，以及基于深度学习的客流预测算法，包括数据的预处理，特征矩阵的生成，深度神经网络结构的构建和下一个时刻不同区域的客流预测等。

第五章介绍了本文针对研究目标所采集的数据，实验平台，实验结果，以及对实验结果的分析。

第六章是对本文的全部总结，并对未来的改进及应用进行展望。

1.6 本章小结

本章从国内外的相关研究背景出发，描述了本课题的研究意义。概述了室内时空客流预测时存在的若干问题以及国内外的研究发展现状，从而引出本文的研究目标，并介绍了本文的主要工作和贡献，最后总结叙述了本文的组织结构。

第 2 章 室内时空客流预测的相关技术

在本文中室内时空客流数据来自于基于 WiFi 的室内定位系统。而基于 WiFi 的室内定位系统相比于室外定位系统有其自身的技术特点。因此本章首先介绍室内无线定位的相关技术，包括室内传播环境，室内无线信道特征，定位算法和基于 WiFi 基础设施的室内定位系统的结构，然后介绍了室内时空客流预测中用到的不同数据结构，最后介绍了室内时空客流采集、预处理和预测的相关技术。

2.1 室内无线定位技术

早期的无线定位技术主要基于全球卫星定位系统（GPS），通过接收 GPS 卫星信号强度的方法计算用户在室外的经纬度。GPS 在室外定位应用中取得了巨大的成功，因此 GPS 芯片也称为移动设备上的标准传感器。而相比之下室内环境会受到建筑物的遮挡，接收的 GPS 卫星信号往往很弱甚至没有，导致最终定位的结果会有很大的偏差。而随着城市化进程的加快，人们在室内的活动时间和活动范围也越来越大，室内定位技术将变得越来越重要。

在所有的室内定位技术中，无线定位尤其是基于 WiFi、蓝牙等常见无线设备的定位技术占据了研究领域和应用中的主导。不过室内环境无线电波的传播相比于室外环境更复杂。此外室内无线信道特性对无线信号的传输也至关重要。本节将首先分析室内无线传播环境和无线信道的特性，然后描述了基于无线局域网（Wireless Local Area Network, WLAN）的室内系统及其定位算法，最后在此基础上进一步描述了本文使用的基于无线基础设施的室内定位系统。

2.1.1 室内传播环境

室内无线传播环境往往比较复杂，其原因主要分为两个方面：一方面是室内无线设备例如 WiFi 接入点的传输功率小，覆盖距离近；另一方面是对于大型建筑而言，其室内不同区域的空间结构、障碍物和建筑材料都对无线信号的传播、衍射有不同程度的影响。表2.1 给出了 2.4GHz 频段的 WLAN 穿透不同材料的穿透损耗[1]。

表 2.1 不同建筑材料对 WLAN 的损耗

建筑材料	损耗
红砖水泥墙（15-25cm）	13 ～ 18dB
空心砌块砖墙：	4 ～ 6dB
木板墙（5-10cm）	5 ～ 6dB
简易石膏板墙	3 ～ 5dB?
玻璃窗（3-5cm）：	6 ～ 8dB?
木门：	3 ～ 5dB??
金属门：	3 ～ 5dB?

基于这两个方面的原因，WiFi 无线信号在室内的传输会出现较大损耗和空间上的不均匀，甚至是盲区。另外室内空间也是一个动态空间，例如大型建筑物内会有很多不定期开合的门和窗口，这些都会在很大程度影响接收端的信号强度，另外人作为室内主要的移动物体，其本身也是一个非常重要的干扰源。因为人体构成成分中 70% 是水，而水的共振频率也在 2.4GHz 左右，恰好与常用的 WiFi 信号同频率[2]。

因此，室内无线信号的传播环境十分复杂，如要想要通过常规的参数化模

[1]http://www.hrszyl.com/study/gg99fhb7i95cda8fgjc8bgj7.html

[2]http://blog.chinaunix.net/uid-20525239-id-1654195.html

型方法计算无线设备的信号强度信息几乎是不可能的，只能通过 WiFi 探针等接收设备直接测量接收到的无线信号强度。

2.1.2 基于 WiFi 的室内定位系统

尽管室内无线定位系统存在上述缺点，但由于基于 WiFi 的室内定位系统只需要现有的 WiFi 设施，无需其他额外设备，是一个低成本且适用范围非常广泛的技术方案。随着室内环境逐渐成为 WiFi 主要的应用环境，手机、平板等移动设备的普及，以 WiFi 为代表的无线局域网技术是目前世界上部署最为广泛的室内无线网络基础设施[9]。传统的 WiFi 的室内定位算法主要有接收信号角度定位法（Angle of Arrival，AOA）[10]、到达时间定位法（Time of Arrival，TOA）[11] 和到达时间差定位法（Time Difference of Arrival，TDOA）[12]。然而这三种方法受限于实际应用过程中复杂的室内环境，成熟商用的应用比较少。因此目前国内外的研究者把注意力转移到基于接收信号强度的方法。在这其中应用非常广泛的主要有指纹定位（Fingerprinting-based Localization）和加权质心法（Weighted Centroid Localization，WCL）。

WiFi 信号指纹是指室内某个特定位置 p_i 与多个 WiFi 探针接收到该点无线设备的信号强度向量 $\vec{v_i}$。基于 WiFi 信号指纹的室内定位系统通常需要两个阶段：训练阶段（offline phase）和服务阶段（online phase）[9]。训练阶段，专业人士首先对定位分服务区域 p_i 采集该处的信号强度向量 $\vec{v_i}$，然后将服务区域 p_i 的室内相对坐标 (x, y, z) 和信号强度向量 $\vec{v_i}$ 作为一条关联记录存储到数据库中。这个过程定期的到室内各个区域进行，从而在数据库中形成非常多的区域与信号强度向量的关联记录，这个关联记录集合也被称为信号地图（radio-map）。在服务阶段时，用户发送期间所在位置的信号强度向量 v_i' 即 WiFi 信号指纹到定位服务器，定位服务器通过计算与之最匹配的信号强度向量对应的位置返回给用户，从而完成一次定位服务。由于室内环境的变化，这个过程需要定期进行，以保证信号地图的准确性。

除此之外，加权质心法也是常用的室内定位算法。相比于指纹的方法，加权质心法没有训练数据的采集，但需要知道室内 WiFi 探针的坐标。加权质心法通过计算与“可见锚点”（visible anchor nodes）的质心，从而得到物体的近似位置。“锚点”实际上就是坐标已知的室内 WiFi 探针设备，其坐标一般记为 $a_i = (x_i, y_i)$。如果某个设备在某一时刻接收到了多个锚点的信号，则这些锚点相对于该设备而言就是“可见锚点”，该设备此时的位置可以由公式2.1 得到：

$$\hat{p} = \frac{1}{m} \cdot \sum_{i=1}^{m} a_i, \qquad \text{公式（2.1）}$$

因为实际可见锚点的个数会有多个，并且公式2.1中把所有的可见锚点的权重认为是等同的，但很明显各个可见锚点的权重是不同的，所以通常计算的时候，m 一般采用的是可见锚点集子集中信号强度比较强的若干个点。尽管这样，2.1背后的假设还是显得过于简单，所以有了加权质心法。加权质心在利用锚点的时候会考虑到物体此时接收到的锚点信号强度，最终的定位可以由公式2.2得到：

$$\hat{p} = \frac{\sum_{i=1}^{n} (\hat{d_i}^{-g} \cdot a_i)}{\sum_{i=1}^{n} (\hat{d_i}^{-g})}, \qquad \text{公式（2.2）}$$

其中的 $\hat{d_i}$ 是物体与可见锚点 a_i 的距离，该距离可以通过信号接收强度与距离的关系估算出来。常用的信号强度与距离的关系可参阅相关文献[13]，指数 $g > 0$ 决定了不同锚点的重要性，增大 g 将增强最近可见锚点的权重。加权质心法的精度与室内 AP 部署的个数与位置密切相关，一般来说部署的个数越多其定位精度就越高[13]。

2.1.3 基于无线基础设施的室内定位系统

上一节提到的室内定位系统默认定位在手机客户端完成，也就是基于无线终端（client-based）的主动定位。但实际应用部署的过程中，这样的行为系统会面临两个大的问题：1、首先若要实现室内定位，需要在用户的智能手机上安装

定位应用或登录定位的服务 Web 页面；2、其次用户在室内主动发起定位的情况非常少。据国内某商场的 APP 所有的功能模块使用频率统计，室内导航部分的使用频率不到 3% [3]。因此期待用户在手机客户端主动发起定位从而获取室内的客流数据将远低于实际值。由此基于室内无线基础设施（infrastructure-based）的定位系统应运而生。

基于无线基础设施的室内定位系统的基本思路是在室内空间中部署多个 WiFi 探针（sniffer），通过探测收集智能手机等无线设备客户端发射的无线信号强度。WiFi 探针可以侦测到无线客户端发射的管理帧，控制帧和数据帧，并从中解码得到无线客户端的 MAC 地址。因此对于某个时刻内可以得到无线客户端来自多个相对于 WiFi 探针的信号强度集合 $(MAC, rss_1, rss_2, ...rss_n)$，这些值以日志的形式统一发送到定位引擎服务器上，则可使用上节提到的定位算法实现位置计算。

基于上述过程，商场等大型建筑的管理方只需在商场中部署 WiFi 探针和统一的定位引擎服务器即可获取到室内所有打开 WiFi 开关的无线终端设备，从而间接获取到室内的客流数据。这种定位系统也被称为被动定位系统，其主要优势有两方面，首先是部署和管理非常方便，WiFi 探针在很多企业级的 AP 上都有搭载，可以在为用户提供免费 WiFi 服务的同时搜集用户在室内的位置; 其次，被动定位不需要用户主动连接无线 AP，只需要打开手机上的无线开关即可检测到用户以及对应的 MAC 地址，因此系统最终得到的客流数据会更接近真实数据。这里需要指出的是在 WiFi 定位系统中，总体上基于无线基础设施的方法精度上是要弱于基于无线终端的定位方法，因为基于无线终端的定位方法是手机主动发送信号，此时手机信号更稳定和可控。其详细对比数据可以参照有关文献[14]。因此在目前的实践应用中往往会将二者结合，在对室内客流进行统计分析时使用基于无线基础设施的方法，而当用户需要室内导航的时候可以采取主动定位的方法。本文若无特别说明，后续用于客流统计和预测时原始数据的采集方法均为基于无线基础设施的方法。

[3]http://mt.sohu.com/20150911/n420894958.shtml

2.2 室内时空客流预测的基本描述

由于商场等大型建筑内免费 WiFi 和用户智能手机的普及，获取室内用户的位置变得更为容易。通过基于 WiFi 的室内定位系统统计分析室内客流的时空分布，预测商场未来一段时间内的客流时空分布将变得越来越有意义。

2.2.1 室内时空客流相关数据的结构

室内无线定位技术的进步产生了大量的室内移动物体轨迹信息。通常情况下，上文提到的室内定位系统通过 WiFi 探针探测到手机等无线客户端信号生成 WiFi 信号日志，然后通过 WiFi 信号日志抽取得到各个 WiFi 探针对应该设备的信号强度向量 $< rss_1, rss_2..., rss_m >$，然后通过定位算法计算移动物体此刻在室内的位置。室内定位引擎接收到 WiFi 探针扫描到的无线信号强度信息后会在 10~60 秒左右产生一次定位日志。相比于室外环境的定位频率，室内的定位频率往往会低不少。这是由于人们在室内活动范围相较于室外会更小，移动的速度和频率也较低，大部分的时间处于静止的工作或等待状态。另外一方面也是基于减小室内定位引擎的服务压力的考虑，实践过程中商场、医院或机场等大型建筑的管理方会根据需要调整其定位频率[15]。

通过上述过程，最终我们将得到用户在室内不同区域一系列有序的定位点组成，即用户进入室内后的一系列轨迹点，例如，$a_1 \rightarrow a_2 \rightarrow ... \rightarrow a_n$。这些点往往由包含相对于室内的空间坐标和时间戳的四元组构成，$a_i = (x, y, z, t)$。其中的 z 对于多楼层建筑而言用于标识楼层，而对于单楼层而言其值往往为 0。下面我们结合图2.1给出室内时空客流预测中的相关概念定义。

（1）WiFi 信号日志：WiFi 信号日志是最原始的数据，一条 WiFi 信号记录是一个包含五个元素的元组，包含了客户端的 MAC 地址，探针的 ID，信号强度，发包数，探测到的信号的时间戳，记为 $R =< MAC, ap, rss, packets, timestamp >$。WiFi 信号记录是室内定位中的关键数据，许多预处理和数据的清洗工作都需要通过原始记录完成。

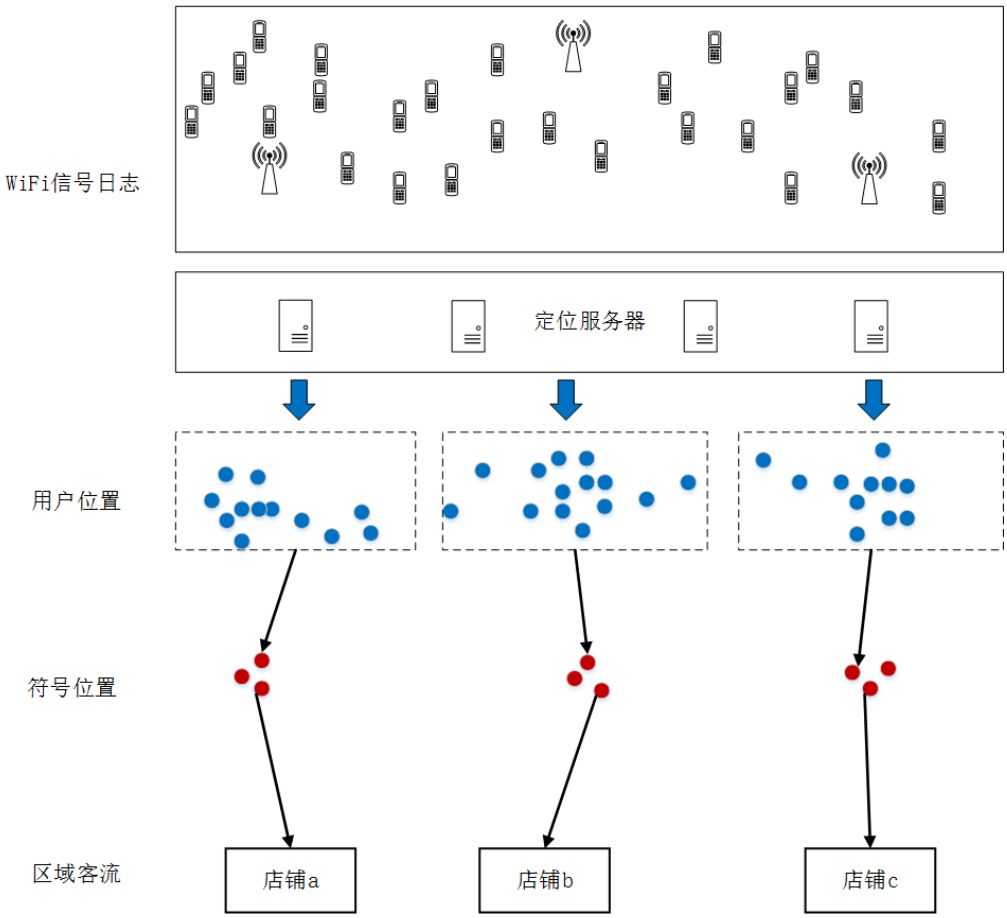

图 2.1　WiFi 信号日志、用户位置、符号位置和区域客流

（2）信号强度向量：信号强度向量是从 WiFi 信号强度日志中提取出来的，包含了某个时刻一系列 WiFi 探针探测到的信号强度及发包数，记为 $\vec{v} =< rss_1, packets_1, rss_2, packets_2, ...rss_n, packets_n >$。

（2）用户位置：用户位置是室内定位引擎根据定位算法和信号强度向量 $\vec{v}$ 计算出的位置点，包含了用户的 MAC 地址，室内坐标和时间戳，记为 $p =< mac, x, y, z, t >$，其中 (x, y) 是相对于室内空间的水平坐标值，z_i 在多楼层的建筑中对应的是楼层标识，对于单楼层来说一般记为是 0。若将一个用户位置按时间序列排列，则可得到用户在室内的时序轨迹 $Track =< p_0, p_1, p_2, ..., p_k >$。

（3）符号位置（Symbolic Position）：符号位置[4][15] 是对空间区域进行建模

的常用方法。其步骤是先对室内空间进行划分，从而用户在不同的时刻的位置点被划分到不同的区域，这些区域用一些代号或名字命名，之前的用户位置里的坐标信息就被转换为一系列标区域代号，即 $S =< mac, region, t >$。其中 $region$ 代表室内空间的某个区域。符号位置在室外地理模型方面用的最多的例子是邮政编码，一个邮政编码代表了地理上一块区域。

（4）区域客流：客流是按照时间段对符号位置内进行聚合，对区域内的用户 MAC 进行独立计数，从而得到某个时间段内某个区域的用户数量，即区域客流，记为 $Flow =< region, count, t >$。大厦内把不同区域客流按照时间先后顺序进行排列，则可以得到商场内不同区域不同时间段客流的分布，即我们关心的商场客流时空分布，常见的可视化结果为热力图的流动变化，如图2.2

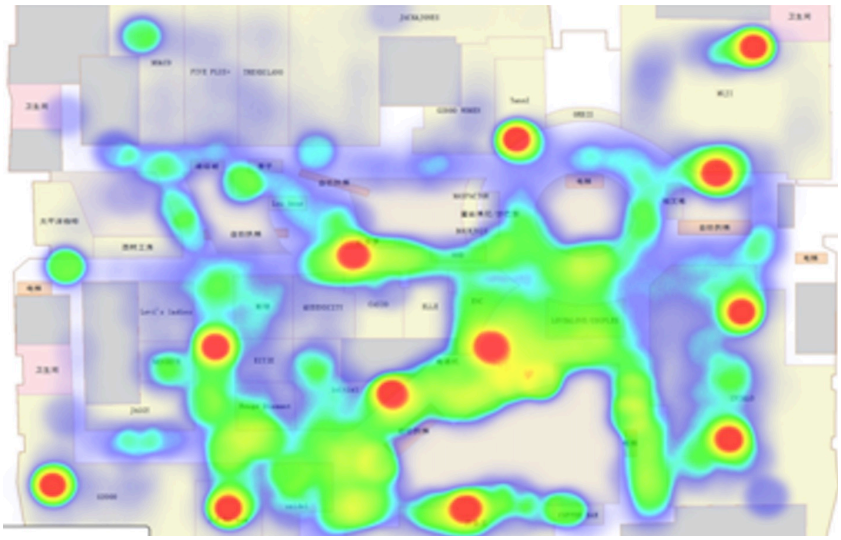

图 2.2 某商场客流热力图

2.2.2 室内时空客流预测的基本过程

传统的客流信息是商场或公交站点中不同时间到访的顾客总数的统计，反映的是商场内总的顾客数量随时间变化的一个分布。而室内时空客流信息是室内空间不同区域客流数量随时间变化的一个分布。其变化因素不仅随时间变化，

也跟具体的空间拓扑结构相关，因此预测难度更大，对商场管理者更具有建设性意见。不同于普通客流预测过程中的步骤，室内时空客流的预测一般包含如下步骤：

1）噪声过滤等 WiFi 数据清洗

由于室内空间环境的复杂性以及 WiFi 信号传输的特点，原始的 WiFi 信号数据不可避免的会带有部分噪声数据，例如在室内的 WiFi 探针容易探测到来自室外行人手机发射的无线信号，部分无线设备发射的信号会带有伪 MAC 地址等。这些噪声数据不仅会对室内客流统计带来比较大的影响，而且会引入未知因素影响我们对时空客流的预测和分析，得到不正确的结论。

2）室内定位

经过上述清洗过滤后的 WiFi 信号数据，定位服务器会将同一时间内的 WiFi 信号数据进行归并，形成以 MAC 地址和时间戳为标识的信号强度以及发包数的特征向量，然后使用指纹定位或加权质心等定位算法计算该时刻此 MAC 对应的室内空间坐标。在这个过程中，若用加权质心法，有时得到的位置可能位于不可达位置，例如商场某根柱子或柜台上，此时需要对室内空间建模，把此时的位置纠正到最近合法的可达的位置上。对于某个时刻由于信号强度数据不足无法产生定位的点，可根据实际需要舍去或进行插值。经过定位后，原始的 WiFi 数据往往会被定期压缩转存或直接清理，因为这部分的数据量往往会非常大。

3）客流统计

经过室内定位后，我们可以得到不同时刻用户在不同位置的坐标点，这些坐标点直接拿来使用往往会显的没有规律，我们更期待这些坐标值有具体的含义。因此在做客流统计之前首先会将这些坐标值按照室内空间进行划分分配到不同有意义的区域，形成符号位置。最后对符号位置同一时间段内的 MAC 进行独立计数，最后得到商场内不同区域不同时间段内的客流，即商场时空客流。

4）预测室内的时空客流

通过前面几个步骤，我们将原始的 WiFi 信号强度信息转化为商场历史的时

空客流序列，接下来我们需要构建自己的模型，结合商场的空间结构信息和商场内的事件等信息，预测接下来一段时间内商场的客流时空分布。

以上四条是商场等大型建筑内时空客流预测的基本步骤，室内时空客流的预测基本基于上述思路展开。每一个步骤都涉及到很多相关的技术，尤其是噪音过滤中如何判别无效的信号数据、室外无线设备的信号，以及预测模型的构建等。

2.3 相关技术概述

本节介绍客流预测领域和深度学习相关的一些技术，主要如下：室内客流的获取、数据的预处理、时空客流的预测技术和深度学习相关技术。

2.3.1 室内时空客流数据的采集

数据搜集是研究人员运用数据驱动方法（Data-Driven method）解决实际问题过程中第一步，数据搜集质量的好坏直接影响到最终的结果以及结论的准确性。数据搜集需要研究人员根据具体问题，分析所需数据格式、内容，设计收集方法，并对其中的噪声加以清洗，缺失数据加以填充估计等等系列步骤，最终得到规范化数据用于后续模型使用。

室内客流数据的搜集可以有很多种形式，最传统的方式是通过人在入口进行计数，例如电影院、机场的安检、车站等有固定入口的场景。这种方式虽然得到数据比较准确，但难以实现自动化，并且能获取的信息仅限于一部分区域，比如入口。而用户进入区域后在室内的游走等行为则无法有效的监测和记录。在总客流的时序宏观预测（按天、按月等较大时间尺度）上，这种方法往往可以作为的客流预测数据的标准值[16]。随着室内各类无线传感器的部署，室内客流的实时获取成为可能。常见的传感器主要有蓝牙、WiFi 和摄像头，蓝牙和 WiFi 的原理基本相同，都需要用户打开手机上的蓝牙开关或 WiFi 开关，从目前来看打开

WiFi 的用户数量要远多于打开蓝牙开关的数量。近年来图像识别和理解技术的进步，越来越多关于行人检测（Pedestrian Detection）[17][18] 的方法得以应用，在客流数据收集上可以取得比 WiFi、蓝牙等更好的效果。不过从目前来看，缺点也比较明显，首先是成本比较高，其次若要对客流进行进一步理解分析单个用户行为时会比较困难。因为通过视频的方式取得连续的用户室内轨迹会比较困难，而蓝牙和 WiFi 可以利用 MAC 地址等物理变量精确的标识单个用户，一定时间内用户多次进入商场都可以识别出来。基于以上原因，目前室内监测客流主要以 WiFi 为主。而基于 WiFi 的方法在于室内复杂的无线传播环境和客户端的发送 WiFi 数据等特点，得到的客流数据会有噪声，因此对 WiFi 得到客流数据需要做必要的预处理工作。

2.3.2 室内时空客流预处理技术

在室内移动物体轨迹预测的过程中，我们常常需要对原始数据进行过滤、平滑和插值，而室内客流的统计往往更注重区域客流计数时的准确性，所以在这个过程中最重要也是最主要的预处理技术是噪声数据的过滤。常见的过滤技术有伪 MAC 地址，室外行人的过滤等。

直观上，用户手机上的 WiFi 开关打开后会有如下三个阶段：1、扫描阶段，发现无线接入点 AP，常用的扫描有主动扫描和被动扫描；关联阶段，和相应的无线接入点建立关联，手机和相应的 AP 进行一次握手协议，这时候 AP 会关联到这个手机的 MAC 地址，为下一个阶段数据包的传输做准备；3、传输阶段，手机进行数据上传和下载。对于这三个阶段，室内的 WiFi 探针都可以检测到这个过程中发射的探测帧，控制帧和数据帧，将其中的 MAC 地址作为用户标识，信号强度和发包数用于确定用户位置。在客流获取的过程中，我们很大一部分数据是通过检测探测帧后得到的。然而随着人们对隐私保护的目的，部分手机厂商会在扫描阶段把探测帧中使用随机的 MAC 地址。若不将这部分的数据过滤将会很大程度上影响我们的客流统计精度。

室外行人在经过商场等大型建筑时，若其无线开关处于打开状态，室内的 WiFi 探针将会接收到这些用户手机发射的探测信号，从而将接收到该手机的信号强度信息发送到定位服务器。由于大型建筑的面积往往比较大，其周边的行人数量也不在少数，若不对这部分设备判别出来将会影响到我们的客流统计结果。关于这个问题，来自新加坡南洋理工大学的 Zhou[2] 提出了一个通用的高精度室内室外检测方案 IODetector。该方法主要通过手机上搭载光线感应器、磁感应器和移动蜂窝信号强度三种传感器对室内室外环境进行判别。为了更好的判别室内和室外环境的转换，Zhou 还增加了一种状态半室外（semi-outdoor）用于描述室内到室外之间的转换。Zhou 的理念主要是室外的光源由于是自然光源，白天光照强度大，晚上光照强度小，变化浮动比较大，而室内光源通常比较稳定，光照强度位于室外白天和晚上的光照强度，因此通过比较室内室外光强可以得到一个室内与室外状态的置信概率 $[D_L, C_L]$。室内建筑由于建筑材料中使用了钢材和其他金属，因此磁场变化比较大，而室外即使在南北两极和赤道上的差别也比室内的磁盘变化要小，因此通过磁场的变化又可以得到一个室内室外的置信概率 $[D_M, C_M]$。此外，手机的天线在进入室内后加大发射功率，从室外进入到室内会有一个移动蜂窝信号的增强，而从室内到室外后，会相应减小发射功率，从而会有一个移动蜂窝信号的减小，这样又可以根据移动蜂窝信号变化得到一个室内室外的置信概率 $[D_C, C_C]$。最后将这三者的置信概率相加，从而得到室外、半室外和室内三种状态的置信概率 $C_E \in C_{indoor}, C_{semi-outdoor}, C_{outdoor}$，在其中选取置信概率最大的那种状态作为最后的判定结果。此外，由于室外、半室外和室内之间的状态变化是个连续的过程，因此文中还考虑了引入隐马尔科夫模型进行建模，基于观察数据他们定义了各种状态之间的转移概率（Transition Probability）：

1）$T(S,I) = T(S,S) = T(S,O) = p_1 = \frac{1}{3}$

2）$T(I,I) = T(I,S) = p_2 = \frac{1}{2}$

3）$T(O,O) = T(O,S) = p_3 = \frac{1}{2}$

4）$T(O,I) = T(I,O) = p_4 = 0$

用原始的三种传感器得到不同状态的置信概率做初始发射概率（Emission Probability），最后通过维特比算法（Viterbi algorithm）计算手机所处的状态。Zhou 的方法初衷是通过检测室内室外状态从而关闭 GPS 等其他传感器从而用于节省电量，最终在几款手机上得到了大约延长 30% 的使用时间。Zhou 的方法需要在手机上安装应用软件。对于基于客户端的主动定位方式而言是可行的，然而实践过程我们使用的是基于基础设施的被动定位，因此不能直接使用。尽管如此，Zhou 的方法还是给我们有所启示，主要方向还是基于信号强度变化进行建模对室内室外进行检测。

2.3.3 客流的预测技术

准确的客流预测对现代智能管理系统有着重要意义。国内外相关的研究主要集中在室外相关的客流预测上。研究方法上主要把客流预测看成是一个时间序列预测的问题，例如北京交通大学张春辉关于公交站点的客流预测[7] 和中国铁道科学研究院的关于铁路客流的预测[6]。国外在这方面的研究起源很早，尤其是在上个世纪 90 年代广泛应用的高级交通管理信息系统（Advanced Traffic management and information systems，ATMIS）后，各种相关的客流预测得到了比较全面的研究，应用的方法也比较广泛。其中最早的相关描述在 1997 年总结了客流预测问题的规范定义以及给出了当时主要的几种流行的方法的比较[19]。

客流预测问题的一个规范化定义为：已知当前时刻的客流状态 $V(t)$、当前过去一个时刻的客流 $V(t-d)$、历史相同时刻客流状态 $V_{hist}(t)$ 和历史相同时刻下一个时刻的客流状态 $V_{hist}(t+d)$，求当前时刻下一个时刻的客流状态 $\hat{V}(t+d)$。Demetsky 的这个定义是针对当时普遍的交通管理信息系统中的问题进行定义，其中的 d 是指时间间隔，即预测区间是以一个时间范围为单位，在他的论文中，他使用的是 15 分钟。因为对于一个交通站点而言，客流的变化比较大，直接预测会比较困难，所以当时大部分的系统都把预测目标转换为下一个时间范围的均值。这样可以明显降低预测的难度并且不会对实际的应用造成的太大的

影响。在后续的相关研究上，大部分的客流预测问题基本延续了这一思想，只是预测的区间发生了变化，一般要求当前时间接下来连续的多个点的值，即 $V(t+d), V(t+2*d), V(t+3*d), ..., V(t+m*d)$，并且按照 m 值的大小分为短时预测和长时预测，长时预测一般都会选择 $m >= 20$。另外随着研究范围的扩大，在地图上一般还会研究多个临近站点之间客流的关系，从而在预测中引入了空间的因素，形成了时空客流预测的原型。

在预测方法上 Demetsky 总结了当时主要的四种方法：历史同一时刻均值，统计时间序列模型、神经网络和非参数回归。其中统计时间序列模型 Demetsky 主要用的是经典的自回归求和移动平均模式（Autoregressive Integrated Moving Average，ARIMA）。在这四种方法中，历史同一时刻均值是将历史上下一个时间范围的均值直接作为均值即 $\hat{V}(t+d) = V_{hist}(t+d)$。例如要预测某站点今天 15:00~15:15 的客流，历史均值法会把该站点历史上所有 15:00~15:15 之间的客流进行平均作为预测结果。实际上基于历史均值的方法是一种静态的预测方法，无法对动态或未知事件进行预测，一旦下一时刻出现了不同于历史的事件，则无法得出比较准确的结果。ARIMA 是统计时间序列预测时使用的经典方法，ARIMA 的缺点是要求历史的客流时序记录都是连续的，并且经过差分运算后序列是平稳的，否则则不能应用 ARIMA 方法进行计算。而一般来说对于一个交通站点的客流，总的客流来说一般并不是平稳的，有季节、天气以及节日等等相关的因素。因此在 Demetsky 中的结论中，ARIMA 的效果往往比较差。而对于时空客流，由于不同区域之间的客流随着时间相互影响，体现了一定的微观性，ARIMA 的结果可能会更差。此外人工神经网络由于可以任意逼近一个非线性函数，因此在客流预测中也常常使用。神经网络是一种黑箱方法，训练过程往往比较复杂。对于传统的客流预测上效果需要比较多的调参工作才能得到比较令人满意的结果。在当时所有的方法中，基于非参数的回归方法的预测效果是比较好的，非参数的回归方法的思想与最近邻的思想是相同的，它的过程可参考算法1：

算法的过程是扫描历史数据中与当前时间段内客流数据变化最相似的 k 条记录，然后用这 k 条记录各自后一个时刻的值的均值作为预测值。非参数方法

算法 1 (非参数回归算法)

1: At time t, given the state of the system, $X(t)$:
2: Initialize list of nearest neighbors, NB, to contain cases $1, 2, ..., k$ of the development database D
3: **for** each element $c \in D$ **do**
4: Calculate $DISTANCE(X(t), X_c)$
5: **if** $DISTANCE(X(t), X_c) < MaxDISTANCE(NB)$ **then**
6: Remove the element which has max distance from NB
7: Add c to NB
8: **end if**
9: **end for**
10: Estimate $V(t+d)$ as: $V(t+d) = \frac{\sum_{q \in NB} V_{fq}}{k}$

与聚类算法类似，它需要定义比较好的距离测度，才能够比较好的预测未来的结果，同时它也需要比较丰富的历史数据。

上述方法主要把客流的预测当成是一个时间序列的问题，因为其研究对象往往只是一个交通站点的总体客流，并不考虑其内部客流的分布。而人们的生产活动有比较明显的周期性：白天工作，晚上休息；工作日工作，周末休息；对于大部分人来说，四季分明，不同的季节的作息会有相应的影响。因此从客流的数据上来看，这些周期性的规律都比较好的体现在了客流的波动上，在后续的研究中 Williams[20] 在预测未来多天的交通客流中使用了基于季节模型的 ARIMA，把平均绝对百分误差率（Mean Absolute Percentage Error，MAPE）从 11.5% 降低到了 8.6%。刘建军[5] 在对南京某商场基于 WiFi 统计的小时客流预测过程也引入了季节模型，在多个评测指标上都取得了更好的结果。

随着研究对象的扩大，人们对客流预测的研究不再局限到一个地点，逐步拓展到多个相邻地点的客流预测，在这种情况下，地理位置信息也是对客流一个非常明显的影响因素。徐薇[6] 等人提出了利用神经网络对北京附近的 8 个地点 10 年春节前后 40 天的历史数据建模。站点之间的关系直接通过神经网络的输入

节点进行学习，最终把平均相对误差率从 1.55% 下降到了 0.97%。此外 Zhang[21] 也在客流预测中不仅引入地理信息，还进一步引入了一些特定的外部信息，例如交通事故的数量，天气，道路限速等，这些对一些特定的区域的客流信息预测是有帮助的。

综上所述，我们可以看出对时空客流的预测将逐渐成为客流预测问题的焦点。尤其是对室内客流而言，由于客流相比于室外更为密集，区域之间的关联更为紧密。及时预测出未来客流的时空分布，对提升商场等大型建筑的管理水平，防止踩踏等突发事件的发生有着重要意义。在预测方法上，神经网络和非参数的回归方法会有更高的准确率，这也是本文方法的基本出发点。

2.3.4 深度学习相关技术

深度学习是机器学习研究中的一个崭新的领域，传统的机器学习方法往往是浅层学习[22]，例如常见的支撑向量机（SVM，Support Vector Machine）、随机森林、逻辑回归、前馈神经网络等等。这些模型的结构基本可以看成是带有一层隐含节点的方法或者没有隐层节点（逻辑回归）的方法。这些模型在理论上和应用中都获得了一定的成功，但在实践过程中面对比较复杂的任务时准确度和泛化能力往往有限，并且需要耗费大量的精力进行特征的构建和筛选等。Hinton[23] 在 2006 年时指出解决这类问题的方法是深层次的神经网络，即深度学习。深度学习相比于“浅层”学习能够自动学习特征的分布式表达，具有更好的全局和局部特征表达能力。同时深度学习的中间层表示可以作为特征表达应用在不同领域中，应用范围非常广。最重要的是深度学习相比传统的前馈神经网络能在更少的数据样本上学习更深层次的抽象化表达，产生的特征更为有用。在实践过程中，卷积神经网络和循环神经网络是处理图像和时序模型中最为常用的方法。

2.3.4.1 卷积神经网络

卷积神经网络（Convolutional Neural Networks，CNNs）是近年来深度学习领域非常流行和常见的深度网络。卷积神经网络首先在 2012 年的图像处理领域取得了巨大的成功[24]，从而引起了学术界极大的兴趣。卷积神经网络相比于传统的神经网络最大的特点是它卷积核结构中的权值共享（Shared Weights）。这种结构来源于早期对生物神经网络的研究。权值共享结构通过减少权值数量可以有效降低深度神经网络模型的复杂度。这种网络结构对于多通道的图像处理任务时显示了很大的优势，因为卷积神经网络可以将多维图像直接作为输入，避免了传统图像识别领域中复杂的特征提取过程，例如 SIFT[25] 等。

除了权值共享外，卷积神经网络另外一个非常重要的特性是稀疏连接性（Sparse Connectivity）。稀疏连接性能使卷积神经网络在神经元中实施局部连接提取输入空间中的局部性。例如第 m 层神经元的输入来自第 $m-1$ 层 k 个神经元。我们可以通过图2.3 来表示：

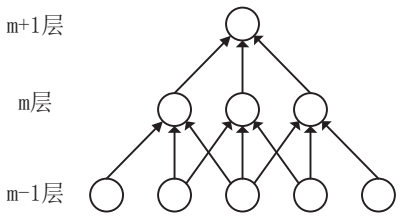

图 2.3 CNNs 的稀疏连接

从卷积神经网络的生理学基础上而言，第 $m-1$ 层是输入视网膜，第 m 层的神经元有大小为 $k=3$ 的感受区，因为 m 层的神经元只连接视网膜层中 3 个局部相邻的神经元。第 $m+1$ 层的神经元与 m 层的神经元有相似的连接。这种结构保证了下一层的神经元对一个空间的输入图像有局部最强的响应。而在 $m+1$ 层的神经元收到信号可以形成 $m-1$ 层神经元全体关联，也就是随着层数的堆叠，卷积神经网络可以获取全局化的非线性学习能力。在 CNNs 中，每一个

卷积核共享同一组权值，这样该卷积核经过一层映射后可以形成一个特征映射（feaature map），传统的梯度下降算法可以很容易学习这种共享参数。卷积核的权值共享减少了需要学习优化的参数，这样的设定，使得 CNNs 在视觉和图像识别领域应用非常广泛。

上述两个直观上的概念对应了卷积神经网络中最重要的两个层: 卷积层和池化层（Pooling Layer）。其中共享权值主要指的是卷积操作。卷积中引用的是数学上的概念，对于定义在 t 上的离散输入函数 x 和卷积核函数 w，其卷积为2.3：

$$s(t) = (x * w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a), \qquad \text{公式（2.3）}$$

卷积层是卷积神经网络的核心，每个卷积核 w 是一个可以学习的过滤器，通过卷积操作，从而得到某一维度上的特征，卷积核可以一次定义多个，输出到下一层的结果不断叠加，成为特征映射。同样卷积可以定义在输入是二维的函数及卷积核上，其定义为2.4[4]

$$S(i,j) = (K * I)(i,j) = \sum_{m}\sum nI(i-m, j-n)K(m,n), \qquad \text{公式（2.4）}$$

这样卷积操作不仅可以生成一维的局部特征，还可以同时生成二维的局部特征。这种操作可以在下一层再继续加入卷积核，从而形成多层次的卷积层。在图像识别[24] 上，这样操作的结果在图像识别中会产生类似于点、线、面等不同层次的特征效果，对图像的识别带来非常大提升。

稀疏连接性往往指的是卷积中的池化层。卷积神经网络的池化用的最多的是最大池化（Max-pooling），一个非线性的降采样形式。最大池化就是将输入图像分割为一系列不重叠的矩阵，然后对每个子区域，输出最大值。最大池化在视觉相关的研究中应用明显[24]，通过消除非最大值，减少了更上层的计算量，同

[4]http://www.deeplearningbook.org/

时提供了一种平移不变性。一个最大池化层一般级联在一个卷积层之后。最后再接上一个全连接的层用于预测目标函数。其整体结构如图2.4：

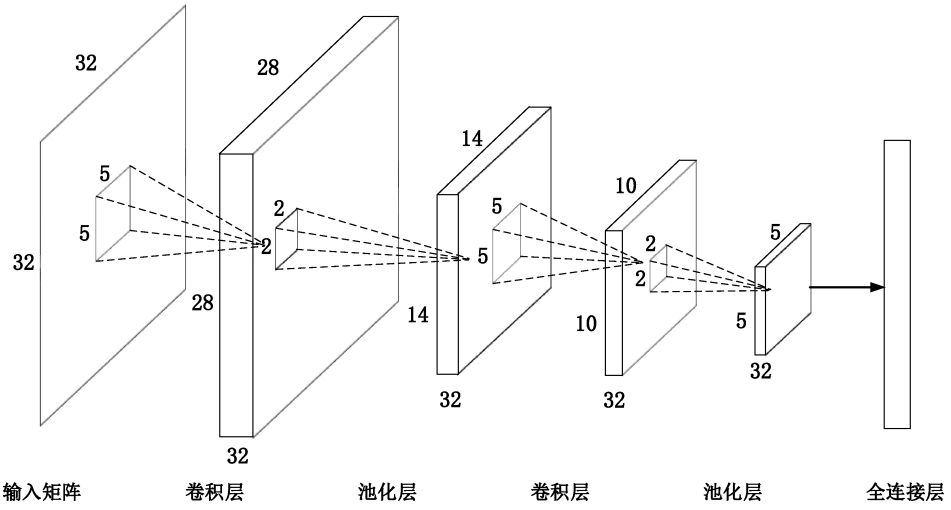

图 2.4 卷积神经网络整体结构图

2.3.4.2 长短时神经网络

卷积神经网络在处理图像语音等问题上发挥了很大作用，但是在自然语言处理以及时间序列相关的问题上发挥的作用要相对小。另外一种更适用的深度神经网络是循环神经网络。循环神经网络是一种带环的神经网络，这些环展开后会形成输入前后的链式结构，从而揭示了序列前后之间的相关性，如图2.5。

在所有的循环神经网络中有一种特殊类型–长短时神经网络（Long short-term memory，LSTM）[26]，它可以学习到时间序列上的长期依赖信息。在神经元的构建上，长短时神经网络中单个神经元引入了“输入门”（input gate）、“遗忘门”（forget gate）、“输出门”（output gate）三种门结构，如图2.6

通过对神经元中这三种门结构的精心设计，LSTM 默认能学习时间序列中长期的变化趋势。

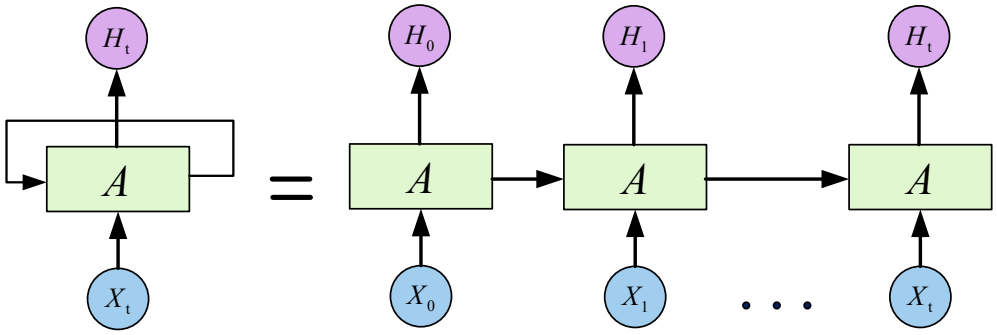

图 2.5 展开的循环神经网络

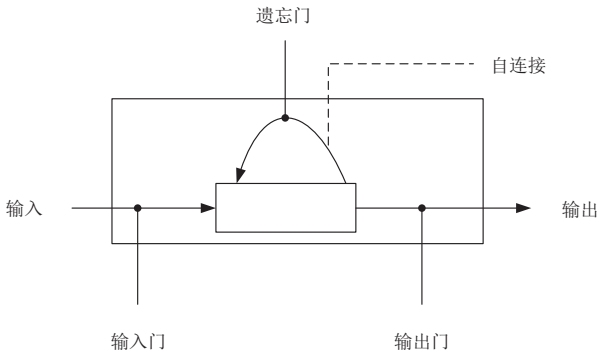

图 2.6 LSTM 神经元结构

2.4 本章小结

本章首先介绍了室内无线定位的相关技术，分析了室内相比于室外环境有着更复杂的传播环境，接着介绍了常见的两种定位算法：基于指纹的定位算法和加权质心法，并引出了在实践过程中更具可行性的基于基础设施的室内定位系统，然后描述和定义了室内客流相关的数据结构，以及室内客流时空预测的基本过程，再详细描述了国内外关于室内客流数据的获取，预处理技术以及常见的客流预测技术。最后介绍了深度学习的技术特点，以及在图像和自然语言处理方面的应用，从而为后续使用深度学习方法预测室内空客流做铺垫。

第 3 章 基于机器学习的室内时空客流统计

在上一章中我们介绍了由于无线信号的特性在统计室内时空客流数据时需要做必要的预处理。其中影响最大的是室外行人无线设备的过滤。因为室外行人的无线设备的信号强度影响因素比较复杂，其中包括室内 WiFi 探针的特性，室外行人的用户设备，行走速度，与 WiFi 探针的距离等。若直接对室外行人的无线信号强度进行建模得到若干确定性的规则，则显得比较困难，尤其是室内环境定期可能会发生相应的变化。为此，本章首先介绍我们在实验室中搭建的基于 WiFi 基础设施的室内定位数据平台以及后续算法中需要用到的数据，然后提出并详细描述基于机器学习的方法对室内室外物体进行有效识别。

3.1 无线信号数据的获取

在描述基于机器学习的室内室外物体判别算法之前，我们先介绍我们在实验室搭建的室内定位系统以及其产生的原始数据。该系统是典型的基于基础设施的室内被动定位系统，与我们调研的一些大型购物中心中使用的室内定位系统的工作原理相一致[1]。如图3.1:

在 WiFi 室内定位系统中，首先需要将搭载 WiFi 探针的多个 AP 部署在实验室的各个位置。由于手机等无线设备的 WiFi 开关一旦打开（并不需要连接上对应的 AP），手机便会自动周期性的发送广播信号寻找附近的可连接的 WiFi 接入点，其中手机发射广播帧的频率与手机操作系统厂商的实现有关。因此在上述系统中，实验室的工作人员以及经过实验室的其他人员进入我们的 WiFi 探针探测范围后，则可以被我们的探针探测到其设备发射的广播帧信号。

[1]http://www.barcodegiant.com/motorola/part-com-adsp-std.htm

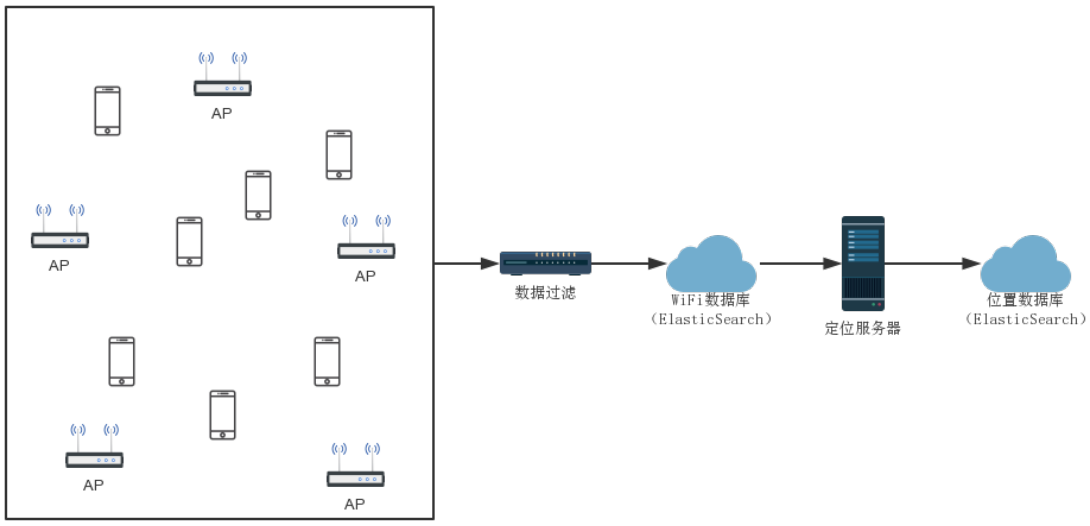

图 3.1　室内定位系统结构图

表 3.1　WiFi 信号强度记录

mac	ap	rss(db)	packets	timestamp
00:26:82:84:b3:d9	8	-59.33	21	1463673841486
00:26:82:84:b3:d9	6	-61.89	9	1463673842029
00:26:82:84:b3:d9	4	-58.3	10	1463673842028
94:39:e5:2d:9c:87	6	-75.94	19	1463673846100
38:bc:1a:35:07:ac	4	-81	2	1463673847357

探针将探测到的手机 WiFi 信号强度和数据包数经过过滤程序过滤掉非法的 MAC 地址和室内的固定 AP 设备后会实时发送到 WiFi 记录数据库中，形成一条信号强度日志记录，其数据结构和样例数据如表3.1，主要包含了用户设备的 MAC 地址，接收信号的 AP 探针标识，信号强度，接收数据包数和探针扫描到设备信号的时间戳。定位服务器定时（一般为 5~60 秒）读取 WiFi 数据库中的

信号强度数据，将相同 MAC 地址在一段时间内的信号强度数据聚合在一起可以得到某个时刻不同位置 AP 探测到该设备的信号强度向量。根据 AP 探针的室内坐标和定位算法（这里我们使用加权质心法）定位生成定位数据存入位置数据库。在这个过程中，数据包数变量往往被许多室内定位算法所忽略不用。

3.2 基于机器学习的室内室外物体识别

在第二章中我们描述相关研究时提到过直接识别室内室外物体是一个比较困难的问题。首选我们的定位系统是基于 WiFi 基础设施的，因此我们不大可能使用像 IODector 那样的方法利用用户手机的传感器进行判别。我们能使用的数据只能是 WiFi 探针探测到的信号强度及发包数的信息。直观上我们可以得到的是 WiFi 探针接收到室外物体发送的 WiFi 信号会比较弱，而室内物体收到的 WiFi 信号强度会比较强。基于此，我们可以为通过 WiFi 接收到的信号数据强度进行建模。然而本文第二章曾介绍过室内 WiFi 信号传输的环境非常复杂，如果我们使用启发式的规则来确定各个 WiFi 探针接收信号强度的阈值，从而推导规则得出物体是否在室内则会显得非常困难。尤其是 WiFi 探针数量变的更多的时候，大型建筑的边界范围很大时，例如我们调研的机场和商场中，WiFi 探针的数量都达到了 500 个以上，建筑面积都在 10 万平米左右[2]。

基于上述分析，我们可以推出，我们需要从复杂的数据中找出多个不同条件下合适的阈值判别室内室外物体，因此这是一个比较典型的二分类问题。该问题需要的准确度要求比较高，而对各个维度的噪声数据不敏感，同时特征的维度可能比较大，因此我们可以考虑使用机器学习领域中的随机森林作为该问题的基础模型。下一节我们将首先简要介绍随机森林的基础决策树，然后再分析随机森林与本问题的关联。

[2]http://www.gbiac.net/gywm/airportsituation

3.2.1　分类器模块

分类决策树[27][28] 是一种描述对实例进行分类的树形结构，如图3.2。决策树由节点（node）和有向表（directed edge）组成。结点有两种类型：内部节点（internal node）和叶节点（leaf node）。内部节点表达的是一系列还未分类的点集，而叶节点表示标记好的一个类。决策树可以认为是 if-then 的规则集合。不同一般的人工设定规则，决策树的规则是通过从训练数据中运用算法计算得出来的，因此效果和泛化能力一般要比人工制定的规则更具可行性。

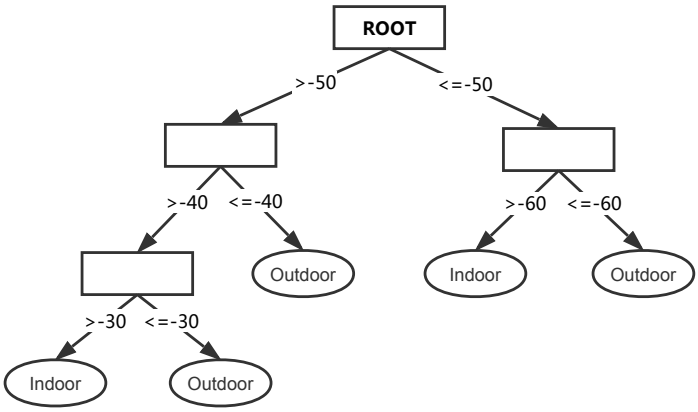

图 3.2　决策树逻辑结构

决策树可以很好的实现类似 if-then 的规则功能：由决策树的根节点到叶节点的每一条路径构建一套规则；路径上从内部节点分出来的有向边对应着规则条件，而叶节点的类对应着规则的结论。决策树的路径或者其对应的 if-then 规则结合具有一个非常重要的性质：互斥且完备。这就是说每一个实例都被一条路径或一条规则所覆盖，而且只被一条路径和规则所覆盖，这里所谓的覆盖是指实例的特征与路径上的特征一致或实例满足规则的条件[28]。

在规则的生成上，决策树采用一种非常有效的方式 –信息增益（Information gain）。信息熵是表示随机变量不确定性的度量，设 X 是一个取有限个值的离散随机变量，其概率分布为：

$$P(X = x_i) = p_i, i = 1, 2..., n, \quad \text{公式（3.1）}$$

则随机变量 X 的熵定义为3.2：

$$H(X) = -\sum_{i=1}^{n} p_i log p_i \quad \text{公式（3.2）}$$

随机变量 X 给定的条件下随机变量 Y 的条件熵（conditional entropy）$H(Y|X)$ 定义为：

$$H(Y|X) = \sum_{i=1}^{n} p_i H(Y|X = x_i) \quad \text{公式（3.3）}$$

信息增益表示得知特征 X 的信息而使得类 Y 的信息的不确定性减少的程度。对于一个特征 A 对训练数据集 D 的信息增益，定义为3.4：

$$g(D, A) = H(D) - H(D|A), \quad \text{公式（3.4）}$$

决策树在分裂节点的时候选择信息增益最大的特征或者切分点进行划分，切分的过程中可以使用不同的特征和切分点，从而形成不同特征之间的组合，例如在本文数据中，我们通过启发式的设定规则也是基于若干个 AP 的信号强度小于等于某个阈值而另外若干个 AP 的信号强度大于某个阈值时物体应该在室外。因此我们可以很清楚的看到基于决策树的学习在有良好的训练数据的情况下是可以生成非常好的规则。

单棵决策树的效果在小数据量的情况下往往比较好，而数据量大的情况下往往在训练集上比较好，而测试集上表现的比较差。针对这一问题，Breiman 在 2001 年提出了随机森林（Random Forest）[29]。随机森林是多棵决策树的组合，不同的是为了提高模型整体的泛化能力，随机森林在使用训练数据时使用了自助采样（Bootstrap Aggregating），保证了每次取的训练数据是独立同分布的，同时在选取特征的时候一次只选若干个特征作为一棵决策树的特征候选集，最

后训练出的多棵决策树的预测结果进行取平均，进一步降低模型的预测方差（variance）。因此我们可以采用基于随机森林的机器学习方法对室内室外物体进行判别。

3.2.2 训练数据构建

对于使用机器学习的方法而言，最重要也是最关键的是带标签的训练数据的采集。训练数据的好坏直接决定了最终分类效果的好坏。在本文中，我们首先使用基于指纹的定位算法的方式通过人工拿多台不同品牌的手机设备在实验室内部和外部分别模拟用户打开开关、连接 WiFi、打开手机应用等多种行为，每次为期两小时，通过在室外和室内的时间段进行标记室内室外，最后从 WiFi 数据库中提取这段时间手机对应接收的信号强度。

从第二章中我们可以知道，不同时间室内人员、门的环境的变化会极大的影响室内的 WiFi 信号数据，因此为了最大程度保证数据的分布能够与实际情况相符，提高我们模型的泛化能力，这样的采样需要在不同时间段进行多次采样。然而这样的数据采集是低效的，因为实际情况中，我们在室外的时候，WiFi 探针收到的信号往往很弱，部分时间甚至没有。因此往往在室内 2 小时内能搜集到 3000 条左右的数据（5*60*6*2=3600，5 台手机搜集两小时，每 10 秒 WiFi 探针输出一次数据，部分时间手机不会发射信号，因此比理论上的少）。而在室外的情况下，同样的时间我们只能收到 300 条左右的数据。这样的结果无疑会带给我们的数据采集工作很大的工作量。

实际生活中，我们都希望数据的采集能够非常符合日常中用户的使用习惯。因此像前面提到的数据采集工作并不合理，实际情况中不大可能一个人会携带多台设备，一人携带单台设备是比较符合实际情况的。而这类型的数据采集是一个非常枯燥和单调的过程。如何避免这样的过程？我们考虑采用众包的思想解决这样一个问题。

首先对于室内的用户，这是一个非常好采集的过程，因为对于一个建筑物

而言，内部必然会有一定的工作人员。他们在内部工作的时候，手机的 WiFi 开关一般都是打开的，因此 WiFi 数据库中会搜集到相当多关于他们的手机设备的信号数据。把他们手机的 MAC 地址登记到我们的服务中，联合他们的信号数据就可以得到室内标签。当然我们必须要考虑到一个问题，室内的工作人员也不可能每时每刻都在室内，若不加选择的全部搜集过来，数据中会存在不少他们中途离开，早上还未到达等相关的噪声数据。因此我们考虑将他们工作时间段最集中的小段时间对应的数据用于室内标签数据。

关于室外标签数据，我们统计分析了 WiFi 数据库中每天用户的信号强度数据，可以发现室内若在室内待的时间超过 5 分钟，在 12 台 WiFi 探针的探测下，会产生 100 条左右的信号强度数据。而在预处理后，我们的 WiFi 数据库中会出现大概 5000 个左右的 MAC 地址，而建筑物中固定工作人员的数量在 200 人左右。因此可以断定剩下的绝大部分都是室外路过的行人。如果直接对这些 MAC 地址按数据包数的大小进行排序，取数量最多的 MAC 地址为室内样本，其余为室外样本则会非常武断。因为建筑内每天都会有临时来该建筑的人员短暂停留过。因此更合理的方式是使用 WiFi 数据包的个数，为其设置一个阈值 r，小于该阈值的对应的 MAC 地址记录都标记为室外。这样我们搜集到的信号强度数据可以分为三部分：室内、未知、室外。其中未知部分的数据是主要的，我们需要利用上述方法中标记为室内和室外的数据作为训练数据来预测未知部分数据记录对应的状态。从每天的统计记录上来看，该阈值越小越好，极端情况下，若当天 12 个 WiFi 探针只接收到该 MAC 地址设备发送了一个数据包，那该设备肯定是室外路过的行人。但阈值太小会造成搜集到的室外标签数据对应的模式不够丰富，不能识别比较靠近建筑物的行人。训练数据采集流程如图3.3：

经过上述工作后，对于每一天而言，设定数据包的阈值为 $r = 50$，我们在实验室大概有 30000 条左右的历史记录可以标记为训练数据，而此外有 150000 条左右的记录需要通过分类器进行标记室内或室外。因此本文创造性的通过启发式的构建规则以及借助室内固定人员的信息避免了人工采集标签数据的过程。最终的标签数据的数量远多于人工采集的标签数据，并且随着时间推移，我们

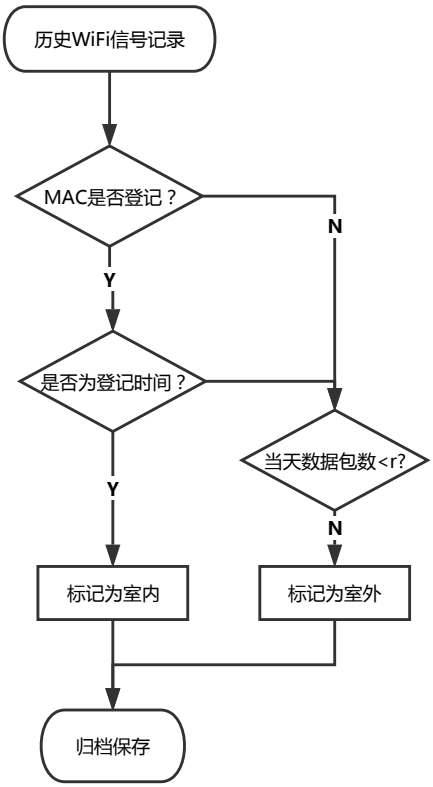

图 3.3　训练数据构建流程

可以重复这个过程，并且使用最近半个月的标签数据作为训练。从而时刻保持模型能学习到室内无线信号的环境特征，大大减少了我们的工作量。

3.2.3　特征构建

有了带标签的训练数据后，我们需要把原始的训练数据的信息通过数值化的方法转换为特征让模型能够学习得到。IODetector[2] 中根据其他传感器信号强度建模的方法，需要寻找比较明显的传感器信号，然后通过一系列实验观察信号强度的数值变化，定义一系列状态概率和状态转移概率，最后通过隐马尔科模型求解。

然而在最常见的基于 WiFi 的室内定位系统中，我们所能知道只有各个 WiFi

探针在一段时间内接收到移动设备发射的无线信号。从第二章的室内传播环境的介绍中我们可以知道，无线信号在室内传播环境非常复杂。要想直接通过观察得到状态概率非常困难。尤其是无线探针数量非常多，而建筑周边的无线覆盖范围比较广的情况下。从图3.4中我们可以直观上知道 WiFi 探针接收到室外移动设备的信号要弱一些，尤其是靠近内部区域的 AP，而靠近建筑边界的 WiFi 探针接收到室外移动设备的信号要强一些。

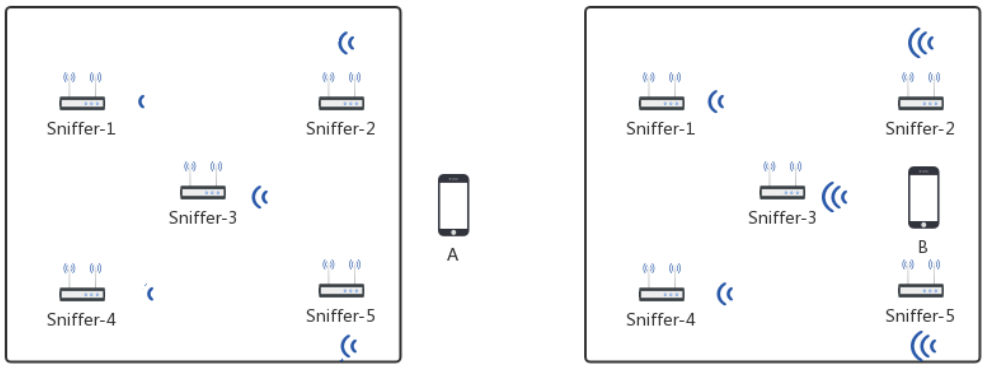

图 3.4　室内室外物体信号差别

因此，我们首先需要构建的特征是各个无线探针在短时间内接收到移动设备的信号强度均值，因为如果室内所有的探针探测到该设备发射的无线信号都很弱，则该设备在室外的概率则非常大。我们统计了搜集到的数据集中所有室内室外物体在各 WiFi 探针的接收信号强度的均值，其结果如表3.2

可以看到室内物体的信号强度统计均值是要大于室外物体的，尤其是在靠近实验室中心区域的 AP 上，室内物体的均值要明显强于室外物体，7 号 AP 由于在室内比较中心地带，其 WiFi 覆盖范围没有超过建筑物的边界，因此导致在实验室现有的环境下该 WiFi 探针对室外物体的接收信号强度基本处于“0”（-100）的状态（由于无线信号强度的单位是 db，数值越小，其对应的信号强度越大，因此如果直接对无信号的数据填 0 则会对分类器的学习产生负面影响。因此我们在计算时对某个时刻没有接收信号强度的数据填充为 -100）。但并不能就此认为

表 3.2　室内室外物体在各 WiFi 探针上的接收信号强度统计均值

ap	indoor(db)	outdoor(db)	ap	indoor(db)	outdoor(db)
ap-1	-60.71	-76.02	ap-7	-68.33	0
ap-2	-68.67	-77.00	ap-8	-60.61	-83.00
ap-3	-64.19	-76.02	ap-9	-75.33	-84.50
ap-4	-58.07	-76.25	ap-10	-68.57	-83.67
ap-5	-59.52	-72.33	ap-11	-70.83	-79.63
ap-6	-61.98	-89.14	ap-12	-78.71	-76.50

7 号 AP 接收信号强度为 -100 的物体就是室外物体。因为在我们搜集到的数据中，也会有一些室内远离 7 号 AP 的物体部分时刻对应的接收信号强度为 -100。随着建筑面积的增大，室内 AP 的个数增多，这种情况会变得更为常见。

其次，室内物体移动速度往往比室外行人移动速度慢，因为室内的顾客主要处于浏览、等待等状态。而移动速度直接影响到接收信号强度的稳定性，因此我们把一段时间内 WiFi 探针接收到的信号强度的标准差也作为特征。与直观意义不同的是室外物体的信号强度的标准差往往是比较小的，因为其信号强度本身就比较弱，在没有信号的时刻我们已经填充为 -100。同样在我们搜集的室内室外物体的信号强度标准差的统计结果如表3.3

再次，室内移动设备经常会连接到 AP 进行上网浏览，因此产生的数据包也会较多，而室外移动设备往往只是路过建筑物，因此其产生的探测数据包往往比较有限，基于此，我们把各个 WiFi 探针一段时间内扫描到 MAC 对应的数据包数也作为特征。同样的方法，我们先在整体数据上分析室内室外物体发送的无线数据包数均值，如表3.4：

而数据包数的方差则室内室外的统计规律并不十分明显，对结果的提升非常有限，这里不再列出。另外探测到移动设备信号的 WiFi 探针的位置也非常重

表 3.3 室内室外物体在各 WiFi 探针上的接收信号强度统计标准差

ap	indoor(db)	outdoor(db)	ap	indoor(db)	outdoor(db)
ap-1	39.85	12.54	ap-7	24.56	0
ap-2	28.27	10.16	ap-8	28.23	4.21
ap-3	26.33	13.22	ap-9	24.23	6.56
ap-4	40.16	9.15	ap-10	30.57	7.45
ap-5	30.22	16.68	ap-11	20.83	10.43
ap-6	25.12	3.61	ap-12	28.14	12.20

表 3.4 室内室外物体在各 WiFi 探针上的接收数据包均值

ap	indoor	outdoor	ap	indoor	outdoor
ap-1	18.31	3.25	ap-7	20.24	0
ap-2	10.25	4.33	ap-8	12.65	5.61
ap-3	8.22	5.15	ap-9	16.54	4.78
ap-4	24.13	6.23	ap-10	15.35	6.51
ap-5	15.22	4.58	ap-11	13.26	5.25
ap-6	12.16	5.21	ap-12	9.14	4.20

要，靠近建筑边界的 WiFi 探针不容易分辨室内室外物体，靠近建筑内部区域的 WiFi 探针容易分辨室内室外物体。因此我们把不同 WiFi 探针接收到移动设备的信号强度按从强到弱进行排序，然后把对应 WiFi 探针的室内坐标填入，让模型学习这样的规律，这种规律直观上并不容易得出，但通过决策树对坐标上的连续数值划分，组合探针的空间信息，可以得到一些准确率上的提升。

前面我们提到在商场等大型建筑中 WiFi 探针的数量是非常多的，如果直接

按照上述方法构建特征，那么我们的特征维度将非常的大。虽然随机森林能学习高维特征的结果，但特征维度的增大将需要大量的样本作为训练。而实际中我们采集到的样本还不足以支撑上述方法中的特征维数。我们知道 WiFi 探针的探测范围并不大，因此我们的特征向量中会有相当多的维度是空值。实际中接收到移动设备的 WiFi 探针数量有限，因此我们把每次接收到一个移动设备信号强度的 WiFi 探针按从强到弱进行排序，取其中的 K 个探针的结果构造特征，由于有这 K 个探针的坐标信息，模型理论能够学习到这种规律，K 的值可以根据实际训练数据的大小进行扩大或缩小。此外上述过程或丢弃一些接收信号弱的 WiFi 探针特征，我们需要特征构建上补充它们的统计信息，例如总和等。

综上所述，我们对室内室外物体的判定构建的特征如下：

1）K 个 WiFi 探针一段时间内分别接收到移动设备信号强度的均值。

2）K 个 WiFi 探针一段时间内分别接收到移动设备信号强度的标准差。

3）K 个 WiFi 探针一段时间内分别探测到该移动设备发射的数据包个数。

4）K 个 WiFi 探针的室内空间坐标。

5）所有 WiFi 探针一段时间内接收到移动设备信号强度的总和。

6）所有 WiFi 探针一段时间内探测到该移动设备发射的数据包总数。

3.3 本章小结

本章首先介绍了基于基础设施的室内定位系统的结构和数据格式，继而引出了室内室外物体判别问题。通过分析 WiFi 探针接收室内室外物体信号强度的特点，提出并分析了基于机器学习的室内室外物体判别方法，并创造性的通过室内固定工作人员的 MAC 以及室外部分行人的信号强度与接收数据包的统计特点，避免了枯燥的数据采集过程，为自动化部署训练数据采集系统奠定了基础，最后基于随机森林模型和 WiFi 信号的特点，提取了丰富的特征，比较好的解决了室内室外物体判别问题。

第 4 章 基于深度学习的室内时空客流预测方法

4.1 引言

室内时空客流是基于用户在室内移动、停留等一系列行为在时间和空间上聚合统计的结果。不同的室内空间对应着具体的事件，虽然从单个用户上来说，他们在室内移动轨迹的规律比较难统一的分析。但从总体上，大量用户在室内的活动则有着明显的规律。例如在商场中，从早上商场开业，室内客流逐渐会有一个上升的过程，这个过程中客流会逐步从低楼层累积到高楼层，到了中午，客流会逐步流向餐饮相关的区域，饭后，客流又逐步分散到商场的其他楼层。而到傍晚，客流又会逐步流向餐饮相关区域，随后又流向影院等娱乐区域。对于机场这样的大型建筑，其规律则更为明显，在候机楼中每天客流伴随着航班的安排从值机区流向安检区再流向登机区。而到达区的客流则从到达区分散到各个机场内部的交通枢纽中，这种规律循环往复。我们定义某个区域 i，其在某个时刻 t 的客流为 $C_t^{(i)}$，则直观上我们可以猜测其下一个时刻的客流为某个非线性函数 $C_{t+1}^{(i)} = f(C_t^{(i)}, C_{t-1}^{(i)}, ..., C_{t-1}^{(i+1)})$。这个非线性函数不仅与当前区域 i 过去的状态 $C_{t-1}^{(i)}$ 有关，还与其他区域 $i+1$ 过去的状态 $C_{t-1}^{(i+1)}$ 有关。在实际情况中，这样的区域和过去的状态可能对应有多个，也就是说这个非线性函数会非常复杂，很难用一个确定的非线性函数来模拟。而神经网络恰好擅长拟合复杂的非线性函数，并且理论上神经网络可以以任意精度拟合任意的非线性函数。

然而传统的前馈神经网络只有输入层、隐藏层和输出层这三层。其中输入层是输入特征矩阵，其中的特征需要像我们在前面描述的通过结合数据特点进行分析得到，对于原因比较明显的问题，这个过程会容易一些，而对于影响因素复杂的问题，这个过程往往充满了试错和不断改进的，往往被称为“特征工程”（feature engineering ）。而在客流时空预测问题中，我们难以定义每个区域

与其他区域的哪个时间段会有关联，或者与历史上哪天的客流有关联。我们可以确定的是某个区域的短期客流往往与其周边的客流有关联，而更长的预测则与其自身比较长的一个时间点的客流有关。例如中午某餐饮店客流很大，则其晚饭时间客流也会比较大。

在图像处理中，卷积神经网络通过多个卷积核在输入矩阵上的滑动，可以很好的学习到图像上不同维度的局部特征，并且不同的卷积核得到不同的类型的局部特征。这一点也非常适用于我们的室内时空客流预测场景，因此我们把不同区域的不同时刻的客流按照时间和空间有序的进行排列作为输入，即我们的输入每一行的内容为 $C_t^{(0)}, C_t^{(1)}, ..., C_t^{(m)}$，然后采用卷积神经网络用多个卷积核来学习结果。

理论上，一个足够深的卷积神经网络可以拟合任意复杂的非线性函数，能够学习输入网络各部分的相关性，但是事实上，标准的卷积神经网络模型对距离较远的点不能很好的记忆，减少记忆的范围会使预测结果不够稳定。而如果可以记忆足够多的历史信息，即使最近的历史预测有一定的误差，但是模型可以考虑更远的历史信息把结果纠正。基于这样的考虑我们加入长短时神经网络模块用于预测。

4.2 基于深度神经网络的预测模型

4.2.1 预测模型框架

正如我们之前提到的，室内不同区域的客流往往存在着时间和空间上的相关性，同时也与自身的长期的历史信息有非常紧密的联系。因此我们采取不同的神经网络学习室内时空客流的这种特点。同时深度神经网络可以在拓扑结构上非常的灵活，可以将不同的输入维度（输出目标要相同）经过不同的模块进行处理后合并到一层神经网络最后输出预测目标函数。我们的预测模型框架如图4.1

通常在深度学习建模环节中最重要的一步就是特征提取，相对于传统的机

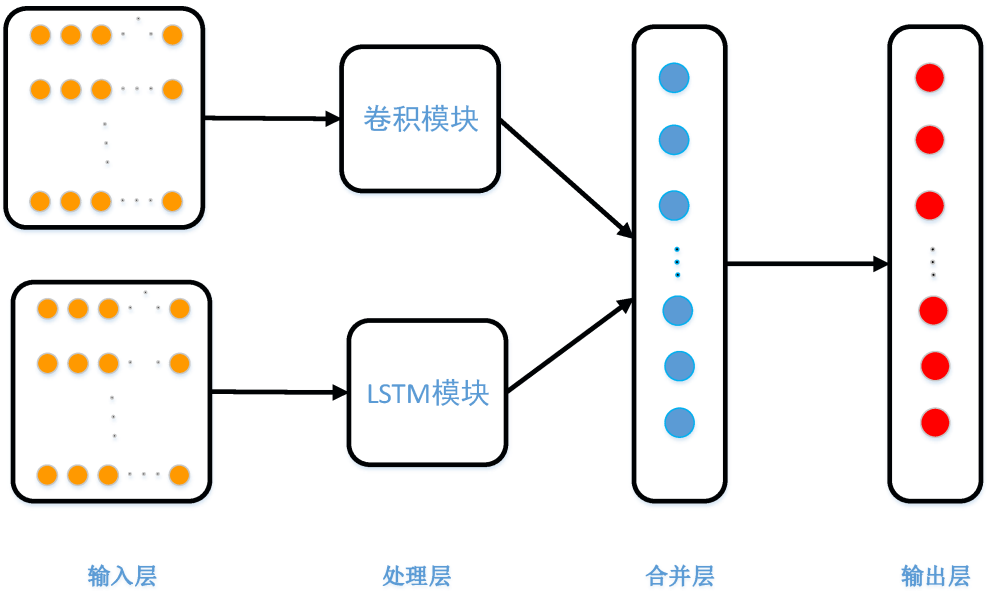

图 4.1　基于深度神经网络的时空客流预测模型框架

器学习的方法。深度学习中所需要的特征一般尽可能的是原始数据，也就是最裸的特征。特征提取完成后将作为输入交由特定的模块进行处理。从图中不难发现我们预测框架中处理模块主要由卷积层、LSTM 层和合并层构成，其中卷积层和 LSTM 层对应了不同形式的输入层，因此我们在详述卷积层和 LSTM 层的工作机制时分别描述他们的输入层：

1）卷积层：卷积神经网络的输入是多维张量，我们已知了室内所有区域的历史客流记录，同时要预测多个区域在未来某一时刻的客流，同时我们希望在输入中邻近之间的元素在实际的物理空间上也是邻近的。基于此我们把同一时刻 m 个区域的客流按空间临近关系排列成行，列按时间先后的 r 个不同单位时间的历史客流，由此我们得到了一个 $m \times r$ 的二维矩阵，对应接下来一个时刻的值作为预测的真值 y。在这样的一个二维矩阵中，卷积核从左往右，从上到下不断滑动，则可形成不同维度上的特征映射。

2）LSTM 层：LSTM 层的输入比较固定，它要求输入的每个样本是一个时间范围（timestep）内的值和对应的 m 维特征形成的二维矩阵。由于室内时空客

流预测的时间片往往比较小，如果需要利用历史上多天相同时刻的客流信息，则时间窗口需要设的很大，这样会造成训练数据的减少和网络参数过多。因此我们首先将数据进行预处理，得到每个区域每个时间片的历史客流，然后使用最近的 K 天同一时间同一区域对应的历史上客流拼接到输入矩阵的行中。例如当前时间如果为 12 点，我们需要预测 12:00∼12:30 的客流，时间片为 30 分钟，则我们的输入中可以设定 9 个近期的时间片对应的客流，即从 7:30 到 12:00 的每半小时的客流。但这样的信息还不够长，因为今天 12:00∼12:30 的客流变化与最近几天的历史趋势也有关系。因此我们可以把最近一周 12:00∼12:30 对应的客流扩充到输入矩阵中。

3）合并层：合并层是将卷积层和 LSTM 层提取出的特征向量合并拼接到同一个向量中，然后接上最后的输出层，完成对目标的预测。合并层的主要作用是融合卷积层和 LSTM 的特征映射，并与输出层的神经元形成线性组合关系进行预测，其形式化表达如公式4.1

$$\vec{y} = \vec{w_o} \cdot [\vec{h_c}, \vec{h_l}], \qquad \text{公式（4.1）}$$

在接下来的两节中，我们结合卷积模块和 LSTM 模块内部工作原理描述对应的计算过程。

4.2.2 卷积模块

卷积神经网络是一种特殊的深度神经网络，其细节部分在第二章中进行过描述。通过最大化采样，输出为室内时空客流的向量表示。在本文框架中，从历史时空客流向量层到卷积输出向量的过程如图4.2所示：

首先通过卷积核在输入时空客流矩阵上进行时间维度上和空间维度上的卷积操作。卷积核在客流向量张量上从左到右，从上到下依次移动。每次会有多个

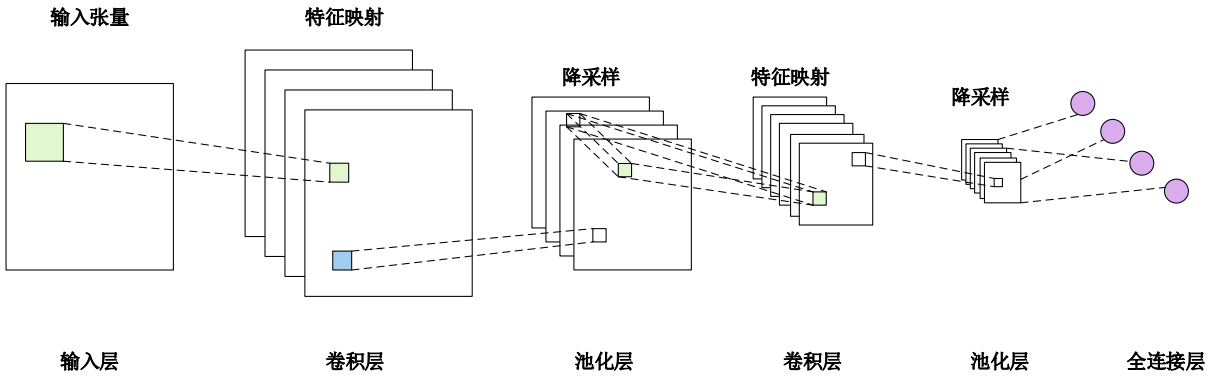

图 4.2　从客流向量到卷积输出向量过程

卷积核，卷积层的计算计算形式如公式4.2

$$x_j^l = f(\sum_{i \in M_j} x_i^{l-1} kernel_{ij}^l + b^l), \qquad \text{公式（4.2）}$$

其中 l 表示层数，M_j 为前一层的局部连接部分，b 为偏置。每一层的卷积操作相当于在上一层输入向量上继续学习高阶特征映射。卷积层后面一半会紧接一层池化层，池化层的主要作用是消除数据中的偏置，其数学表达为公式4.3，其中 n 表示池化窗口大小。

$$x_j^l = \underset{i=1...N}{max} x_i, \qquad \text{公式（4.3）}$$

因为直观上来说，卷积通过在相邻区域的客流向量以及前后时间客流向量上面的卷积计算能够提取客流的在时空上的局部信息，不同的卷积核提取出不同的特征向量，然后经过最大采样，每种卷积的输出选取最大的值，然后将所有卷积的输出拼接在一起，成为客流时空预测的高维特征向量。

4.2.3　LSTM 模块

在第二章中按时序展开的 RNN 模型结构中，我们可以看到神经网络的核心模块 LSTM，读取一个区域客流的输入 X_t，并且输出一个值 H_t，循环神经网络

可以连接先前的信息到当前的任务上，例如使用过去的客流来推测当前的客流。但是如果未来的客流可能需要的上下文信息间隔比较大，比如已知某一区域中午的客流比较大（餐厅），那么它未来晚上的位置还是比较大，因此我们需要往前递推多个时间片。这说明有时候相关的信息和当前的预测时间点之间的距离比较大。遗憾的是随着距离的增大，CNN 或者 RNN 会丧失学习如此远的信息。因此我们需要加入 LSTM 的模块结构，它可以记住自身长期的规律信息。

在时空客流预测模型中，基于当前的区域客流预测下一个时刻该区域的客流中，细胞状态可能已经包含该区域历史客流的大致属性特征，因此可以对短期客流进行推断。当预测较长时间点的客流时，就会希望忘记比较旧的上下文信息，从细胞中丢弃旧的信息，这个决定通过一个称为忘记门完成。该门读取 h_{t-1} 和 x_t, 输出一个概率值给每个细胞状态 c_{t-1}，表达保留信息的概率，神经元状态图如4.3。计算表达式如公式4.4：

$$f_t = (w_f.[h_{t-1}, x_t] + b_f), \qquad \text{公式（4.4）}$$

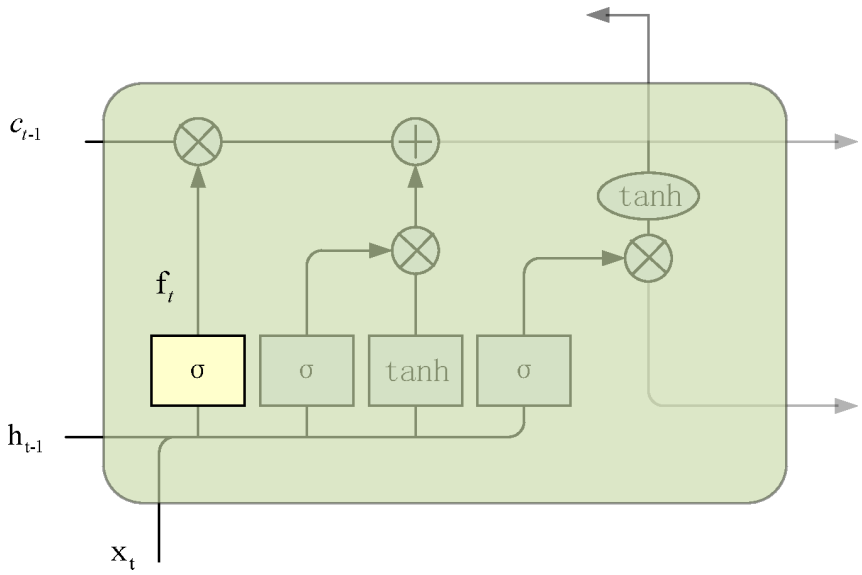

图 4.3　LSTM 舍弃陈旧信息

下一步将确定哪些长期信息将被存放在细胞状态中，这里包含两部分。第一，sigmod 层称为“输入门”决定什么值将要更新，计算表达式如公式4.5，然后一个 tanh 层创建一个新的候选值向量 $\tilde{c}_t$ 会被加入到状态中，计算表达式如公式4.9。在本文的预测模型中，我们希望增加的新的客流信息到细胞状态中，来代替旧的需要忘记的客流信息。对应的神经元状态图如4.4

$$i_t = (w_i.[h_{t-1}, x_t] + b_i), \qquad \text{公式（4.5）}$$

$$\tilde{c_t} = (w_c.[h_{t-1}, x_t] + b_c), \qquad \text{公式（4.6）}$$

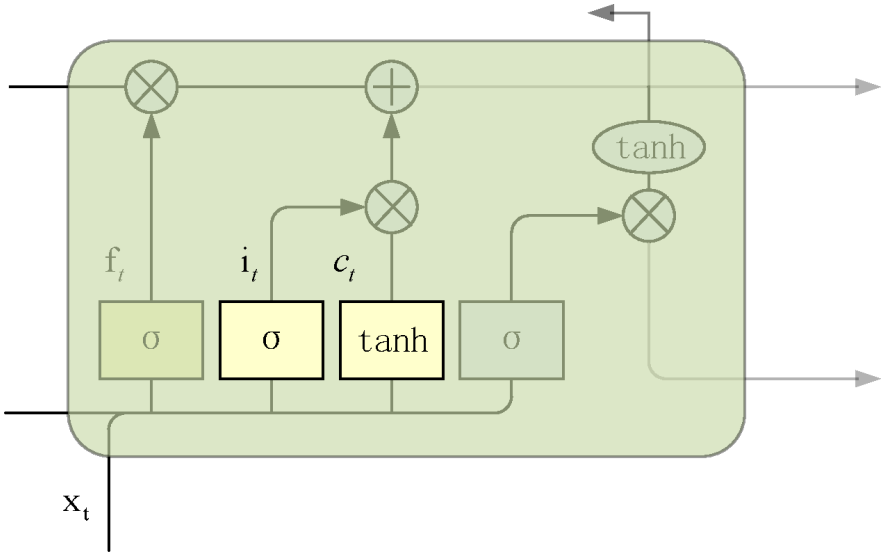

图 4.4　LSTM 更新记忆的状态

再下一就是更新细胞状态，也就是区域客流预测中记忆历史客流信息和未来客流之间的相关性，计算表达式为4.7。对应的神经元状态图如4.5

$$c_t = f_t \times c_{t-1} + i_t \times (c_t), \qquad \text{公式（4.7）}$$

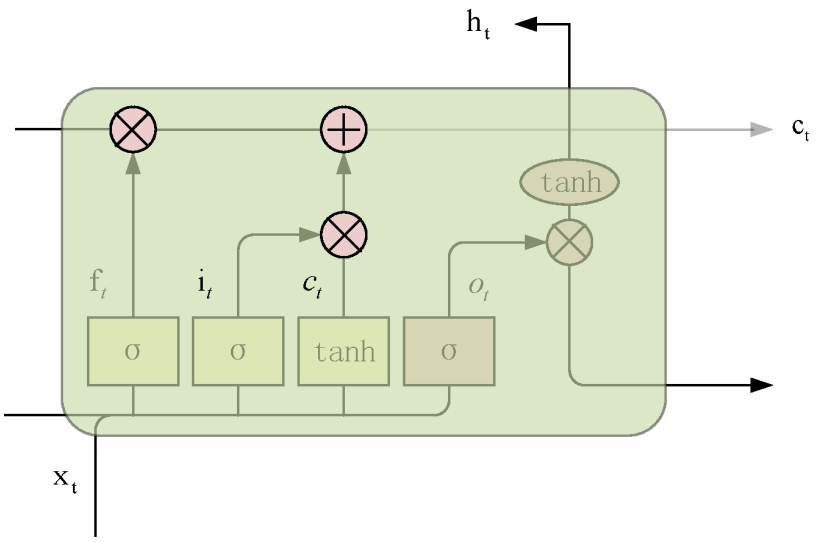

图 4.5　LSTM 更新细胞状态

最终，需要确定输出什么值。这个输出将会基于我们的细胞状态，首先我们运行 sigmod 层来确定细胞状态的那个部分将输出出去，计算表达式如公式4.8。接着，我们把细胞状态通过 tanh 进行处理（得到一个在 -1 和 1 之间的值）并将它和 sigmod 门的输出相乘，最终我们会输出确定输出的那部分，即固定维数不同区域的客流表达，计算表达式为4.9。对应的神经元状态图如4.6

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o), \qquad \text{公式（4.8）}$$

$$h_t = o_t * tanh(c_t), \qquad \text{公式（4.9）}$$

上述的每一步就是标准的 LSTM 模块的计算的过程。

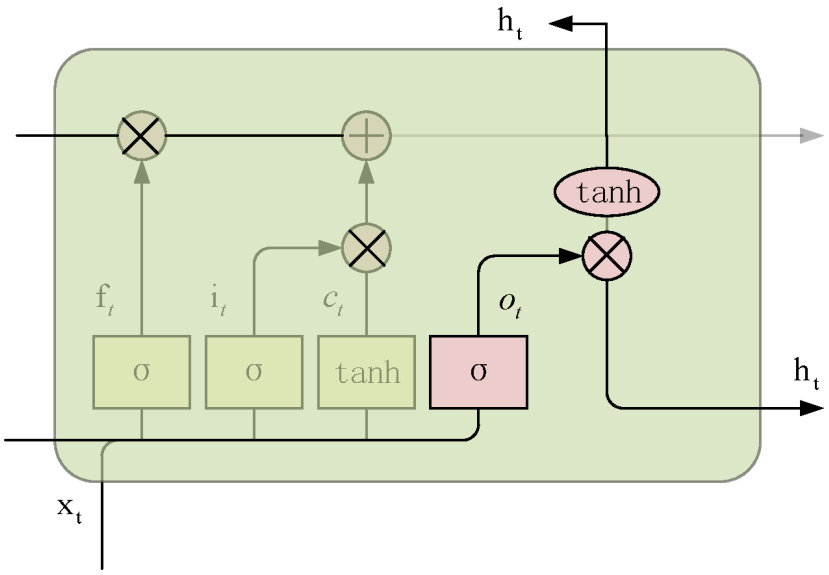

图 4.6　LSTM 确定输出的值

4.3　本章小结

目前，关于客流的预测主要都是基于三层神经网络、ARIMA 等时间序列的方法预测，并且研究方向主要在室外客流的预测，很少关注室内时空客流的预测。室内时空客流不仅有时间前后的关联也有空间上的关联。而深度神经网络无论是在时间局部性和空间局部性上有不错的处理能力，也在处理长期的时间相关性有比较显著的成果。本章通过观察时空客流历史分布，发现室内时空客流存在着这些相关性。我们构建了三层深度神经网络模型，从历史时空客流中提取特征，并应用这些特征预测室内未来的客流时空分布。

第 5 章 实验结果与分析

5.1 实验配置

本文的实验及算法主要在实验室的 4 台小型服务器上完成，其中三台用于搭建基于 WiFi 的室内定位系统，另外一台用于模型的训练和测试，其硬件及对应的系统配置如表5.1：

表 5.1 硬件及系统配置

名称	型号	名称	型号
CPU	Intel Xeon E5-2670	内存	64GB
GPU	NVIDIA Geforce GTX TITAN X	OS	Ubuntu 14.04 LTS Server

相应的我们的实验环境主要的软件及算法包如表5.2

表 5.2 软件配置

名称	版本	名称	版本
Python	2.7.12	Numpy	1.11.2
Pandas	0.16.2	Sklearn	0.14.1
Cuda	8.0	Cudnn	5.0
Keras	1.0.7	Theano	0.7.0

其中 Python 为我们的算法实现语言，此外在搭建实验室的室内定位系统时，还需要用到 WiFi 探针 Netgear WNDR4300 及在上面的 OpenWrt Barrier Breaker 系

统上刷入我们的 WiFi 探测程序，其余的用于搜集和保存 WiFi 信号强度日志及用户位置等数据的软件有 Redis，Logstash 和 Elasticsearch 等。

5.2 数据集描述

本文中提到解决的两个问题，其中室内室外物体判别分析没有基于 WiFi 信号强度的专用数据集，为了评测我们的算法性能。我们通过在实验室手工采集数据构建标准数据集。而对于室内时空客流的预测，我们采用的是广州白云机场候机楼开放的室内客流数据[1]。广州白云机场是中国三大国际枢纽机场之一，平均每天起飞 1000 多个航班，接待的客流量在 10000 人左右。白云机场的建筑结构是一座大型的四层建筑，建筑面积 52.3 万平方米，内部不仅包括了值机区、安检区、登机区、到达区等，也包括了各类中小型商场等服务设施。该数据集包括了白云机场候机楼中 749 个 WiFi 探针对应区域的客流数量，时间从 2016 年 9 月 10 日至 2016 年 9 月 25 日 16 天里 15768815 条数据。正常情况下每个 WiFi 探针每分钟会输出一条该区域的用户数量，即该区域的客流量，其数据样例如表5.3：

表 5.3 白云机场候机楼客流数据样例

WiFi 探针编号	客流量	时间标记
WC-3G<WC-3G-05>	7	2016-09-25-14-59-01
WC-3G<WC-3G-06>	20	2016-09-25-14-59-01
WC-3G<WC-3G-07>	0	2016-09-25-14-59-01
WC-3G<WC-3G-08>	1	2016-09-25-14-59-01
WC-3G<WC-3G-09>	0	2016-09-25-14-59-01

将不同楼层所有的 WiFi 探针的位置投影到平面上，如图5.1，可以看到这些

[1]https://tianchi.shuju.aliyun.com/competition/information.htm?spm=5176.100067.5678.2.NvQtVZ&raceId=231588

WiFi 探针的位置比较均匀的分布在机场内部空间，在比较好的为用户提供免费 WiFi 服务的同时，也很好的统计了室内不同区域的客流信息。

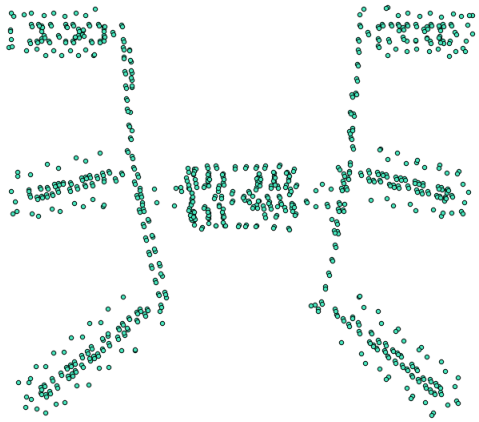

图 5.1 白云机场 WiFi 探针位置投影图

5.3 实验结果与分析

5.3.1 室内室外物体判别结果

5.3.1.1 评测标准

室内室外物体判别是典型的二分类问题，因此我们采用经典的查准率（precision）、查全率（recall）和 F 值（F-Score）三个指标用于评测我们的判别结果。我们定义在标准数据集中室内物体的记录为 $Truth$，预测集合中标记为室内物体的集合为 $Predict$。则上述三个评测指标的定义如下：

查准率指的是预测结果中正确的正样本比例占预测结果的比例，体现的是模型的精度，这里的正样本我们定义为室内物体。见公式5.1

$$Precision = \frac{Truth \cap Predict}{Predict}, \qquad \text{公式（5.1）}$$

查全率指的是预测结果中正确的正样本的数量占标准集中正样本的比例，体现的是模型的覆盖范围。见公式5.2

$$Recall = \frac{Truth \cap Predict}{Truth}, \qquad \text{公式（5.2）}$$

单独的查准率和查全率不足以描述模型的效果，因为如果一个模型把所有的结果预测为正样本，则查全率为 100%，而查准率很低。而如果只预测一条正样本，且该正样本恰好预测中，则查准率也会为 100%，查全率会非常低。因此常常会引入 F 值是对上述两个指标的综合评价，它是差准率和查全率的调和均值。见公式5.3

$$Fscore = \frac{Precision \times Recall \times 2}{Precision + Recall}, \qquad \text{公式（5.3）}$$

5.3.1.2 判别结果及分析

在测试室内室外物体判别实验时，我们需要做两组实验，第一组实验将众包化采集到的数据作为训练数据，把人工采集标记的数据作为测试集。另外一组实验将人工采集标记的数据 80% 作训练集，剩下的 20% 作测试集。设置第二组实验的目的在于检验本文提出的众包化提取的训练数据训练出的模型是否能达到和我们人工采集标记数据集的效果。为了表现训练数据的多样性，我们分别在不同的天里不同时间段采集了室内室外物体的 WiFi 信号强度数据。其中人工标记的数据有 13422 条室内记录和 12214 条室外记录，众包化采集标记的数据有 35436 条室内记录和 125781 室外物体记录。登记的室内固定人员的 MAC 地址有 22 个，这些人员工作时间不一定会打开 WiFi 开关，如果打开则会搜集到他们手机的 WiFi 信号强度数据作为正样本，收集数据的时间集中在工作日固定的四个小时。由于我们的实验室处于道路的附近，所以搜集到数据相比室外物体记录要少。尽管这样众包化以及通过启发式规则采集标记数据的总量还是比我们耗费大量时间和精力采集的标记的数据要多。考虑到中自动标记样本中正负

样本的不平衡性，我们在训练数据时通过对室外物体记录采样，从而使室内室外物体的正负样本比例达到平衡。最后我们通过文中前面提到的方法提取了 98 维特征作为随机森林的输入，训练采用了 500 棵树，最小叶节点为 10，决策树深度为 7。最终的实验结果如表5.4：

表 5.4 室内室外物体判别结果

类别	Precision	Recall	F-score
人工采集	0.9912	0.9909	0.9910
自动采集	0.9893	0.9894	0.9894

最后我们把我们的算法部署到实验室搭建的室内定位系统平台上，每天凌晨自动统计标注训练数据，训练并更新所得到模型，实时预测标记扫描到的移动设备，运行了 7 天，得到室内室外独立 MAC 计数统计如图5.2。由于我们的室内环境每天来往的人数对应的移动设备不会超过 1000 台。从从独立 MAC 计数上来看，每天我们的算法可以很好的过滤大量的室外行人对应的 MAC。

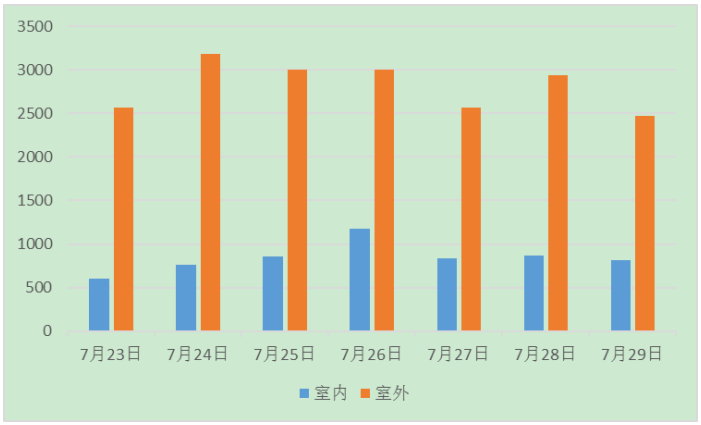

图 5.2 室内室外物体独立 MAC 预测结果

但在对室内时空客流的统计时，我们关注的是每一个时刻区域的客流数量的准确性，因此定位记录的总数中过滤掉的室外行人记录数是更有意义的。基

于此我们也统计了相应的结果，如图5.3

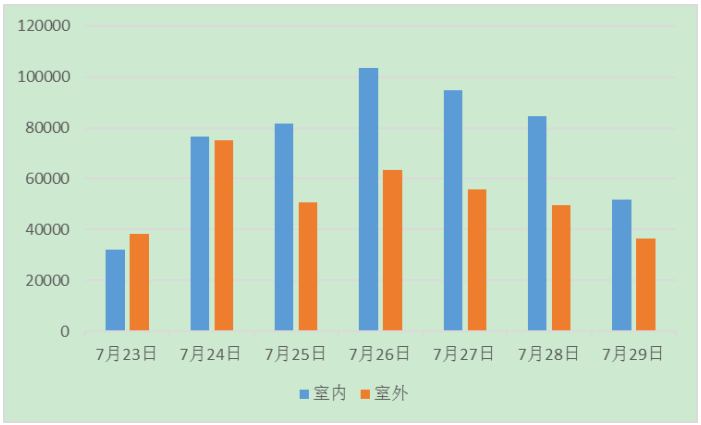

图 5.3 室内室外物体定位记录预测结果

从图中，我们可以看出室内记录数要多于室外记录数，这也非常符合我们的预期，因为室内人员打开 WiFi 开关后每 10 秒就会产生一条定位记录。从实际的结果来看，室外行人的记录数大概占所有记录的 40% 左右。我们的算法比较好的过滤了室外客流，从而为准确统计了室内不同区域不同时间段的客流奠定了基础。

5.3.2 室内时空客流分布预测结果

5.3.2.1 评测标准

室内时空客流的预测是预测未来多个区域的客流量，因此可以使用回归问题中常用的均方误差（Mean Squared Error，MSE）来评测我们的结果。由于未来客流变化越大的区域，其预测难度也越大，而这也是实际运营中更为关心的。均方误差会对预测结果中偏差大的部分施加二范数的惩罚，因此相对于其他常用的平均绝对误差（Mean Absolute Error，MAE）等评测指标可以更好的评测我们的结果。其定义如公式5.4

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2, \qquad \text{公式（5.4）}$$

其中 $\hat{y}_i$ 代表我们对某一区域未来客流的预测值，y_i 代表相应区域未来客流的真实值。

5.3.2.2 预测结果

在第四章种我们介绍了基于深度神经网络来提取室内时空客流的特征及训练方法。我们采用白云机场室内时空客流的历史记录作为模型实验的数据集，并选取了最后 10 分钟，30 分钟，60 分钟，90 分钟和 120 分钟 5 种时间片作为预测区间，而将前面所有的数据记录用于训练模型评测我们的模型。

通常来说影响模型预测结果的因素比较多，尤其是对深度神经网络这种参数较多的模型而言，例如提取特征的维度，特征维度越高，对应的参数个数也越多，训练结果也会不断变好，但在测试集上则可能会出现相反的情况。此外，深度神经网络对应的一些参数也是一个非常重要的因素，例如卷积核的个数，训练过程的轮数等。我们将分不同的情况分析预测的结果。

我们首先针对近期的不同的特征数进行了实验，其结果如表5.5 不同的特征数代表了我们使用该区域最近客流数的时间范围，该值越大，则代表关联的时间范围越大。直观上未来短时间的客流与近期的关联性比较大，因此较少的特征数即可得到不错的结果。实验的结果也充分说明了这一点，随着近期特征数的不断增加，从 2 增长到 32，预测未来 10 分钟的客流对应的误差减小的有限，在特征数超过 8 后效果反而下降，而在预测较长一段时间内的客流时，近期特征数增加带来的效果则比较明显。例如在预测 120 分钟时，近期特征数为 32 的时候比特征数为 2 时下降了 20% 左右。

在模型的构建中，我们提到了室内建筑内的活动每天都有其固有的规律，因此未来的客流与历史上同时刻的客流会有相当的关联，基于此我们构建了最近 K 天预测时间范围对应的客流作为特征。其结果如表5.6。我们选取上了上一实验

表 5.5 近期特征数对预测结果的影响

近期特征数	10 分钟	30 分钟	60 分钟	90 分钟	120 分钟
2	3.568	4.621	5.202	5.713	6.259
4	3.447	4.718	4.971	5.653	6.013
8	3.257	4.265	4.811	5.215	5.726
16	3.341	4.546	4.671	5.105	5.425
32	3.363	4.618	4.712	5.095	5.234

中的最佳的近期特征数，同时在输入特征举证中拼接最近 K 天的历史特征。K 值越大，则代表关联的天数越多。从结果上来看，加入近期的历史特征对短期的客流预测基本没有帮助，在 10 分钟的预测上基本没有提升，但是随着预测区间的不断扩大，历史特征数的增加对预测结果的提升有比较大提升，相比于只使用近期特征的预测结果，整体上提升了 10%。

表 5.6 历史特征数对预测结果的影响

历史特征数	近期特征数	10 分钟	30 分钟	60 分钟	90 分钟	120 分钟
2	8	3.261	4.272	4.511	4.915	5.326
3	8	3.183	4.108	4.271	4.653	5.112
4	8	3.207	3.911	4.113	4.815	4.975
5	8	3.315	3.924	4.071	4.605	4.772
7	8	3.286	3.954	4.012	4.595	4.679

此外卷积核的个数和训练轮数等超参数对预测的结果也有一定的影响，一般来说对于图像识别等任务，卷积核的个数往往设的比较大，比如 128,256 等，这是因为图像的构成往往比较复杂，而在客流预测的过程中我们尝试了不同的

卷积核个数，发现了在客流预测等问题上不需要过大的卷积核个数即可得到较好的效果，卷积核的个数设的比较大反而容易过拟合，训练轮数和结果如表5.7

表 5.7 卷积核数与训练轮数对预测结果的影响

卷积核数	训练轮数	10 分钟	30 分钟	60 分钟	90 分钟	120 分钟
32	30	3.286	3.954	4.012	4.595	4.679
32	50	3.192	4.218	4.211	4.723	4.915
64	30	3.416	4.251	4.513	5.136	5.175
64	50	3.815	4.124	5.121	5.205	5.221
256	30	4.226	4.253	5.212	5.295	5.279

5.3.2.3 模型实验对比结果

为了进一步验证基于深度学习的预测模型的有效性，我们对比了室外客流预测中常用的方法，不同的是由于二者的影响因素并不相同，因此并不适合在同一数据集上做横向对比。因此我们在现有数据的机场数据集上重复室外客流的预测方法。由于室外客流中基于时间序列的预测方法中要求数据是连续的，而白云机场提供的数据集中部分区域存在一定时间段的缺失，因此我们过滤了 32 个存在数据缺失的区域，最终得到 717 个室内区域的客流，仍旧以最后 10 分钟，30 分钟，60 分钟，90 分钟和 120 分钟 5 种时间片作为预测范围。分别采用当前区域客流 $current$、历史同一时间的平均客流 $history$、ARIMA 和随机森林做对比实验，其中随机森林按历史均值，当前的客流值、区域标识、当前小时数、当前分钟数生成了 60 维特征，实验结果如表5.8

对应的预测均方误差曲线图如图5.4，从预测结果来看，基于规则的静态预测方法误差率一直比较大，其中使用当前的客流预测未来的客流在非常短的时间内是可行的，这是因为室内用户的移动速度和频率比较小。但随着预测范围的

表 5.8 不同模型的预测均方差

预测方法	10 分钟	30 分钟	60 分钟	90 分钟	120 分钟
$current$	3.565	10.731	14.408	17.451	20.652
$history$	8.386	6.294	6.202	7.758	6.568
ARIMA	3.687	5.352	7.235	8.231	10.175
随机森林	3.815	4.524	5.421	5.605	5.721
CNNs&LSTM	3.286	3.954	4.012	4.595	4.679

扩大，当前客流与未来客流的误差就越来越大。历史均值的预测误差结果比较稳定，说明室内有比较明显的周期性规律，但相对于我们的模型而言误差要高 50% 左右。而最常用的时间序列客流预测方法在室内预测问题上效果并不稳定。这也比较好理解，室内不同区域的时空客流并不是一个稳定的时间序列，而是不同区域客流时空相互影响的结果。所有对比方法中只有使用了历史特征，空间信息特征等信息的随机森林方法预测结果比较好，但仍旧比我们的方法要差 20%。尤其是随着预测范围的扩大，差距逐步明显。

同样的方法，我们对比了单独使用 LSTM 和 CNNs 两种神经网络的预测结果，实验结果如表5.9。

表 5.9 不同深度神经网络的预测均方差

预测方法	10 分钟	30 分钟	60 分钟	90 分钟	120 分钟
CNNs&LSTM	3.286	3.954	4.012	4.595	4.679
CNNs	3.322	4.012	4.352	4.913	5.142
LSTM	3.518	4.154	4.112	4.823	4.962

对比三者的误差曲线如图5.5，我们可以看出在时空客流的预测上，卷积神

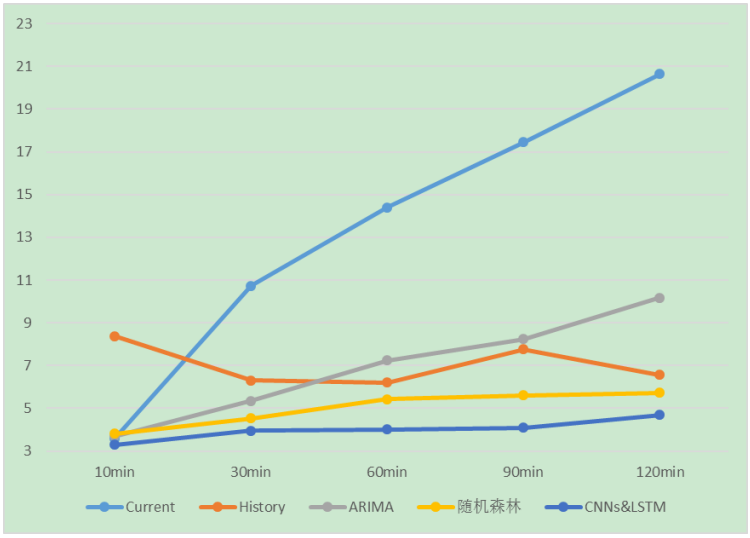

图 5.4 不同模型的预测均方差

经网络和 LSTM 预测结果相比于传统的预测模型都要好不少。两种深度网络合并后得到的预测结果相比于单个模型的结果都要好，说明卷积神经网络和 LSTM 从数据中学习到的特征相关性比较低，从集成学习的角度来说，这样的学习器更容易产生更好的融合结果。

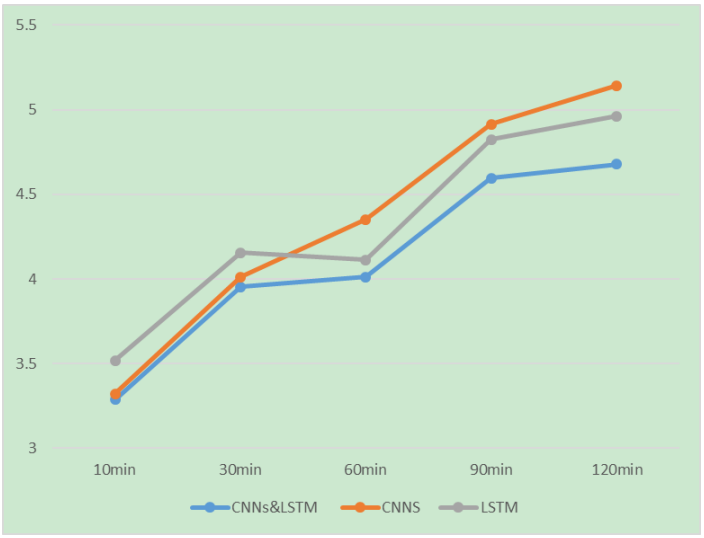

图 5.5 不同深度神经网络的预测均方差

综上所述，基于深度神经网络的方法在拟合预测室内时空客流时相比传统的方法要更有效，具有一定的研究意义。

5.4 本章小结

本章首先介绍了实验室环境和软硬件配置，然后分别介绍和分析了室内室外物体判别和室内时空客流预测使用的数据集和评测指标。接着详细描述了使用不同训练数据下，室内室外物体判别的实验结果，并将实验中的算法提交到实验环境中运行了多天，统计分析了 7 天的运行结果。证明了我们方法的正确性和有效性。最后我们对实际中大型建筑的代表广州白云机场的室内客流进行短时预测，从不同时间片上进行了实验和分析，并结合了多种传统的方法作为对比实验。结果显示基于深度学习的方法效果要明显优于传统的预测方法，尤其是考虑长期的历史信息后对结果有比较明显的提升，达到了我们的预期目标。

第6章 总结与展望

6.1 本文工作总结与贡献

近年智能手机、平板等个人移动设备的普及，商场等大型建筑内的WiFi基础设施部署日趋完善，用户室内的位置获取变得更为便捷。这些位置信息经过时间和空间上的聚合，形成室内不同区域的不同时间段的客流分布。对于大型建筑的运营方而言，及时的了解当前客流和未来客流的时空变化对于室内各项设施和人力的合理分配有着重要意义。同时对于特定的场景例如商场而言，客流的时空分布为其在室内部署广告牌，安排促销活动等日常工作有着很好的指导意义。另一方面来说，客流数据的获取很大程度上避免了对单个用户位置数据的直接处理，对于用户隐私的保护也有着积极的意义。

本文对室内时空客流的预测技术进行了深入研究，结合实际调研，分析得出其中两个重要的子问题，即室内时空客流的统计和室内时空客流的预测。主要的工作内容和贡献如下：

1）提出了一种室内室外物体的判别方法

在基于WiFi的室内定位系统中，室外打开WiFi开关的行人对室内客流的统计有着相当的影响。从本文中的调研和实验中可以看到，室外的行人对应的独立MAC数往往会多于室内的用户。这些行人产生的数据对应的行为模式和内在规律相对于室内的顾客而言有着较大差异。例如室外行人的移动速度往往比室内人员的更快，信号持续的时间比较短。如果不对这部分的数据进行清洗，则我们对室内客流的时空统计结果有比较大影响。另一方面，客流统计过程中如果引入了噪声，则会影响预测过程中模型学习客流内在的规律，从而影响预测精度。因此时空客流的数据清洗过程非常重要，它能为后续的预测工作和进一步的客流挖掘分析工作奠定了基础。

2）提出了一种基于深度神经网络的客流时空预测方法

室内客流的时空相关性往往要高于室外的客流。因此传统的室外客流预测方法中往往只把客流的预测当成一个时间序列的预测问题，而通过我们的实验结果可以得出在短时间内，客流的变化与上一个时刻的相关性确实比较大。然而随着预测时间跨度的增大，室内客流更多的与其他区域产生了更大的相关性，同时也与自身的长期趋势也密切相关。这种相关性往往是非线性的，传统的ARIMA方法比较难处理这类情况。而使用不同的深度神经网络，通过提取不同区域的时空特征，能对预测结果有比较明显的提升。

6.2 未来研究工作展望

本文基本实现了室内时空客流的统计和预测目标，然而在实践过程仍然存在这一些不足和待改进的地方。具体来说有以下几点：

1）室内室外物体的判别方法的通用性

本文提出的室内室外物体的判别方法主要基于的是用户手机等移动设备发射的WiFi信号强度。而随着4G资费的降低，蓝牙可穿戴设备的兴起，未来室内的无线信号可能会越来越多的是蓝牙无线信号。蓝牙信号相对于WiFi信号强度更小，传播距离也更短。室外客流的影响相对会变小，但如何在这种情况下过滤室内客流是本文下一步关注的方向。

2）影响室内客流的因素很多。室内客流的时空影响因素很多，除了时间和空间上的信息外，还有室内发生的一些事件。例如商场的营业时间，店铺的促销活动；机场的航班排班、天气和航班的延误等。这些事件的发生对于室内客流的短期和长期预测都非常有意义。本文主要关注的客流预测集中在短期的预测上，因为长期的预测需要深入到特定场景中，结合事件发生的特点才能合理的建立模型，从而准确的预测未来的时空客流。同时在数据足够的情况，对发生突发事件时，室内客流的变化和预测也是本文下一步努力的方向。

参考文献

[1] Matthew Gast. Analysis of iOS 8 MAC randomization on locationing[M]. ”ZEBRA TECHNOLOGIES”, 2005.

[2] Mo Li, Pengfei Zhou, Yuanqing Zheng, Zhenjiang Li, Guobin Shen. IODetector: A Generic Service for Indoor/Outdoor Detection[J]. ACM Trans. Sen. Netw., 2014, 11(2):28:1—-28:29.

[3] Henrik Blunck, Mikkel Baun Kjærgaard, Thomas Skjødeberg Toftegaard. Sensing and classifying impairments of GPS reception on mobile devices[J]. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011, 6696 LNCS:350–367.

[4] Hua Lu, Chenjuan Guo, Bin Yang, Christian S Jensen. Finding frequently visited indoor pois using symbolic indoor tracking data.[C]. In EDBT. 2016, 449–460.

[5] 刘建军, 廖闻剑, 彭艳兵. 两种时间序列模型在客流量预测上的比较 [J]. 计算机工程与应用, 2016, 9:042.

[6] Wei Xu, Yong Qin, Houkuan Huang. A new method of railway passenger flow forecasting based on spatio-temporal data mining[C]. In Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on. IEEE, 2004, 402–405.

[7] Chunhui Zhang, Rui Song, Yang Sun. Kalman Filter-Based Short-Term Passenger Flow Forecasting on Bus Stop[EB/OL], 2011.

[8] Pratap S Prasad, Prathima Agrawal. Mobility prediction for wireless network resource management[C]. In System Theory, 2009. SSST 2009. 41st Southeastern Symposium on. IEEE, 2009, 98–102.

[9] 杨铮, 吴陈沭, 刘云浩. 位置计算: 无线网络定位与可定位性 [EB/OL], 2014.

[10] Peng Rong, Mihail L Sichitiu. Angle of arrival localization for wireless sensor networks[C]. In 2006 3rd annual IEEE communications society on sensor and ad hoc communications and networks. IEEE, 2006, 1, 374–382.

[11] Quentin H Spencer, Brian D Jeffs, Michael A Jensen, A Lee Swindlehurst. Modeling the statistical time and angle of arrival characteristics of an indoor multipath channel[J]. IEEE Journal on Selected Areas in Communications, 2000, 18(3):347–360.

[12] Farshid Alizadeh-Shabdiz. Time difference of arrival based estimation of direction of travel in a wlan positioning system[EB/OL], 2007. US Patent App. 11/696,833.

[13] Paolo Pivato, Luigi Palopoli, Dario Petri. Accuracy of RSS-based centroid localization algorithms in an indoor environment[J]. IEEE Transactions on Instrumentation and Measurement, 2011, 60(10):3451–3460.

[14] Sachin Ganu, AS Krishnakumar, P Krishnan. Infrastructure-based location estimation in wlan networks[C]. In IEEE Wireless Communications and Networking Conference (WCNC 2004). 2004, 1, 465–470.

[15] Filipe Meneses, Adriano Moreira. Large scale movement analysis from WiFi based location data[J]. International Conference on Indoor Positioning and Indoor Navigation, 2012, (November).

[16] Antonio J Ruiz-Ruiz, Henrik Blunck, Thor S Prentow, Allan Stisen, Mikkel B Kjaergaard. Analysis methods for extracting knowledge from large-scale wifi monitoring to inform building facility planning[C]. In Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on. IEEE, 2014, 130–138.

[17] David Gerónimo, Antonio M. López, Angel D. Sappa, Thorsten Graf. Survey of pedestrian detection for advanced driver assistance systems[J]. IEEE Transactions

on Pattern Analysis and Machine Intelligence, 2010, 32(7):1239–1258.

[18] Piotr Dollar, Christian Wojek, Bernt Schiele, Pietro Perona. Pedestrian detection: An evaluation of the state of the art[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 34(4):743–761.

[19] Traffic Flow Forecasting, Comparison Of, Modeling Approaches, By Brian. TRAFFIC FLOW FORECASTING: COMPARISON OF MODELING ApPROACHES By Brian L.Smith and Michael J.Demetsky, zFellow, ASCE (Reviewed by the Urban Transportation Division)[J]. Transportation, 1997, i(August):261–266.

[20] Billy M. Williams, Lester a. Hoel. Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results[J]. Journal of Transportation Engineering, 2003, 129(6):664–672.

[21] Yanru Zhang, Yunlong Zhang, Ali Haghani. A hybrid short-term traffic flow forecasting method based on spectral analysis and statistical volatility model[J]. Transportation Research Part C: Emerging Technologies, 2014, 43:65–78.

[22] Yoshua Bengio. Learning Deep Architectures for AI[J]. Foundations and Trends in Machine Learning, 2009, 2(1):1–127.

[23] Yann LeCun, Yoshua Bengio, Geoffrey Hinton. Deep learning[J]. Nature, 2015, 521(7553):436–444.

[24] Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks[C]. In Advances in neural information processing systems. 2012, 1097–1105.

[25] Pauline C Ng, Steven Henikoff. Sift: Predicting amino acid changes that affect protein function[J]. Nucleic acids research, 2003, 31(13):3812–3814.

[26] Sepp Hochreiter, Jürgen Schmidhuber. Long short-term memory[J]. Neural computation, 1997, 9(8):1735–1780.

[27] Jerome Friedman, Trevor Hastie, Robert Tibshirani. The elements of statistical learning[M], 1. Springer series in statistics Springer, Berlin, 2001.

[28] 李航. 统计学习方法 [J]. 清华大学出版社, 北京, 2012.

[29] Leo Breiman. Random forests[J]. Machine learning, 2001, 45(1):5–32.

攻读硕士学位期间的主要研究成果

致谢

两年半的时光转眼即逝，学习道路即将告一段路，在这里留下了我太多的回忆和不舍，在即将离别之际，感谢所有给予我帮助的老师和同学们。

首先感谢我的导师寿黎但教授在研究生生涯里给予的指导和关怀，使我能够顺利地完成论文的选题和实践。您严谨治学的态度、对科研工作孜孜不倦的精神是我求学道路上不断取得进步的动力和榜样。感谢陈珂老师在我遭受生活和工作上的挫折时，不断给予的细心关怀和支持，使我重拾前进的勇气。

感谢李环、骆歆远、顾晓玲、彭湃等各位博士在科研和学习生活上给我的指导和帮助。特别感谢李环和骆歆远博士在日常的科研工作中给予指导以及对这篇论文做出的贡献，为我完成本论文提供了坚实的基础。

感谢何平、刘博文、唐思、李幸超、柴一平、唐晓瑜、朱一聪、俞骋超等师兄师姐在日常的学习以及求职期间为我提供的宝贵经验。感谢史飞超、陈欣、赵萍、喻影等师弟师妹，和你们一块工作和学习的一年多时间里，我感觉很开心。

感谢我的小伙伴们：冯杰、吴联坤、胡凡、张也、王改革、王伟迪、于志超、金明健、周俊林、吴晓晓、钱宇、朱华、任伟超、张静恬、刘伟、朱清华。也非常感谢我的同学钊魁在我科研和学习道路上不断对我的支持和陪伴，谢谢你们与我一起在最美好的年华中努力奋斗过，不在研究生生涯中留下遗憾。

最后，感谢我的父母、我的姐姐在我成长过程中的关怀和陪伴，谢谢你们的理解和支持，让我在负面情绪时能及时开导和鼓励。谢谢你们让我能健康快乐的成长。祝你们生活快乐！

署名：李邦鹏

2017 年 1 月于浙大求是园