硕士专业学位论文

大规模基站网络流量模式挖掘和预测

Mining and Predicting Traffic Patterns in a Large-scale Base-Station Network

作者：苏健

导师：陈一帅

北京交通大学

2019 年 5 月

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通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，提供阅览服务，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。学校可以为存在馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：苏健　　导师签名：

签字日期：2019年5月30日　　签字日期：2019年5月30日

学校代码：10004 密级：公开

北京交通大学

硕士专业学位论文

大规模基站网络流量模式挖掘和预测

Mining and Predicting Traffic Patterns in a Large-scale Base-Station Network

作者姓名：苏健 学 号：17125052

导师姓名：陈一帅 职 称：副教授

工程硕士专业领域：电子与通信工程学位级别：硕士

北京交通大学

2019 年 5 月

致谢

本论文的研究工作是在我的导师陈一帅老师的悉心指导下完成的。陈一帅老师严谨的科学态度，积极探索和求真务实的科研精神深深感染了我。在学习中，陈一帅老师丰富渊博的知识、敏锐的学术思维、精益求精的工作态度以及侮人不倦的师者风范是我终生学习的楷模，这也对我今后的学习、工作生涯起到了重要的指引作用。在此衷心感谢两年来陈一帅老师对我生活、学习上的指导和关怀。

衷心感谢郭宇春老师、赵永祥老师和实验室里的所有老师在我研究生学习阶段对我的无私帮助和关怀。特别感谢郭宇春老师和赵永祥老师对我论文的审阅并提出宝贵的修改意见，在老师们的帮助下，我不断完善论文内容并最终成功完成毕业论文。同时，导师们的高深精湛的造诣与严谨求实的治学精神，将永远激励着我，在此向各位老师表示诚挚的谢意。

衷心感谢王雨师姐对我的支持与帮助。感谢她对论文的数据支持、科研问题的深刻讨论，对我完成论文提供了宝贵的指导意见。在此向王雨师姐表示衷心的感谢。

衷心感谢唐伟康师兄、冯梦菲、艾方哲、王珍珠、曹中等同学在实验室工作和撰写论文期间对我的研究工作给予的热心帮助。在此向他们表示我的感谢之意。

最后，特别感谢一直无微不至的关心、支持我的父母，正是他们对我不断的鼓励与默默地付出，才使得我有足够的信心与毅力克服科研途中的困难，并顺利地完成学业，成为社会有用之才。

摘要

随着 4G 无线技术普及，无线基站流量持续增加。作为承载无线网络的基础设施，分析基站流量的静态和动态特征、挖掘基站流量演变模式对无线基站的运营方案制定、参数合理配置至关重要。目前已有的测量和模型工作，聚焦在短时间粒度（如分钟、小时、天）的流量变化规律，至今尚缺少一个长时间尺度（如一年）的城市基站流量变化的准确测量和模式挖掘结果。

为此，本文基于一个中国大型无线网络运营商在一个大型城市的基站网络的流量测量数据，对超过 7 千个基站、以月度为单位、持续一年的基站流量的静态和动态特征进行了观察与测量，并对基站流量的时间变化模式进行了聚类、分析和预测。本文贡献如下：

(1) 提出了一种新的基站流量演变模式的聚类方法。该方法基于基站月度总流量值在一年内的排序序列进行聚类。在我们数据集上的实验结果表明：该方法对短时间序列（不存在周期性）的波形涨跌特点有很好的描述，能得到比传统方法更容易理解的聚类结果。

(2) 基于该聚类方法，我们对运营商的 7 千多个基站的流量演变模式进行了大规模聚类分析，获得了 6 种典型基站流量演变模式，最主要的一种流量模式涵盖了 38.6%的基站，特点为流量总体呈上升趋势，11 月份达到高峰，次年 2 月降到低谷。其它模式包括：“春节返乡”、 “双 11”电商购物模式等。结合该城市的特点，我们对各模式的形成原因做出了解释，这些发现为运营商掌握其基站的流量演变的规律提供了有益的指导。

(3) 提出了一种基于基站地理位置和地址语义信息的基站模式预测方法，对一个新建基站的流量模式进行预测。因为新建基站的初始信息很少，预测困难，因此我们创新性地将基站语义标签信息引入到基站流量模式预测中。实验结果表明：通过加入基站词向量表征，预测模型的 F1-score 提升了 5%，其中两种模式的预测准确度较高。

本文对大规模实际基站网络流量的分析结果对于网络运营商有重要的实际价值，本文提出的长期流量变化模式模型、聚类算法、和预测算法，具有通用性，能够被应用于类似的各种应用场景，具有重要的理论价值和实际意义。

图 20 幅，表 8 个，参考文献 43 篇。

关键词：网络测量；模式挖掘；模式预测；机器学习

ABSTRACT

With the popularization of 4G wireless technology, the traffic of wireless base station continues to increase. As the infrastructure of wireless network, it is very important to analyze the static and dynamic characteristics of base station traffic and excavate the evolving mode of base station traffic for the formulation of operation scheme and reasonable configuration of parameters of base station. At present, the existing measurement and model work focuses on the rule of traffic change in short time granularity (such as minute, hour, day). So far, there is no accurate measurement and pattern mining results of traffic change in a long time scale (such as one year) for urban base stations.

Therefore, based on the traffic data of a large Chinese wireless network operator in a large city's base station network, this paper observes and measures the static and dynamic characteristics of base station traffic of more than 7,000 base stations in monthly units for one year, and clusters, analyses and predicts the time-varying patterns of base station traffic. The contributions of this paper are as follows:

(1) A new clustering method for base station traffic evolution pattern is proposed. The method is based on the ranking sequence of the monthly total traffic value of the base station in one year. The experimental results on our data set show that this method can describe the fluctuation of waveform in short time series (without periodicity), and get clustering results that are easier to understand than traditional methods.

(2) Based on this clustering method, we conducted a large-scale clustering analysis on the traffic evolution modes of more than 7,000 base stations of operators, and obtained six typical base station traffic evolution modes. The most important one covers 38.6% of the base stations, which is characterized by an overall upward trend of traffic, peaking in November and falling to a low in February next year. Other modes include "Spring Festival returns home" and "Double 11" e-commerce shopping mode. Combining with the characteristics of the city, we explain the reasons for the formation of various modes. These findings provide useful guidance for operators to master the rules of traffic evolution of their base stations.

(3) A base station mode prediction method based on the location and address semantics information of the base station is proposed to predict the traffic mode of a new base station. Because the initial information of the new base station is very little

and the prediction is difficult, we innovatively introduce the base station semantic label information into the base station traffic mode prediction. The experimental results show that the F1-score of the prediction model is improved by 5% by adding base station word vector representation, and the prediction accuracy of the two models is higher.

The analysis results of large-scale actual base station network traffic have important practical value for network operators. The long-term traffic change model, clustering algorithm and prediction algorithm proposed in this paper are universal and can be applied to similar application scenarios. They have important theoretical value and practical significance.

Figure 20, table 8, reference 43.

KEYWORDS：Network measurement; pattern mining; pattern prediction; machine learning

目录

缩略词表

英文缩写	英文全称	中文全称
NLP	Natural Language Processing	自然语言处理
CDF	Cumulative Distribution Function	累积分布函数
PDF	Probability Density Function	概率密度函数
SMOTE	Synthetic Minority Oversampling Technique	合成少数类过采样技术
RTT	Round-Trip Time	往返时延
RNN	Recurrent Neural Networks	循环神经网络
LR	Logistic Regression	逻辑回归
KNN	K-Nearest Neighbor	k 近邻
MSE	Mean Squared Error	均方误差
NLP	Natural Language Processing	自然语言处理
CBR	Case-Based Reasoning	基于案例推理

1 引言

1.1 研究背景和意义

随着 4G[1]技术商用普及，基站作为承载网络流量的基础设施，分析挖掘其网络流量模式对无线基站的运营方案制定、参数合理配置至关重要。Markets and Markets 网站分析指出，到 2020 年全球超过 50%的设备和连接将在移动端完成（例如：智能手机、无人驾驶车，智能穿戴设备等），每月全球移动数据流量将超过 30.6×10^{18} bytes，同时该趋势将在可预见的未来继续下去。同时网络流量分析的市场规模将从 2015 年的 7.6 亿美元增长到 2020 年的 23.2 亿美元。每年 SIGCOMM、SIGMETRICS 等顶级会议中，都会有相当数量的文章致力于网络测量的研究。目前 Cisco、华为等大型网络设备供应商以及很多初创公司（例如：Pluribus Networks, Logic Monitor, Big monitoring fabric）也都致力于网络流量分析的研究和产品开发当中。5G 站点峰值下载速率可达 1.5G/秒，最高可达 4G 普通下载速率的 500 倍，面对海量数据，基站流量的测量分析与模式挖掘十分必要。

基站网络流量模式挖掘的主要目标是通过测量与分析基站数据，利用模式挖掘方法从大量基站数据中提取具有实际分析价值和意义的模式。这是一种无监督学习过程，我们并不知道能够从数据中提炼出的模式数量，同时每种模式具体特点也事先未知。基站流量数据与其他类型数据有所不同，首先它具有明显地理分布特点，即不同的地理位置其流量变化特征；其次，基站流量数据来源于海量用户的移动设备使用记录，它在一定程度上反映了人类活动的特征，所以像节假日、用户所在城市、季节变化等只要能够对用户造成影响的特征都能够影响到基站流量变化。所以在制定模式划分规则时，一定要结合当地区域特性、特殊事件等社会因素，只有这样我们提取的基站网络流量模式才能够有实际意义。因此分析测量基站流量数据特点、制定切实合理模式规则是基站网络流量模式挖掘的重要内容。

在模式挖掘的过程中，基站流量模式能够具有结合实际的解释性尤为重要。现实的基站使用运营情况中，基站的流量动态变化与其安置地的地域特征，社会事件高度相关，同时具体研究过程也要以运营商和管理者的角度去分析哪些基站流量特点需要格外关注。比如作为大型电商城市，在“双 11”电商购物节时期的基站流量模式特点和模式分布情况就需要重点考虑，这是对运营和管理有意义的部分，所以在模式挖掘的过程中，基站流量模式能够具有结合实际的解释性尤为重要。

除了模式挖掘，对基站流量模式的预测也非常重要。对于新建基站，如果能利用已有的信息预测出该基站的流量模式，就能够提前掌握该基站流量变化的特性，对此可以提前规划该基站的参数配置，运营方案等。

本文重点从模型的可解释性上出发，通过对大规模实际数据的多角度测量分析，发现数据的一般规律，然后提出针对数据特点的新的聚类算法，获得具有良好可解释性的基站流量模式聚类结果，然后系统评估基站流量模式和基站地理位置及基站地址语义信息的相关性，提出基于基站地理位置和地址语义信息的模式预测模型。这些工作对运营商提升基站运营效率、节约基站运营成本具有重要意义。

1.2 国内外研究现状

1.2.1 研究现状

本小节介绍现阶段国内外基站网络流量模式挖掘以及应用研究情况，说明其不足，具体来说，可以分为以下几个方面：

(1) 目前基站流量模式挖掘主要做法之一是通过对网络流量时空模式的分析测量，建立统计模型，这些模型反映了流量的分布规律，但不能刻画基站流量的演变过程。文献[2]中首先对我国商业蜂窝网络的流量测量数据进行了分析测量，论证了基于时间和空间的对数正态分布和威布尔分布可以近似表示流量密度(单位面积的流量负载)的空间分布。该文章提出了一种时空流量模型，该模型通过一组正弦信号的叠加模拟大规模时空流量变化。文献[3]测量和描述了移动互联网流量的时空动态,提出了一种类似 zipf 的模型来捕捉应用流量的体积分布，以及一种马尔科夫模型来捕捉聚合互联网流量的体积动态。文献[4]首先统计了该区域基站流量，表明对数正态混合分布可以精确地模拟基站连接密度和用户体验数据速率的空间分布，并且构建了一个能够生成不同基站连接密度和用户体验数据速率的综合基站的网络能力模型。文献[5]建立了一个 5 周的模板来描述时间波动的半周期模式，并提出了一套定量指标来度量假日流量对该模板的偏差。结果揭示了公共假期的不同效应。在短假期(最长 3 天)，网络流量会比正常时间略高一些，而这种影响在相邻的几天内是有限的。对于长假(长达 7 天)，网络流量会有较大的拉动或推动作用，前一周和后一周的交通量也是如此。

(2) 目前对无线网络流量数据的模式挖掘工作，主要目标是研究用户的网络使用行为或短期活动模式带来的短期网络流量模式（如以天为周期的模式变化），尚没有以年为跨度的长期流量模式的预测和模型研究。文献[6]根据移动流量数据

对表现出相似特征的行为进行了分类，并对异常行为进行了分析和表征，开发了一种方法用于提取和表征正常的用户行为模式以及识别异常行为。文章将这些行为与同一时期发生的重要事件(如国家和宗教节日)联系起来，并研究所观察到的行为与这些事件的相互作用。文献[7]利用全国 3G 蜂窝数据网络内采集的大规模数据集，对网络流量使用情况和用户行为进行了详细的测量分析，分析了用户移动与时间的活动模式并确定了它们与业务量的关系。文献[8]通过分析蜂窝网络运营商内部收集的数据集，对蜂窝网络的主要用途提供了一种独特的、互补的分析。研究发现调用持续时间与常用的指数分布有明显的偏差，这使得基于调用的建模更加复杂。研究还展示了一个不使用调用持续时间的随机游走过程，它通常可以用来建模聚合单元的容量。文献[9]使用聚合的移动手机信息来挖掘游客的行为以代替人工分析。他们使用一个传播模型对研究区域进行划分，提取网络的不同时间和空间模式，最后他们绘制了堆叠的柱状图来显示每天、每季度的模式。文献[10]的目的是挖掘基站流量模式，揭示城市环境中人类活动与基站流量模式之间的关系。该研究提出了一种新方法并设计了一个功能强大的分析系统。基站真实的 6400 个基站数据，该文章确定了与不同人群日常活动模式相对应的五种日常流量模式。其次，该文章从每周的移动流量消费趋势中又可以提取出两种自然模式，从另外的角度反映了人类的活动模式。

(3) 在基站流量模式预测方面，现有的预测方式主要依据长时期、多周期的历史流量数据中体现的流量演变规律，对未来的流量进行预测，尚没有预测基站所属流量演变模式的预测研究。文献[4]对 3G 网络流量模式的预测进行了大量的研究。介绍了一种基于马尔可夫模型的接入点中心方法。对该方法的可行性进行了评价，并对不同阶值预测器的预测性能进行了比较。最后得出结论：阶值为 4 的马尔可夫模型[11]对于单步提前和多步提前的流量模式预测都是有效的。文献[12]中提出了一种新的移动蜂窝网络流量均衡方法，通过预测无线覆盖的变化来提前预防网络拥塞。基于案例的推理(CBR[13])用于学习拥塞期间的流量模式，并在相似的流量模式重新出现时应用最合适的天线模式。与以往的工作不同，该方案不需要每次计算最优模式，大大降低了计算复杂度。此外，CBR 还提供了拥塞预测的能力。对不同流量场景下的系统性能进行了仿真，并给出了仿真结果。文献[14]将 k 均值聚类、Elman 神经网络[15]和小波分解方法[16]相结合，对流量预测的性能进行了比较。其他的工作则利用序列本身的特点，采用聚类的方法进行预测。文献[17]采用聚类算法对序列的前半部分进行聚类，最后在各自类别中分类，预测时间序列的后半部分。文献[18]利用时空层次结构，对时间序列进行聚合，提出了一种多尺度预测系统(MSFS)来预测不同粒度的时间序列。文献[19]利用时间序列预测来预测基于规则成分的流量模式。研究验证了利用时间序列分解方法进行预测的

有效性，并展示了规律性和随机性分量的地理分布。此外，还揭示了规则分量的高可预测性是可以实现的，并证明了对流量数据的随机性分量的预测是不可能的。

1.2.2 研究难点及挑战

下面简述我们在基站流量模式挖掘和应用研究的过程中遇到的难点和挑战：

(1) 数据时序点较少（7110 个基站，每个基站有 5 个非连续月度流量汇总数值，即短时间序列）。这将导致适用于长时间序列的分析方法不能借鉴使用，比如循环神网络 LSTM[20]，RNN[21]、时序分解模型 STL[22]，它们都需要很长的历史时序点作为数据支撑。在此条件下，能否找到一种方法很好描述短时间序列的特点，同时反映基站网络流量特点的表征方法是模式挖掘的一大难点。

(2) 传统的聚类算法往往通过计算类内距离作为样本划分成各种类型的标准，但这种聚类方法在对基站变化的一致性方面有所不足。例如：有两个基站的变化情况是：“涨—涨—跌”，“涨—跌—涨”，它们可能仅仅因为内类距离较小而划分成一组。但在运营商的角度，他们更希望发现的是在“双 11”电商节、“春节”是基站涨跌一致的基站群落，这样运营商能够在流量一致性变化时做好应对措施。

(3) 基站位置所处的宫格名称是其特有的属性即地理语义信息理语义，在以往的基站数据分析中鲜有研究，我们首次将 NLP 语言模型分析方法引入基站模式挖掘中。在我们处理这样的语义信息中，我们发现有如下其有如下特点：第一，宫格名称通常较短，一般可以拆分成两个或者三个较短的地理位置单词（比如：“某市某小区物业处”短句可以通过分词工具拆分成“某市”、“某小区”、“物业处”三个词语）；第二，各个宫格名称极少有重复存在，这是因为一个地点往往只会建立一个基站，同时地理名称存在的意义就是区分不同的地点，避免重复。这导致我们在对宫格名称分词、汇总后进行大量的后续处理，保证其单词语义的有效性。最后，我们如何利用单词的嵌入式表达（Embedding[23]）作为该基站的向量表征，也是需要面对的问题。例如：如何获取宫格名称为“某市某大学图书馆”的基站的词向量表征。首先如何经过统计计算得得到“某市”、“某大学”、“图书馆”的单词向量？之后向量之间如何进行操作能够获得完整“某市某大学图书馆”的基站的词向量表征都是需要研究和考虑的。

(4) 对新建基站来说，能否利用少量仅有的初始信息（例如经纬度、宫格名称）以及如何利用这些初始信息去预测基站流量模式是存在的另一大难题。基站流量模式的预测在前文研究现状中表示，对领域的研究尚少，其主要原因是基站流量模式与基站地理位置分布、社会因素（人口移动，节假日）高度相关，造成其模式分布及其复杂，同时我们不是对已有的基站的流量模式进行预测，而是对新建

基站的流量模式进行预测（新建基站的特点是可用特征较少）。所以在模式预测之前需要我们对模式的空间分布进行分析，处理好少量可用的特征。这些工作的完成也是我们分析研究过程中的难点。

1.3 本论文的主要贡献

本文基于一个中国大型无线网络运营商在一个大型城市的基站网络的流量测量数据，对超过 7 千个基站、以月度为单位、持续一年的基站流量的静态和动态特征进行了观察与测量，并对基站流量的时间变化模式进行了聚类、分析和预测。本文贡献如下：

(1) 统计分析了该城市基站流量数据的时空分布。在空间上，我们发现全市范围基站流量和流量密度分布不均衡，市中心和非市中心流量差异明显；在时间上，我们总结了基站流量总体增长模式的特点是：基站流量总体呈上升趋势，各基站在关键时间节点（11 月、次年 2 月）的变化有所不同。现象促使我们进基站流量模式挖掘。

(2) 提出了一种新的基站流量演变模式的聚类方法。该方法基于基站月度总流量值在一年内的排序序列进行聚类。在我们数据集上的分析结果表明：该聚类方式对较短时间序列（不存在周期性）的涨跌特点有很好的描述，能够得到比传统聚类方法更容易理解的结果。

(3) 基于提出的聚类方法，我们对运营商的 7 千多个基站的流量演变模式进行了大规模聚类分析，获得了 6 种典型的基站流量演变模式，最主要的一种流量模式涵盖了 38.6%的基站，特点为总体流量呈上升趋势，期间 11 月份流量由高峰直转为次年 2 月的流量低谷。其它还包括：“春节返乡”、 “双 11”电商购物模式等。结合城市特点，我们对所有模式的特点与形成原因做出了解释，这些发现为运营商理解其基站的流量演变情况提供了有益的信息。

(4) 提出了一种基于地理位置和基站地址语义信息的基站模式预测方法，能够对一个新建基站的流量模式进行预测。因为应该新建基站的初始信息往往较少，因此我们创造性地将基站语义标签信息引入基站流量模式预测中，改进了基站流量模式预测的准确性。实验结果表明：基站词向量表征能够明显提升 LR 线性模型的性能，同时我们将两种特征组合送入模型中，性能都比单个特征有所提升，“经纬度+基站词向量表征（Average）”的特征组合在 XGBoost[24]上 F1-score 为 35%比仅仅加入“经纬度”特征高出 5%，提升(0.35-0.3)/0.3=16.6%；在 LR 上的 F1-score 为 35%比仅仅加入“经纬度”特征高出 20%，提升(0.35-0.15)/0.15=133%，其中两种模式的预测准确度较高。

1.4 本论文的组织结构

本文的组织结构如下：

第二章主要介绍本文所涉及的网络测量、模式挖掘等基本内容与基本概念，同时我们也介绍了本文中所用的机器学习模型与算法，并给出算法模型所使用的模型评价方法。最后我们频繁使用的开发工具做出了简单的介绍。

第三章首先对所用数据集进行介绍，对所使用的预处理方法进行介绍。基于真实的大规模运营商基站数据，我们分别从两个范围（全市角度和市中心角度）分别观察测量了：分析了基站流量的空间分布情况、区块面积的分布情况、流量密度的分布，并通过可视化的方式显示。之后我们从时间角度分析了基站流量的增长模式，并针对其特点做出解释说明。

第四章首先分析了基站流量的分布特点，明确了聚类所要关注的问题。之后介绍了 Rank-Based 聚类方法的内容，基于真实基站数据，我们多步骤地将众多基站聚类并发掘其中规律。针对每一类基站流量模式，我们根据该市的特点和中国的节日分析其形成的原因。最后我们将聚类结果与经典的聚类算法 K-means[25]比对。

第五章主要评估了作为基站流量模式的两类特征：经纬度和基站宫格名称。当需要预测新建基站的流量模式时，能够使用的信息往往很少，我们选取了两类典型的特征：地理位置特征（经纬度）、语义信息特征（宫格名称），并分析评估其作为预测基站流量模式特征的能力。我们首先可视化了六类流量模式的地理分区情况，得到了其空间分布特点。对语义信息特征，我们从宫格名称分词到词向量的获取，逐步建立了基站的词向量表征。我们将两个特征送入 XGBoost、LR 分类预测模型进行预测，并对结果做出分析。

第六章讨论了基站流量模式的多方面应用。前文得到了六种基站流量模式，我们对各类型在时间上的特殊变化结合该市特点、国内节日等做出了解释，也分析了各模式的空间分布特点。基于所有观察和测量，我们将从应用角度分析基站流量模式特点，并给出其应用场景。

2 技术背景

本章介绍论文中所应用的模式挖掘和机器学习算法。首先介绍网络测量的基本概念以及目前的 3 个研究领域；然后介绍模式挖局的相关概念和特点；之后又对论文中所使用的几种机器学习模型的原理和特点进行了简要的介绍；我们介绍了模型性能评估的方法以及评价指标；最后介绍了开发过程中使用的科学计算库与开发平台。

2.1 网络测量

网络测量是对网络各要素的量化过程。在网络建设过程中网络测量是必不可少的一部分，同时也是其他网络应用的基础。目前网络测量与分析主要分为 3 个研究领域：

(1) 网络参数测量：精准、定量地获取网络及其网络应用中的参数数据。通常网络测量的主要参数包括有：RTT[26]、带宽、时延、数据丢包率、网络流量、网络服务质量 QoS[27]等。网络参数测量一方面可以发现网络中存在的明显问题，例如某个参数突破阈值等，另一方面大量的网络参数也是后续网络模型化的基础。

(2) 网络的模型化：这是网络测量关注的核心问题，即建立准确的网络描述与模拟。这种网络模型能够抽象出当前的网络特点，同时也能够预测网络将要发生的行为。建立网络模型现在有多种手段，包括：利用统计学知识建立数学模型、利用机器学习中的聚类算法提取网络模型等。

(3) 网络控制：网络测量的应用部分，利用测量结果和网络模型中获取的信息，实现对网络资源的合理配置与使用。

2.2 模式挖掘

模式挖掘的概念是：从大量目标数据中提取典型行为的过程。模式挖掘是数据挖掘的子任务之一，同级的任务还有聚类分析、异常点检测、关联规则发现等，但它们都有相同的特点就是事先并不能确定最终的分析结果的某些特点。例如对某一领域数据进行模式挖掘，但是我们事先并不能知道最后提取出的模式的数量。所以模式挖掘是一个知识发现的过程，它结果与数据的处理与分析密切相关，其中包含的主要数据操作有：数据清理、数据集成、数据选择、数据转换、模式发现、模式评估等。

模式挖掘中的两个重要概念分别是：支持度和置信度。支持度是对某事件的支持程度，置信度是对某事件的可信程度。支持度和置信度从它们的计算公式我们就可以发现它们的作用就是对模式中隐含规则的强度进行量化，例如无关、正相关、负相关这样的

数值评价。支持度和置信度的计算公式如下：

支持度的计算公式：

$$\text{support（A=>B）= P（AUB）} \tag{2-1}$$

置信度的计算公式：

$$\text{confidence（A=>B）=P(B|A)= support(AUB)/support(A)} \tag{2-2}$$

模式挖掘主要研究、关注：模式类型、挖掘方法以及应用。模式挖掘目前主要挑战是在在海量数据中，发现的模式数量也十分巨大。为了解决该问题，我们可以对所提取的模式进行压缩、合并近似模式、求 Top-K 模式等

2.3 机器学习

机器学习是对算法和统计模型的科学研究，它的特点是计算机系统利用模式和推理而不是具体的计算机指令来执行特定任务。机器学习算法从“训练数据”中学习和构建关于样本数据的数学模型，在通过“测试数据”去执行任务中所需要的预测或决策。目前机器学习被广泛应用，例如机器翻译、人脸识别技术、无人驾驶技术等。机器学习目前主要有三种分类：监督学习、非监督学习和强化学习。下面我们简要这些类型的特点以及所包含的优缺点：

(1) 监督学习的特点是：数据包含了标签信息。数据的“标签”可以被认为是人为或者客观的数据属性，例如对一群人的标签可以是：男或者女。监督学习有两种任务：分类和回归。在分类任务中：机器在训练集中通过对样本特征的分析与计算，并比对样本的标签信息，不断调整数学模型的参数使其能够有最好的分类效果。典型的分类应用是工厂中的产品分拣系统；在回归任务中：与分类任务不同的是，回归任务往往能够产生“新的结果”。分类任务通过学习后能够将测试数据划归到某个已知的类型中，但是回归任务能够产生新的数值结果。比如通过学习过去几天的天气信息，我们能够根据预测未来几天天气，而不是将未来几天分成各种类别。

(2) 无监督学习的特点是：数据没有标签信息。现实生活中的数据通常符合这样的特点，所以这些无监督学习算法就是解决这种没有数据标签的数据如何完成指定任务的。无监督学习可以分为聚类和降维。聚类算法主要用于根据属性和行为对样本进行分组，这与之前提到的分类算法不同，聚类算法事先并不知道样本的组别。降维算法主要是通过找到共同点来减少数据集的变量，目前大多数数据可视化操作将使用降维算法来识别趋势和提取规则。

(3) 强化学习的特点是：通过“智能体”本身的学习历史和经验来做出决定。与监督和非监督学习不同，强化学习没有绝对意义上的对错之分。智能体通过不断的“试错”来和外界环境产生交互，以此来获得奖赏或惩罚，智能体主要目的就是获得利益最大化。

强化学习的由来也是借鉴了人类在学习新知识时候的过程。

本文涉及多种机器学习算法，下面我们将简要介绍各种算法的基本原理与特点。

2.3.1 K-NN 算法

K-NN 算法的英文全称是 k-NearestNeighbor，中文名称是 K 近邻算法，是一种用于分类和回归的有监督机器学习算法。算法的主要思想是：一个样本的类别主要由与它最近的 K 个样本的类别决定，同时这个样本也将具有该类别所有的特征。K-NN 算法不但可以参与分类任务同时也可以参与回归任务，之前提到样本类别由邻居样本的类别决定，同时也有邻居样本的特征值，如果将所有邻居样本的特征值进行加权平均，那么就能够得到样本的特征回归值。K-NN 的特点对数据的局部结构敏感，该特点决定了它的决策边界是非线性的，那些具有重叠、非线性可分的数据将特别适合使用 K-NN 算法处理。K-NN 算法的具体步骤如下：

1. 将初始化距离设置为最大值；
2. 计算目标样本和其他训练样本的距离 distance；
3. 得到目前 K 个最近邻样本中的最大距离 max；
4. 若 distance<max，视该训练样本为 K-Nearest Neighbor 样本；
5. 重复 2、3、4，遍历所有未知样本并计算距离；
6. 统计 K- Nearest Neighbor 样本中各类标号的出现次数；
7. 给未知样本标记为 6 中出现频率最大的类标号。

算法的优点是：不需要进行参数训练、简单容易理解。

算法的缺点是：每次迭代需要进行大量的计算、不能够像决策树那样给出具体规则。

2.3.2 K-Means 算法

K-Means 算法中文名称是 K 均值算法，是一种用于聚类的无监督机器学习算法。算法的主要思想是：在样本内随机选择几个初始点，随后所有的样本点计算与初始点的距离，并将样本点分配给最近的初始点作为一类。之后通过计算更新每一类中的质心作为下一次的“初始点”，一直迭代直到符合某个结束条件。

K-Means 算法中最重要的步骤是初始质心的选取。通常使用的方法是在样本点中随机设置初始质心，但是这种方法的效果一般不好。第二种方法是：首先利用层次聚类的方法将样本点分成若干小类，最后从这些类别中选取初始质心，由于先验知识的加入通常这种选取质心的方法会取得一定的效果。但是这种方法有使用条件：其一是样本的数据量较小，这是因为层次聚类的计算量大；其二是 K 的数值远远小于样本点的个数。第

三种方法是：第一个初始质心可以选择所有样本点的质心，之后每一个新增质心都是距离初始质心最远的一个点，直到质心数量符合我们设定的值。这种方法避免质心选择的随机性，保证初始质心都是在数据中心分散开来的，但这种方法仍有缺陷是有可能选取到离群点。

基本的 k-means 算法流程如下：

1. 选取 k 个初始质心（作为初始 cluster）；
2. 重复：对每个样本点，计算得到距其最近的质心，将其类别标为该质心所对应的 cluster 重新计算 k 个 cluser 对应的质心；
3. 结束：直到质心不再发生变化。

对于欧氏空间的样本数据，通常以平方误差和（sum of the squared error, SSE)作为聚类的目标函数，同时也是衡量不同聚类结果好坏的指标：

$$SSE=\sum_{i=1}^{k}\sum_{x\in C_i} dist\left(x, c_i\right) \tag{2-3}$$

K-Means 算法的优点是：算法的可解释型强、算法简单且需要的调节的参数少。

K-Means 算法的缺点是：K 值的选取不好把握，迭代式算法只能得到局部最优解、对类别严重失衡的数据聚类效果不佳。

2.3.3 XGBoost 算法

XGBoost 算法的英文全称是 Xtreme Gradient Boosting，中文名称是极端梯度提升算法，是一种可以大规模并行的 Boosting[28]集成算法。算法的基本思想是：在学习任务中不断地添加树，同时不断地进行特征分裂来生长一棵树。每次添加一个树，其目的就是学习一个新函数，去拟合上次预测的残差。当我们在训练集中训练出 K 个树后，需要给样本一个分数，分数的计算过程是将把样本的特征落在所有树上的叶子节点分数加在一起，就是样本的预测值。

XGBoost 是目前训练速度最快、效果最优的开源机器学习算法之一，在各种数据竞赛中取得了不俗的成绩。XGBoost 最大的优势是并行化操作，原因是 Boosting 算法的特点是当前的树计算依赖于之前的结果，所以是一种串行结构，但是 XGBoost 巧妙的解决了该问题，它在枚举特征选择最佳分裂点时将各个树串行操作，同时枚举特征计算最佳分裂点也是树计算中最耗时的部分。XGBoost 的另外一大优点是对缺失值的处理：当样本的某个特征缺失时，模型会默认将其默认地分到某个节点，具体是左边节点还是右边节点需要通过计算，具体方法是假设该特征是左边或者右边节点，然后分别计算其信息增益，最终选择信息增益大的一边节点。

XGBoost 模型特点包括以下：

(1) 有多重防止过拟合策略：正则化项、列采样等；

(2) 目标函数可自定义：只要是目标函数二阶可导数即可；

(3) 内置交叉验证[29]、Early Stopping[30]：预测结果很好时提前停止建树加快训练；

(4) 样本权重自定义：该权重体现在一阶导数 g 和二阶导数 h，通过调整权重可以改变样本的重要程度。

2.3.4 LR 算法

LR 算法的英文全称是 Logistic Regression，中文名称是逻辑回归算法，是一种基于 Sigmoid 函数的二分类有监督机器学习算法。Sigmoid 函数[31]是函数曲线左右两端分别趋近于 0 和 1 的单调递增曲线，通常的用途是模拟二进制变量，可以将无穷大的区间线性投影到 0 到 1 范围内。算法的基本思想是：先对样本数据建立线性模型，然后借用 sigmoid 函数将其取值范围投影在 0 到 1 范围内，这样就可以把输出结果当做该样本属于某类型的概率，最后针对目标函数使用“极大似然法”去进行参数优化。“极大似然法”是统计模型中估计参数的通用方法，这里不做赘述。

LR 算法是一种二分类的模型，但是可以通过一些技巧实现多分类功能。第一种策略是：一对多，即使用 LR 算法每次只从中队类型中选择一种，而算法重复 N 次（N 代表总分类数）；第二种策略是：每次只训练一对类别，比如有三种类别分别是 A、B、C 那么每次需要训练的分类器是：AB，AC，BC。

LR 算法的特点包含以下几个方面：首先 LR 算法原理简单，所以计算的效率很高，对内存的占用不高，很适合大型数据处理和在线机器学习。同时 LR 算法对数据中的小噪声具有很好的稳定性，轻微的数据异常不会对算法的性能产生大的影响。另外普通的线性回归产生的分类决策面是所有样本共同决定的所以在类别失衡时，分类决策面会严重倾斜，但是 LR 算法很好的克服了这个问题，其思想是 LR 算法通过线性映射，使得远离决策面的样本点其作用减弱。最后 LR 可以配合多种正则化[32]手段来防止过拟合，包括 L0、L1、L2 正则化。

逻辑回归的目标函数是：

$$L(w,b)=\sum_{i=1}^{N}\left[y_i\log(\phi(x_i))+(1-y_i)\log(1-\phi(x_i))\right] \tag{2-4}$$

其中 Xi 为样本的观测值，Yi 为对应的标签值，W 是每个 Xi 的权重，b 是每个 Xi 的偏置值。

2.4 模型性能评估

2.4.1 分类与聚类

在机器学习和统计中，有两大类任务分别是：分类与聚类。接下来我们将简要介绍各自的特点以及应用场景：

(1) 分类：其在机器学习中的基本概念是在数据集中学习样本特征与样本标签的映射关系，并能够将新样本划分到合理的类别中。分类是一种有监督机器学习任务，它最大的特点是事先明确了样本所有的类别。分类的主要应用场景是给样本打上相应的标签，但它也是在回归的基础上的，比如机器学习模型将利用 Softmax 函数[33]将连续值转化成样本属于某个类别的概率，最终实现分类的功能。但是分类没有取近似的含义，类型之间区分的很明确。分类的局限性在于它的前提假设是所有的数据都是有相应的标签来对应，但是现实数据中并很多时候不能够满足这个条件的。面对海量的数据，想通过数据处理的手段使得所有数据都具有相应而且准确的标签，其代价是巨大的。目前分类的主要应用是：图像目标识别、异常点检测、产品分拣系统等。

(2) 聚类：其在机器学习中的基本概念是实现并不知道样本的任何标签信息，通过某种机器学习算法可以将这些样本数据划分成若干类型。聚类分析的主要问题是如何将混在一团的数据聚合成若干个有意义的类型，因为事先并不知道需要聚类的个数，所以需要我们制定合理、可解释的聚类规则。目前聚类算法在多个领域有应用，例如机器学习，模式挖掘，信息压缩，数据降维，蛋白质结构分析等。

聚类分析是模式挖掘的主要手段，本文对基站模式挖掘主要采用了聚类思想。所以对聚类方法的划分情况做简要的介绍：

(1) 基于划分的方法：计算聚类对象的某些衡量参数以及簇的表示方法不同，基于划分的方法主要包括有 K-Means[34]，k-中心点算法。

(2) 基于层次的方法：其主要原理是将数据划分层次，在每个层次上将数据聚类。层次聚类的方法主要有两种分别是：汇聚法和分裂法。

(3) 基于网格的方法：基于网格的聚类方法，主要使用多分辨率的网格数据结构，把对象空间量化为有限数目的单元，形成一个网格结构，所有操作都在这个网格结构上进行这种方法的主要优点是处理速度快，处理时间独立于数据对象的数目，只与量化空间中每一维的单元数目有关代表性的是 STING 算法[35]。

(4) 基于密度的方法：之前几种聚类方法主要使用各种定义的距离来描述数据之间的相似性，这样的聚类方法对于大部分的球形分布的数据效果较好，但往往对任意形状的数据往往结果较差，甚至无法进行有效聚类。针对这样的不足，提出了基于密度的聚类方法。这类方法将簇看作是数据空间被低密度区域分割开的高密度区域该类算法除了

可以发现任意形状的类，还能够有效去除噪声。典型的基于密度的聚类方法包括DBSCAN[36]。

2.4.2 模型评价方法

在建立模型之后我们需要对模型的泛化能力进行评估，这就需要有切实可行的估计方法，同时还要有度量模型泛化能力的评价指标。通常我们会根据不同的任务选出适合的任务指标。目前的对于有如下的评价指标：

(1) 对于回归任务：RMSE(平方根误差)、MAE（平均绝对误差）、MSE(平均平方误差)、Coefficient of determination[37] (决定系数)。

(2) 对于分类任务：精度、召回率、精确率、F 值、ROC-AUC 、混淆矩阵、PRC。

(3) 对于聚类任务：兰德指数[38]、互信息、轮廓系数。

在文中我们首先给每个基站找到了合理的模式标签，最后预测了新建基站的模式类型，所以使用的是分类模型。这里需要提前说明的是分类器在测试数据集上的预测或正确与否，我们记录了四种情况出现的总数分别记作：TP (True Positive)，将正类预测为正类的样本数、FN (False Negative)，将正类预测为负类的样本数、FP (False Positive)，将负类预测为正类的样本数、TN (True Negative)，将负类预测为负类的样本数。接下来我们将重点介绍分类模型的各种评价指标，具体如下：

(1) 错误率和精度：这是分类任务中最常用的两种性能度量指标，既然可以用于二分类任务也可以用于多分类任务。所谓的错误率就是分类错误的样本占所有样本的比例，而精度则相反，是分类正确的样本占所有样本的比例。它们反映了分类器整体的分类效果。这种评价方式其实具有一定的缺陷，即在数据类别不均衡的情况下对分类能力的评价并不客观。例如有如下分类问题：总共有 10 个苹果和 1 个梨子，那么将所有样本都分类成苹果的精度是 10/(10+1) = 90.9%但是我们知道这种分类器的分类能力是不好的。

(2) 精确率（Precision）：是指被分类器判定为正的样本中真正的正样本的比重，即被分类器判为正的所有样本中有多少是真正的正样本。计算公式如下：

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{2-5}$$

(3) 召回率（Recall）：是指被正确判定的正样本占总的正样本的比重，即所有正样本有多少被分类器判为正样本。计算公式如下：

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{2-6}$$

(4) F-score：是精确率和召回率的调和均值，计算公式如下：

$$F_{\beta}=\frac{(1+\beta^{2})\text{Precision? Recall}}{\beta^{2}\times\text{Precision+Recall}} \tag{2-7}$$

2.5 开发工具

本节主要介绍本论文测量过程中中所用的开发工具，包括 Pandas 数据处理开源库、Numpy 数组运算开源库等我们用来清洗、整理数据使用的 Python 库，同时在可视化方面，我们使用了 Python 研发中比较流行 Matplotlib 库制作对应的各种示意图，最后我们介绍了预测新建基站流量模式的 Scikit-learn 机器学习集成库。

2.5.1 Pandas 数据处理库

在 Python 是数据科学中对数据进行清洗、整理的工具。Pandas 适合处理多种类型的数据包括：具有不同数据类型列的表格数据，例如 SQL 表或 Excel 电子表格；有序或无序的时间序列数据；带有行和列标签的任意矩阵数据；任何其他形式的观测/统计数据集。Pandas 主要包含三种数据结构，分别是 Series(一维)，DataFrame(二维)，Panel(三维)。其中 Series 和 DataFrame 可以用于处理绝大多数金融，统计，社会科学和许多工程领域的典型问题。对于 R 用户而言，DataFrame 在支持所有 R 的 data.frame 的功能的基础上还能有更丰富的应用。Pandas 库建立在 NumPy 库之上，旨在与科学计算环境和许多其他第三方库完美集成。pandas 本身依赖 numpy 的，而 ndarray 在内存里占据这一段连续的内存空间，任何改变 ndarray 长度的操作都势必让所有 value 改变内存中的位置，也确实比较慢。Pandas 的优势主要集中在：

(1) 直观的合并，连接数据集操作；

(2) 方便重新定义数据集形状和转置操作；

(3) DataFrame 和 Panel 都可以删除或插入列；

(4) 分组功能可以对对数据集进行拆分组合，十分灵活方便；

(5) 基于智能标签的切片，花式索引，轻易从大数据集中取出子集；

(6) 轴(axes)的分层标签（使每个元组有多个标签成为可能）；

(7) 浮点数据及非浮点数据类型的缺失值(NaN)处理非常方便；

(8) 包括 NumPy 等数据结构中不同类索引的数据都可以转换 DataFrame 对象

Pandas 库对于统计科学家来说，在分析数据时是理想的工具，非常适合应用于数据清洗，分析/建模，然后将分析结果组织成适合于绘图或表格显示的形式的全部过程。Statsmodel 库依赖 Pandas 库，使其成为 Python 统计计算系统的重要组成部分。

2.5.2 Numpy 数组运算库

在我们数据处理的过程中，遇到这样的数据类型，它每个元素的数据类型都一致，但还是它们的数量众多（比如基站的坐标数据），在算法中为了加速算法的运算，我们使用的 N 维数组这种数据结构，Numpy 正是专门应对这种数据结构的工具。它本身基于 C 语言的开发，所以它的运算速度非常的快。

NumPy 是使用 Python 进行科学计算的基础包，它拥有一个强大的 N 维数组对象，支持复杂的（广播）功能，而且还可以用于集成 C / C 和 Fortran 代码，同时支持复杂的线性代数功能，有傅里叶变换和随机数功能 除了明显的科学用途外，NumPy 还可以用作通用数据的高效多维容器，可以定义任意数据类型。这使 NumPy 能够无缝快速地与各种数据库集成。NumPy 对象可用来存储和处理大型数学矩阵，比 Python 自身的嵌套列表结构要高效的多。NumPy 提供了一个 N 维数组类型 ndarray，它描述了相同类型的 Ttems 的集合。 因为 ndarray 中的所有元素的类型都是相同的，而 Python 列表中的元素类型是任意的，所以 ndarray 在存储元素时内存可以连续，而 python 原生 list 就只能通过寻址方式找到下一个元素，这虽然也导致了在通用性能方面 Numpy 的 ndarray 不及 Python 原生 list，但在科学计算中，Numpy 的 ndarray 就可以省掉很多循环语句，代码使用方面比 Python 原生 list 简单的多。Numpy 内置了并行运算功能，当系统有多个核心时，做某种计算时，numpy 会自动做并行计算。Numpy 底层使用 C 语言编写，内部解除了 GIL（全局解释器锁），其对数组的操作速度不受 Python 解释器的限制，效率远高于纯 Python 代码。

在实际编写代码的过程中，由于我们使用的数据类型比较一致，所以为了使得代码运行的效率更高，我们在多数时候使用 Numpy 进行数组的计算。

2.5.3 Matplotlib 绘图工具库

Matplotlib 是一个 Python 2D 绘图工具库，出于易用性的设计理念 Matplotlib 适用于各种结构数据、各种开发平台，可用于 Python 脚本，Python 和 IPython shell，Jupyter 笔记本，Web 应用程序服务器和四个图形用户界面工具包。Matplotlib 提供了方便的借口 API，只需几行代码就可以生成高质量的绘图，例如直方图，功率谱，条形图，热力图图，散点图等。其中的 Pyplot 模块提供了类似于 MATLAB 的界面，这让熟悉 MATLAB 的用户可以方便的过渡到 Matplotlib 中。

本文中我们所有图都是利用 Matplotlib 绘图工具绘制完成，方便了分析过程中结果的直观展示。

2.5.4 Scikit-learn 机器学习库

Scikit-learn 是开源的 Python 机器学习库，它基于 Numpy 和 Scipy，提供了大量用于数据挖掘和分析的工具，包括数据预处理、交叉验证、算法与可视化算法等一系列接口。Scikit-learn 的基本功能主要被分为六个部分，分类，回归，聚类，数据降维，模型选择，数据预处理。

从特征的预处理到最后的模型性能验证，它提供了简单的 API，仅仅使用简单的几句代码即可轻松实现一整套的机器学习流程。Scikit-learn 还有很强大的拓展性，包括与深度学习框架 Keras 的结合、吸收了最新的机器学习模型 XGBoost 并将其改造成方便调用的 API。Scikit-learn 的局限性是对大型数据并行化处理的欠缺，同时也不能够独立完成深度学习任务，但是对经典的机器学习任务有着很好的支持。

2.6 本章小结

本章主要介绍了论文研究中的技术背景，主要包括网络测量，模式挖掘的基本概念以及开发所采用的机器学习算法与模型，最后介绍了模型评估与开发平台，具体包括：

(1) 网络测量、模式挖掘的概念和特点；

(2) 论文中涉及的机器学习算法原理和模型性能的评估、度量方法；

(3) 模型的开发平台与使用的算法工具包

3 基站流量时空分析

本章主要进行基站流量时空分布的统计与分析。本章首先详细介绍使用的数据集以及特征处理细节，然后统计基站流量数据并分析其空间分布特点。再从时间角度，我们总结了基站流量的增长和变化模式，并结合该地区的城市特点提出了合理解释。最后我们讨论了统计分析结果对后续研究的作用与意义。

3.1 数据集介绍与预处理

本文研究所用数据集由国内某运营商部署在中国某发达城市的基站收集汇总，提供从 2018 年 7 月至次年 3 月份（包括 7、9、11 月份以及次年 2、3 月份）所有基站按月汇聚的流量使用数据。数据记录了 7110 个基站按月汇总的数据，包括了基站的月度总流量，以及各种详细类目的流量使用情况。下表 3-1 为使用部分的示例数据。

表 3-1 基站特征
Table 3-1 Feature Item of Base Station

特征	数值
基站序号	A1
宫格名称	XX 市高级中学
总流量 MB	363212
经度	E xx°xx′1.72″
纬度	N xx°xx′35.39″
月份	7 月

“宫格”的空间粒度适中，运营商也主要以此粒度制定商业政策，所以本文选择该粒度作为分析特征。“宫格名称”表示的是基站所处宫格的地理名称，“宫格”是运营商对目标空间的划分单位，在它的下级划分单位是“网格”。它的主要特征是在每一个宫格内所有活动都有逻辑的关联性，比如在商业区划分的一个宫格，在该范围内的活动往往都与商业活动相关。为了分析的需要，我们将每个宫格分配一个独立的基站，也即将少数处在同一个宫格的基站选择其中流量最高的一个。对于同一个宫格内存在冗余和冲突的流量日志以及基站位置信息不完整，我们进行了去重处理。

预处理主要包括两个部分：

(1) 数据预处理过程：主要包括两个步骤：首先，消除技术问题导致的冗余日志和冲突基站流量记录；其次，由于不同位置的基站流量差异较大，为了凸显出基站网络流

量自身变化特征，我们对每个基站流量做归一化处理。

(2) 空间划分：我们使用 Voronoi 图划分该市占地空间，确定每个基站所能够影响的区块。Voronoi 图在基站分析时候经常使用，它的含义是两基站的分界线是两点之间连线的铅直等分线，将全平面分为两个半平面，各半平面中任何一点与本半平面内基站的间隔都要比到另一基站间隔小。当基站数量在二个以上时，全平面会划分为多个包罗一个基站的区域，区域中任何一点都与本区域内基站间隔最近，所以这些个区域可以看作是基站的覆盖区域。

3.2 基站流量的空间分析

本小节主要统计分析该市基站流量的空间分布。由于该市占地广阔，市中心与周围地区基站之间差异明显，我们分别从全市和市中心两个范围对基站流量进行统计分析。我们将每个基站各月份的流量相加，获得该基站的流量使用总和，分析了基站流量的空间分布情况、区块面积的分布情况、流量密度的分布情况，最后以 Voronoi[39]图划分的基站区块绘制流量热力图直观展示流量空间分布的差异性。

3.2.1 全市范围空间分析

(1) 全市基站流量分布：其特点是大部分基站处于低流量区间，基站之间的流量极差大。我们对全市基站的基站流量情况做了统计，为了突出分布的偏度，我们将 PDF 图的 Y 轴和 CDF[40]图的 X 轴调整为对数表示，下文亦做此处理。具体情况如图 3-1 所示。

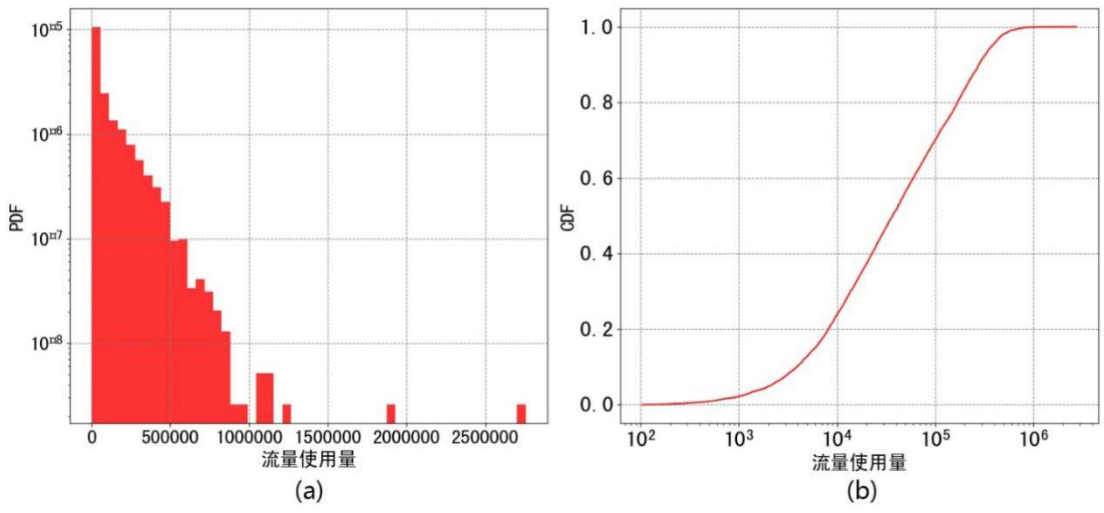

图 3-1 全市基站流量分布

Figure 3-1 Traffic Distribution of Base Stations in the City

图 3-1 中图(a)、(b)分别显示了基站流量的 PDF 和 CDF[41]。在(a)图中，随着流量使用量的上升基站数目会迅速减小，同时在两个流量极高的区间仍有基站分布，最高流量使用量达到了 2.76×10^{6} MB。由于图中纵坐标为对数表示，图中所显示的线性减少现象对应基站流量使用量的 PDF 呈类对数分布，大部分基站处于低流量区间。在(b)图中，50%的基站流量是低于 3.69×10^{4}MB，占最高流量 $3.12\times10^{4}/2.76\times10^{6}$ = 1.13%；75%的基站流量低于 1.31×10^{5}MB，占最高流量 $1.31\times10^{5}/2.76\times10^{6}$ =4.75%，总体标准差为：1.39×10^{5}，基站之间显示出了较大差距。如下文中图 3-7 所示，在地里分布上，流量使用量特点是由市中心流量使用量普遍较高，流量使用量由市中心向郊区方向迅速递减。

(2) 全市基站区块面积分布：我们首先根据 Voronoi 图对整个城市的基站辐射范围划分区块，经统计发现其特点是大部分基站区块处于小面积区间，基站区块面积极差大。我们分别绘制了区块面积的 PDF 图和 CDF 图，具体情况如图 3-2 所示。

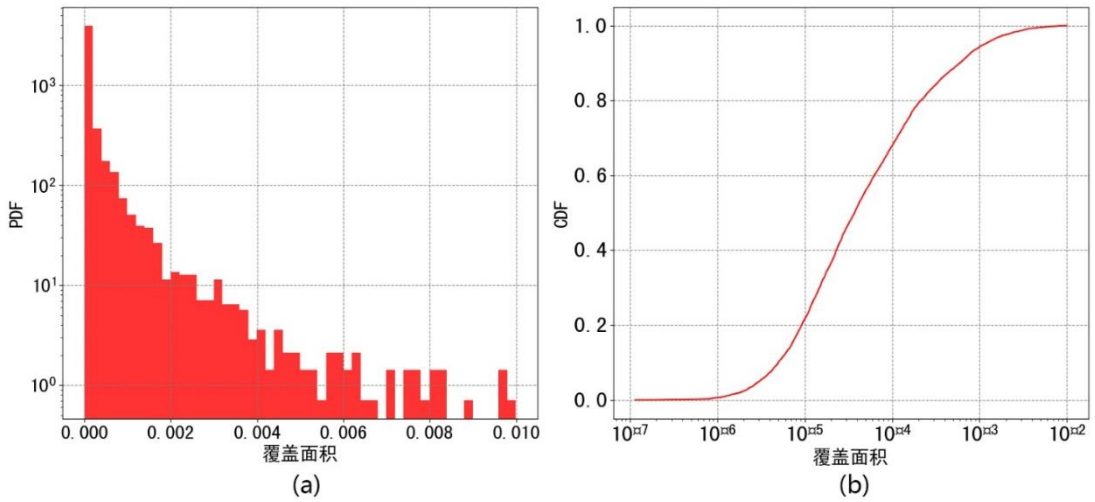

图 3-2 全市基站区块面积分布

Figure 3-2 Area Distribution of Base Station Blocks in the City

图 3-2 中图(a)、(b)分别显示了全市基站区块面积的 PDF 和 CDF。在图(a)中，随着区块面积的上升基站数目会迅速减小，同时在面积为 0.006 处开始平缓缓同时出现不连续现象（值得注意的是区块面积直接由经纬度计算得来，由于比较的是区块之间相对大小，所以没有将其按照比例转换实际面积），同时最小区块面积达到了 1.15×10^{-7}。由于图中纵坐标为对数表示，图中所示的区块面积 PDF 呈类对数分布，大部分基站处于小面积区间。在图(b)中，25%的区块面积大于 1.52×10^{-4}，是最小面积的 $1.52\times10^{-4}/1.15\times10^{-7}$ =1320.6 倍；75%的基站区块面积大于 1.18×10^{-5}，是最小面积的 $1.18\times10^{-5}/1.15\times10^{-7}$ =102.38 倍，总体标准差为：6.88×10^{-4}，基站区块面积之间显示出了较大差距。如下文中图 3-7 所示，在地里分布上，基站区块面积的特点是由市中心区块面积普遍较小，区块面积由市中心向郊区方向迅速变大。

(3) 全市流量密度分布：我们用基站覆盖区块的总流量除以区块的面积得到该区块的流量密度，经统计发现其特点是流量密度呈类对数分布，中间出现断层分成了高低流量密度两个部分。我们分别绘制了区块面积的 PDF 图和 CDF 图，具体情况如图 3-3 所示。

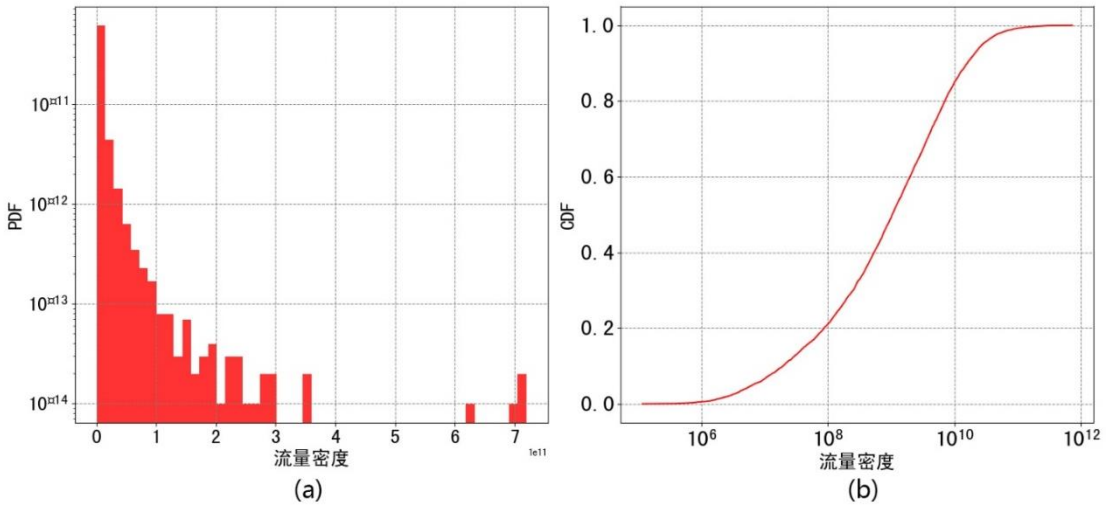

图 3-3 全市基站流量密度分布
Figure 3-3 Citywide Base Station Traffic Density Distribution

图 3-3 中图(a)、(b)分别显示了全市基站流量密度的 PDF 和 CDF。在图(a)中，随着流量密度的上升基站数目会迅速减小，减小的速度甚至超过上文所统计的流量使用量和区块面积，同时在流量密度为 6×10^{-6}（由于区块面积是相对值没有单位，所以流量也没有单位）处开始又出现流量密度极高的基站，最高流量密度可达 7.19×10^{11}。基站流量密度分成两个部分的原因是：市中心基站流量使用量高，同时基站区块面积小，相除之后能够达到很高的流量密度；郊区基站流量使用量低，同时基站区块面积大，相除之后流量密度较低。在图(b)中，50%的区块流量密度低于 1.03×10^{9}，占最高流量 $1.03\times10^{9}/7.19\times10^{11}=0.14\%$；75%的基站流量低于 5.12×10^{9}，占最高流量 $5.12\times10^{9}/7.19\times10^{11}=0.712\%$，总体标准差为：$2.52\times10^{10}$ 基站之间显示出了较大的差距。

3.2.2 市中心范围空间分析

前一小节统计和分析了全市范围基站流量情况，我们发现市中心与市中心周边基站在流量密度上出现分隔。因此我们选择了纬度[120. 219,120.318]，纬度[30.3272,30.3982]的市中心区域进行细致观察，该区域包含了 132 个基站区块。

(1) 市中心基站流量分布：其特点是市中心基站流量 PDF 分布呈“梳子状”，说明该区域的流量配比有一定规律，这启发我们探究基站流量模式。我们分别绘制了区块面积的 PDF 图和 CDF 图，具体情况如图 3-4 所示。

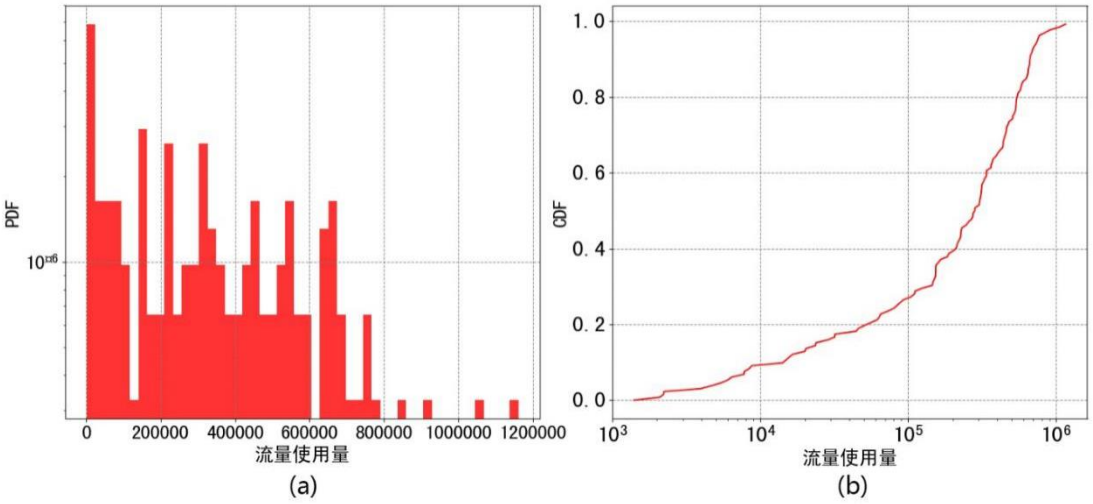

图 3-4 市中心基站流量分布
Figure 3-4 Downtown Base Station Traffic Distribution

图 3-4 中图(a)、(b)分别显示了市中心基站流量的 PDF 和 CDF。在图(a)中，我们能够看出基站流量分布较全市基站流量分布更加平稳，各个基站流量区间的基站数主要分成了一高一低两个数量级，同时一高一低交叉出现呈现出“梳子状”。比较有趣的是高流量区间和相对较低的流量区间 PDF 都有类似分布，形成该分布的原因可能是市中心区域被划分成不同的片区，片区内部流量分布有固定规律，同时片区越向市中心集中流量使用量基数就越来越大。在图(b)中，我们能够得到 25%的基站流量低于 8.31×10^4MB，占比该区域最高基站流量 1.16×10^6MB 的 7.16%；50%的基站流量低于 2.78×10^5MB，占比最高基站流量的 23.96%；总体标准差为：2.56×10^5 比全市范围标准差（1.39×10^5）要大的多，说明市中心基站之间的流量差异更加明显。作为比较，有一些重流量基站，例如该区域最高基站流量 1.16×10^6MB 是最低基站流量 1.4×10^3MB 的 828.57 倍。

(2) 市中心基站区块面积分布：其特点是市中心基站区块大部分面积偏小，基站区块面积差距较小。我们分别绘制了区块面积的 PDF 图和 CDF 图，具体情况如图 3-5 所示。

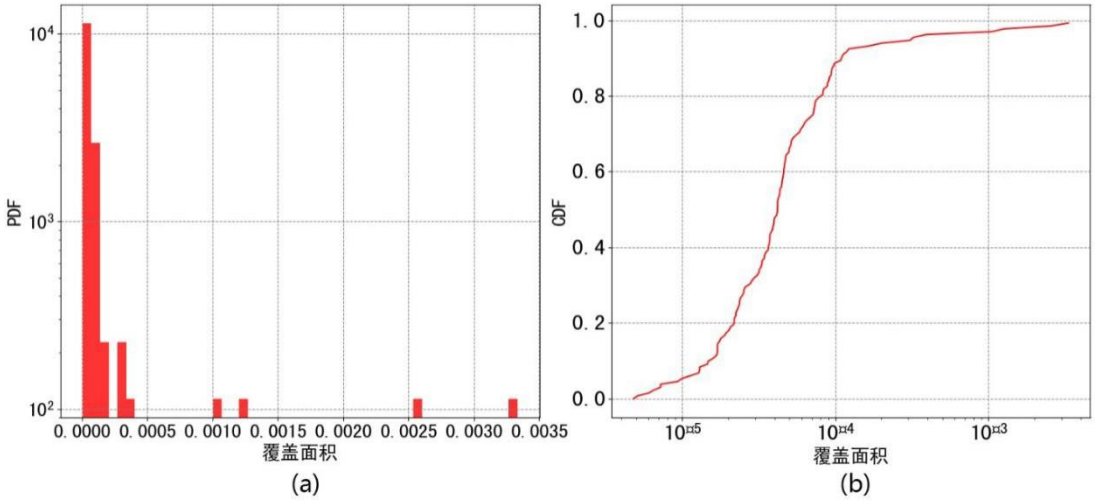

图 3-5 市中心范围基站区块面积分布
Figure 3-5 Block Area Distribution of Downtown Base Stations

图 3-5 中图(a)、(b)分别显示了市中心基站流量的 PDF 和 CDF。在图(a)中，我们能够看出基站区块面积普遍偏小，只有少数大面积区块。在图(b)中，我们能够得到 25%的区块面积大于 6.9×10^{-5}，是最小面积的 $6.9\times10^{-5}/5\times10^{-6}$ =13.8 倍；75%的基站区块面积大于 2.4×10^{-5}，是最小面积的 $2.4\times10^{-5}/5\times10^{-6}$ =4.8 倍，总体标准差为：3.85×10^{-4}，比全市范围的区块面积标准差 6.88×10^{-4} 要小，说明市中心的区块面积更加集中，基站之间区块面积差异没有全市范围的大。

(3) 市中心流量密度分布:其特点是流量密度之间的差距比全市范围内更明显。我们用基站覆盖区块的总流量除以区块的面积得到该区块的流量密度，如图 3-6 所示。

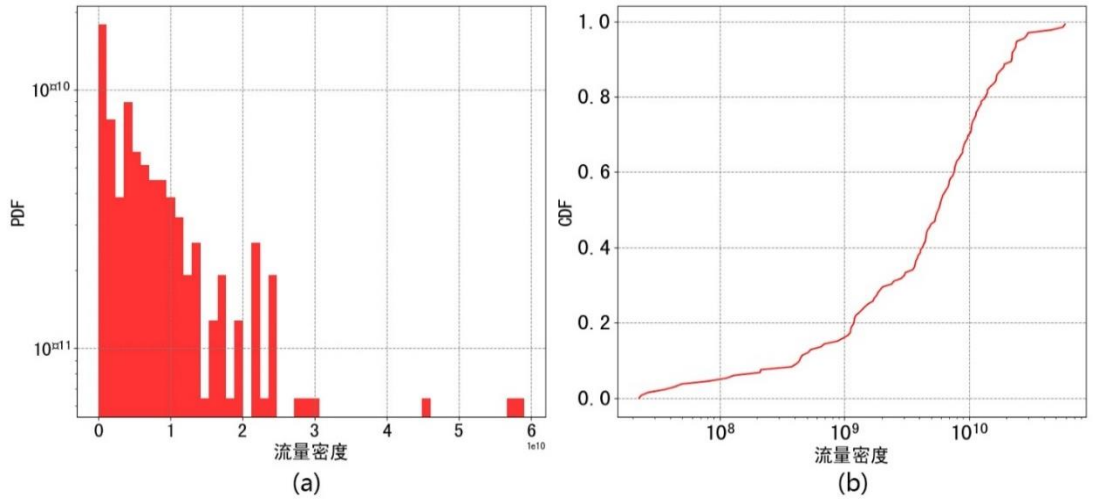

图 3-6 市中心区块流量密度分布
Figure 3-6 Traffic Density Distribution in Downtown Blocks

图 3-6 中图(a)、(b)分别显示了市中心基站流量的 PDF 和 CDF。在图(a)中，我们能够看出由于市中心面积分布较为均匀，所以基站流量密度与基站流量类似，在高流量密

度区间 PDF 出现“梳子状”，说明该区域的流量配比有一定规律。在图(b)中，50%的区块流量密度低于 5.62×10^{9}，占最高流量 $5.62\times10^{9}/5.90\times10^{10}$ = 9.52%；75%的基站流量低于 1.12×10^{10}，占最高流量 $1.12\times10^{10}/5.90\times10^{10}$ =0.712%，总体标准差为：9.99×10^{9}，比全市范围的流量密度标准差(2.51×10^{10})要小，说明市中心的流量密度比全市范围的流量密度分布更加集中。

(4) 流量热力分布：其特点是基站流量从市中心向周边延伸，基站流量渐进式递减，相邻基站流量分布不均匀。如图 3-7 所示。

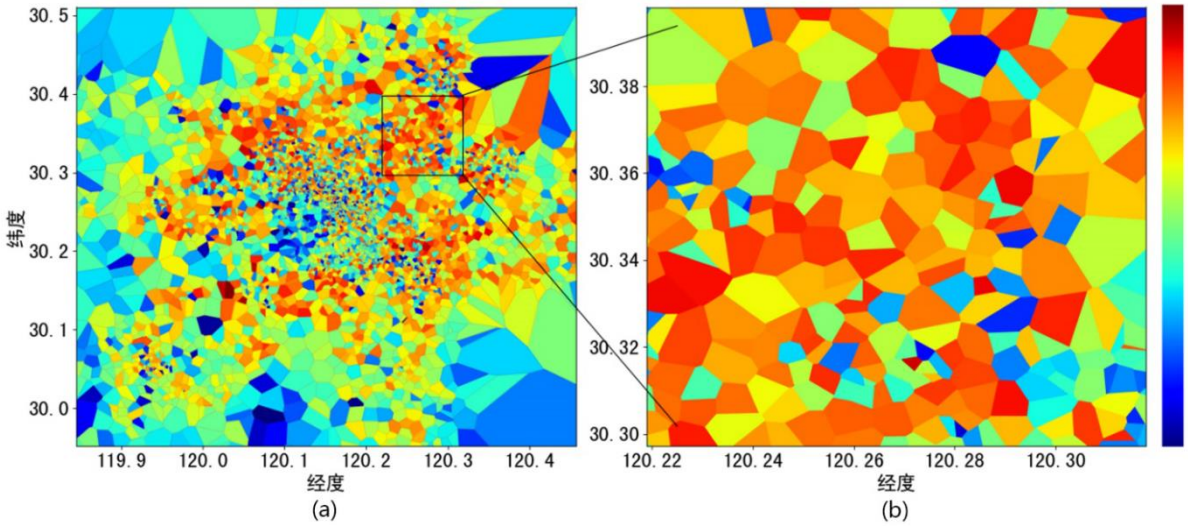

图 3-7 全市流量热力分布
Figure 3-7 Heatmap of the City's Traffic

图 3-7 直观显示基站流量从市中心向郊区的变化。(a)图表示全市范围的基站流量，(b)图为市中心区域的局部放大图。图中蓝色表示基站流量密度较低，红色表示基站流量较高。从(a)图可以看出市中心周围（也即郊区）的基站流量较低，向市中心方向基站流量逐渐升高，但是基站的区块面积却呈现出相反趋势，即越往市中心区域，基站的区块面积越小。(b)图显示的是市中心范围基站流量分布，从中可以看出该区域的基站流量整体偏高，但是分布不均匀，高流量基站周围往往伴随着低流量基站。结合之前统计分析结果、基站区块之间的流量差异与分布特点，我们觉得有必要探索基站的流量模式。

3.3 基站流量的时间分析

该小节主要分析基站流量随时间变化的特点。我们仍从全市和市中心两个角度分析。我们将各范围内的所有基站流量按月累加，它反映了区域内基站流量总体动态变化情况，我们绘制了这 5 个月内全市所有基站和市中心基站的流量变化折线图，如图 3-8 所示。

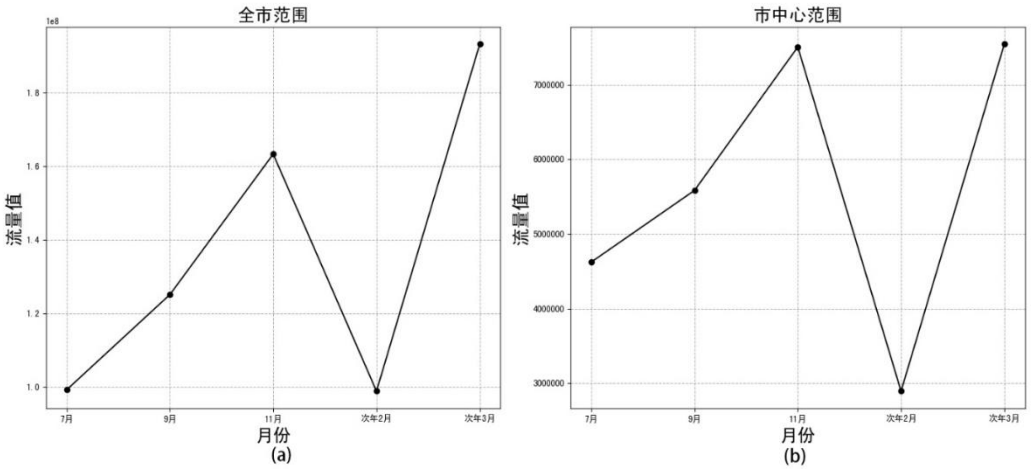

图 3-8 基站流量变化折线图
Figure 3-8 Line Chart of Base Station Traffic

从图 3-8 中，我们得出基站总体增长模式的特点是：基站流量总体呈上升趋势、11 月份与次年 2 月份是产生流量波动的关键点、市中心相比郊区基站流量增减幅度更大。(a)图和(b)图分别代表全市和市中心基站流量变化模式，二者整体波形类似，主要区别在 11 月份和次年 2 月份的流量变化幅度不同，具体来说：

(1) 随着时间的变化，区域的整体基站流量呈上升趋势；

(2) 在 11 月份，基站流量达到当年最高值。其原因是本月包含了一个特别的中国节日“双十一”电商购物节，大量用户使用移动端进行网络购物。该城市电子商务领域发达，导致该月份内基站无线网络流量激增；

(3) 在次年 2 月份，基站流量锐减。其原因是该月包含了中国节日“春节”，目前该城市存在大量外来人口，“春节”期间返乡或部分常住人口迁入郊区，导致基站流量骤降。图 3-8(a)、(b)两图的不同之处是：市中心基站流量比全市基站流量要下降更多，原因是从市中心离开的人口有部分短期（春节期间）迁入郊区，导致郊区基站流量的下降得到缓和；

(4) 在次年 3 月份，基站流量从低谷恢复到较高流量水平。其原因在于“春节”后迎来“返工潮”，外来人口迁回该城市或者常住居民返回市中心工作区，基站流量因为人口的重新聚集而迅速恢复。

我们仍选取纬度[120. 219,120.318]，纬度[30.3272,30.3982]的市中心区域，共包含 132 个基站区块，我们将市中心基站各月份（7 月、9 月、11 月、次年 2 月、次年 3 月）的流量热力图绘制出来，如图 3-9 所示。

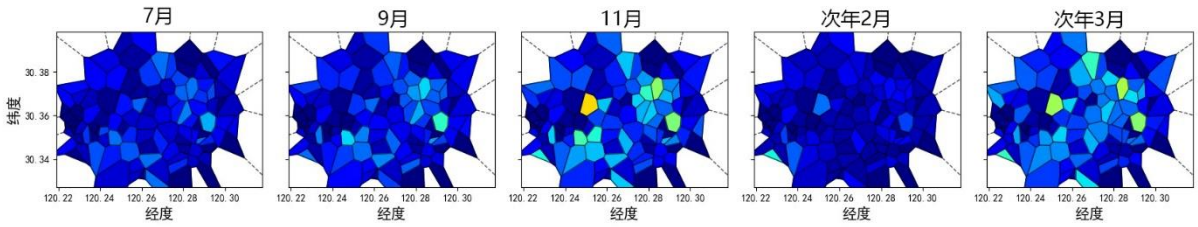

图 3-9 市中心基站流量变化热力图
Figure 3-9 Traffic map of downtown base stations over time

在图 3-9 中，将各子图连贯起来看，市中心基站流量的变化规律与前文所述一致，即在整体上升的趋势中，于“11 月份”、“次年 2 月份”产生了明显的流量升降波动。同时热力图也显示了相邻基站存在的流量差异，造成了基站流量空间分布不均匀性，高流量基站通常有许多低流量基站分布周围。

3.4 本章小结

本章从全市和市中心两个范围对基站流量数据进行统计和分析，同时从时间角度，总结了基站流量的总体增长模式。分析的结论是：

(1) 在空间上，全市范围基站流量和流量密度分布不均衡，市中心和非市中心流量差异明显。市中心基站流量的特点是：基站流量整体偏高、基站流量 PDF 分布有“梳子状”即该区域流量配比存在一定规律、基站区块面积小且分布集中，造成相邻基站之间流量密度变化大。

(2)在时间上，我们总结了基站流量总体增长模式的特点是：基站流量总体呈上升趋势，在关键点波动幅度大，分别对应包含“双十一”、“春节”的月份；对比全市基站，市中心基站流量变化幅度更大。

本章初步探索了基站的流量情况发现：基站流量在空间上差异较大，时间上在关键点处波动明显，并且这种波动和该城市的特点以及中国节日相关，这些启发我们更加细致挖掘基站的流量变化模式。

4 基站长期流量变化模式聚类

本章针对传统聚类方法的不足提出了 Rank-Based 基站流量模式聚类方法，并对 7110 个真实基站进行模式聚类分析。我们首先分析了传统聚类方式在对基站聚类上的不足，之后介绍了 Rank-Based 聚类方法的内容。基于真实基站数据，我们分多个步骤将众多基站聚类并发掘其中规律。根据该城市的特点和中国的节日，我们针对每一类基站流量模式都分析了模式特点和形成的原因。最后我们将 Rank-Based 聚类结果与经典的聚类算法 K-means 进行对比了分析。

4.1 传统聚类方法

传统聚类方法以内类距离作为聚类规则，但是它在基站流量涨跌的方面没有很好的解释性。下面将通过一个例子说明传统聚类方法在基站流量模式挖掘方面的不足，如下图 4-1 所示。

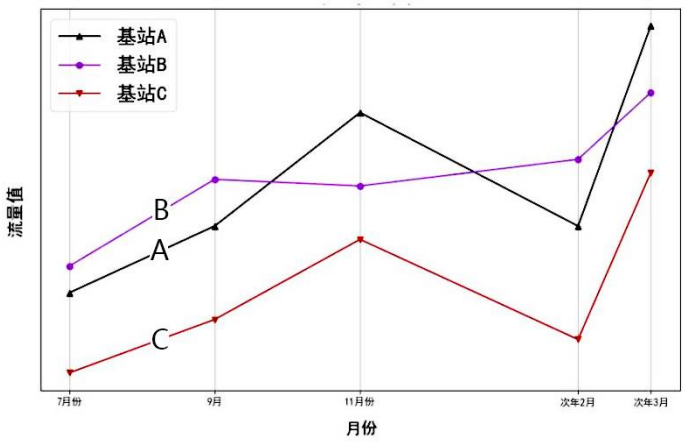

图 4-1 三个基站的流量折线图
Figure 4-1 Flow Chart of Three Base Stations

图 4-1 是三个基站在 7 月、9 月、11 月、次年 2 月、次年 3 月上月度总流量的使用情况，分别命名被为 A、B、C。从图中我们能够看出：

(1) 基站 A 与基站 B 的流量曲线靠的很近，但是这两个基站在 9 月、11 月、次年 2 月的表现却截然相反，基站 A 的流量走势是 9 月份到 11 月份流量上涨、11 月份到次年 2 月份流量下跌；基站 B 的流量走势是 9 月份到 11 月份流量下跌、11 月份到次年 2 月份流量上涨。我们认为将这两种类型基站划分到一类是不合理的，例如在“双十一”期间，两个基站有着相似的流量使用量，一个是流量下跌到该使用量，另一个却是流量上涨到该使用量，这两个基站对该事件产生反应是不同的，可能真是基站所处的地理位置

等因素造成了其不同反应的产生，所以它们不应划分到同一类型。

(2) 基站 A 与基站 C 的流量曲线靠的稍远，但是这两个基站在 9 月、11 月、次年 2 月的流量变化却非常同步。在基站 A 流量上涨时，基站 C 也跟着上涨；基站 A 流量下跌时，基站 C 也下跌，说明了这两个基站对待不同事件的反应是一致。在第三章的统计分析中，我们发现越是靠近市中心基站的流量基数就越来越大，但是基站流量的变化反映了基站自身的属性，例如学校建在郊区与建在市中心其流量基数有可能不同，但是 7 月份的假期来临，流量的变化是一致的。

综上所述，传统聚类方法能够将类内距离相近的基站聚在一起，但是却容易忽略流量变化，特别是“双十一”、“春节”时期基站流量变化。发现基站一致性的变化是运营商能够提前做好防范准备的基础，所有有必要提出符合真实基站数据的新的聚类方法。

4.2 Rank-Based 模式聚类方法

本小节我们首先详细介绍基站的 Rank 向量表征的含义以及构建过程，之后利用我们提出的聚类方法将 7100 真实基站由小类聚成大类，并从大类中提取具体的流量模式。最后基于聚类结果我们对每一个大类流量模式的特点作出分析并给出解释。

4.2.1 基站 Rank 向量

我们首先通过例子解释基站流量的 RANK 向量的含义，如表 4-1 所示。

表 4-1 某基站流量示意表
Table 4-1 A Base Station Traffic Diagram

基站序号	7 月	9 月	11 月	次年 2 月	次年 3 月	RANK 向量
A1	100GB	110GB	200GB	50GB	210GB	[2,3,4,1,5]

表 4-1 是某基站各月份的使用情况，由此我们可以得出该基站流量的 RANK 向量为[2,3,4,1,5]，其中向量里面的“2”表示 7 月的流量在所有月份的排名为第 2 名，其中向量“3”表示 9 月的流量排名为第 3 名，以此类推。各维度的数值越高表示该月份基站流量越高。

使用 RANK 向量表征基站流量的原因是：真实运营商基站流量数据时序点较少，对于短期时间序列（不存在强周期性），描述它的波形主要是依靠它自身流量的相对位置，而我们这种基于 RANK 向量的分类方式正是抓住了短期时间序列的这种特点。

4.2.2 模式聚类过程

下面我们将多步骤地对运营商 7110 个真实基站数据使用 Rank-Based 聚类方法进行聚类，具体的步骤如下：

(1) 对所有基站的流量进行归一化。其原因是：从空间分布测量中我们发现，市中心周边地区分布大量低流量基站，造成该现象的原因是市中心周边人口稀少，基站流量使用量低。市中心基站人口密集其流量普遍偏高，同时基站之间流量极差非常大，部分基站流量能达到极高的值。为了凸显基站流量自身的变化，我们对基站流量按照自身流量的最大值做了大小归一化。

(2) 计算基站 Rank 向量，并将具有相同 Rank 向量的基站合并。7110 个基站将会计算出 7110 个 Rank 向量，但是其中有许多重复的 Rank 向量，基站具有相同的 Rank 向量表示基站的流量变化特点一致。据统计，我们按照每一种 Rank 向量所包含的基站数量排序后，80%的基站都能划归到前 35 种基站模式中。最终我们将基站数量排名前 50 的 Rank 向量作为我们模式聚类的小类模式，合并的结果如图 4-2 所示。

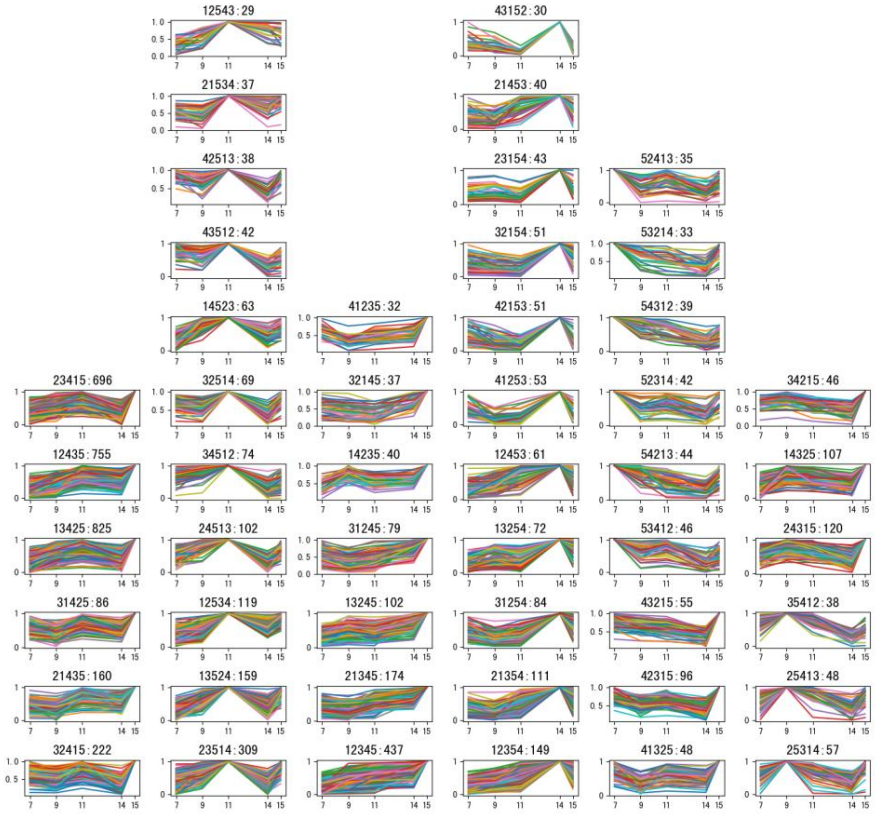

图 4-2 排名前 50 的小类模式
Figure 4-2 The Top 50 Subclass Patterns

图 4-2 中有 50 个子图，表示 50 种不同的 Rank 向量，每个子图包含了具有相同 Rank 向量的基站。子图标题由两部分构成：子图的 Rank 向量和所包含基站的数目。这里值

得说明的是，在图 4-2 中 50 种小类模式的排列位置暂与模式挖掘无关，只是为了与后文协调一致。

(3) 将 50 个小类模式再合并成大类模式，下面我们将分基本思路和具体做法来讨论：

1) 基本思路是；首先从 50 小类模式中找到与总体流量模式特点一致的模式作为“流量模板”，剩余小类模式能否合成出一个新大类主要取决于两点：一个是要有与“流量模板”不同的特点，另一个是包含的基站数量要够。

2) 具体做法是：首先前文介绍了总体流量增长模式的特点主要有：11 月份流量突然增加、2 月份流量突然减小、7,9,11 月份流量呈现增长趋势。根据以上三个特点，我们能够从 50 小类模式中分离出符合特点的小类模式（6 种），并且合成一个大类模式：“主模板流量型”。它的含义是大部分基站符合的流量模式，经统计单这一种流量类型就包含了 38.6%的基站。然后我们逐一比对上述的三个特点，例如特点是“11 月份流量突然增加”，我们在剩下的基站 Rank 向量中搜索不同于该特点的基站，同时在保证该类型具有一定数量基站的要求下，我们发现了如下图 4-3 所示从左数第四列的流量类型，它的特点是流量持续上升，11 月份（包含“双 11”）流量达到所有月份最高值。我们以此类推总共可以得到 6 大类流量模式，同时我们将剩余 20%的基站划归为“其他类型”，所有类型如图 4-3 所示。

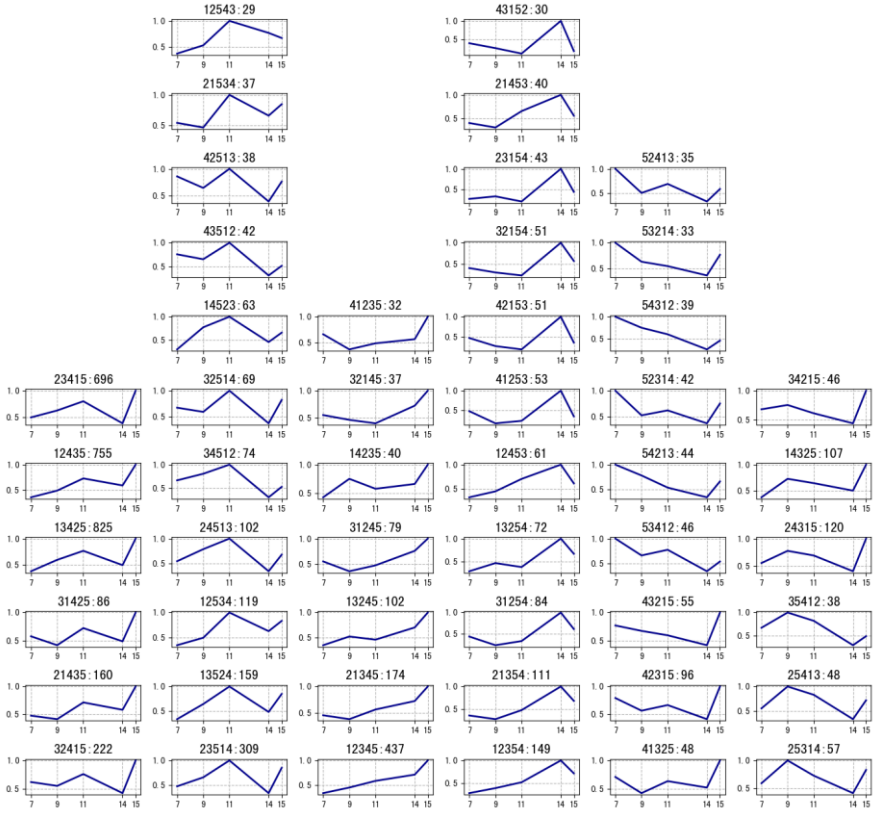

图 4-3 取均值后的前 50 名小类模式

Figure 4-3 The Top 50 Subclass Patterns After Taking The Mean

图 4-3 为 50 个小类模式合并成 6 大类模式的结果图。图中每一列表示一种流量模式大类，同时该大类又包含一定数量的子类流量模式。从图中我们能够看出每个子类之间波形具有相似性，大类之间具有区分性。每个子图是该 Rank 向量下包含的基站在各个月份上求取流量均值的结果，代表了该 Rank 向量实际变化特点。

(4) 获取流量模式大类的波形特点。具体做法是：我们将每一个流量模式大类的基站聚合在一起，求取在各个月份上的流量均值。例如在图 11 中同属第一列的子模式均属于同一个流量模式大类，我们对这些基站在各个月份上求取流量的均值，对上述所有流量模式大类均做该操作即可得到图 4-4：

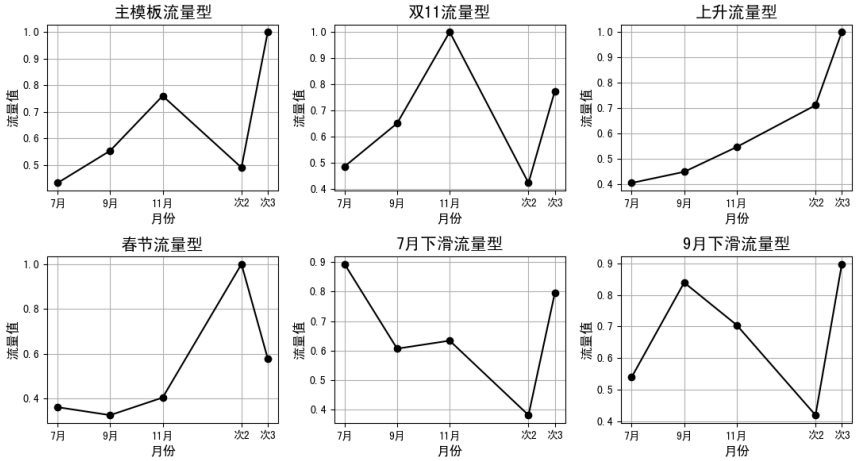

图 4-4 六种基站流量类型波形图
Figure 4-4 Waveform Diagram of Six Base Station Traffic Types

4.2.3 聚类结果分析

经过聚类我们得到了如图 4-4 中六种波形各异的大类流量模式，现在我们总结其大类流量模式的分类规律，如下面表 4-2 所示。

表 4-2 六种基站流量类型
Table 4-2 Six Base Station Traffic Types

流量模式类型	包含基站数目	特点
主模板流量型	2744	持续上升，2 月份“春节”下降，后恢复增长
双 11 流量型	1041	持续上升，11 月份“双 11”流量达到最高
上升流量型	901	持续上升，无流量跌落
春节流量型	745	持续上升，2 月份“春节”流量达到最高
7 月下滑流量型	438	流量下滑，从 7 月份下滑，3 月份“春节”后回升
9 月下滑流量型	416	流量下滑，从 9 月份下滑，3 月份“春节”后回升
其它	825	无明显类型特点

(1) 主模板流量型：该模式与前文基站流量增长模式一致，其主要特点是该类型基站占比高，约一半基站流量遵循该流量模式。形成原因可能为：由于手机移动服务发展，包括运营商推出的优惠套餐和流量价格不断走低，移动流量使用呈现增长趋势。此外，除了基站流量呈上涨趋势外，后面将有两次明显流量波动。第一次波动发生在 11 月份，其原因为该月份包含“双 11”电购物节，作为电商发达城市流量会迅速上升。 第二次波动发生在次年 2 月份，其原因为该月份包含中国特色节日“春节”。该市电商与旅游业发达，流动人口比例高，春节期间将有大量非常驻人口“春节返乡”导致基站流量下降。此外，作为高度发达的电子商务城市，春节期间也有诸多限制影响流量使用，如春节期间快速停运等。3 月份基站流量迅速恢复，主要原因是春节假期过后，流动人口从其他城市返回该城市，恢复正常工作。该流量类型的形成与地区特点，节假日，以及人口流动密切相关。

(2) 双 11 流量型：该流量型是主模板流量型的变体，区别是 11 月份流量达到最高值，同时 2 月份基站流量下降幅度更大。该类型基站将在 11 月份，特别是“双 11”给基站带来巨大流量压力，运营商需将该类型基站作为重点关注基站。同时次年 2 月份流量将大幅度回落，如果不及时重新规划网络资源（频谱，耗电等）将造成资源浪费。

(3) 上升流量型：该流量型是主模板流量型的变体，区别是在关键点处（11 月份、次年 2 月份）没有大幅度涨跌，流量保持稳定增长。该类型基站覆盖区域对流量需求稳定，受到外来人口、特殊事件（双 11、春节）干扰小。

(4) 春节流量型：该流量型特点是流量持续上涨，在次年 2 月份“春节”期间内流量出现大幅上升。通过后面各流量类型地理分布可知，该类型主要集中在郊区。形成原因可能是“春节”期间该市常驻人口由市中心区域向周边转移导致“春节”期间流量上升。运营商需要针对“春节”流量高峰基站做好提前预防的措施，如增加资源、增补临时基站。

(5) 7 月下滑流量型：该流量型特点是从 7 月份开始流量下滑至次年 2 月，次年 3 月回升。具有该流量型的基站数目较少，但是市中心分布范围较广，在风景区能够观察到部分该类型的基站。

(6) 9 月下滑流量型：该流量型特点是从 9 月份开始流量下滑至次年 2 月，次年 3 月回升。具有该流量型的基站数目较少，比较明显的特点学校附近有许多这样流量类型的基站，形成的原因可能是学生暑假结束返校。

我们比对了 K-Means 聚类和文中提出 Rank-based 聚类的结果，如图 4-5 所示。图中显示了 7110 个真实基站按照两种不同的聚类方式所得到的结果，(a)图为 Rank-Based 聚类方法，(b)图为 K-Means 聚类方法。从图中能够看出 k-Means 聚类结果较为混乱，难以对应该市真实的环境；在 Rank-Based 聚类结果中基站行为变化一致，同时在上文中我们也对每一类基站流量型的特点以及形成原因给出了说明。

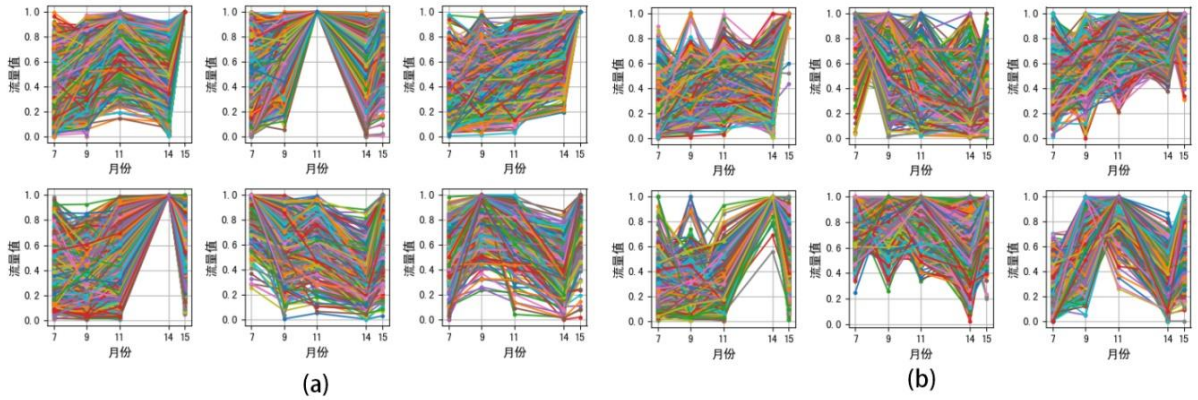

图 4-5 Rank-Based 与 K-Means 的聚类结果
Figure 4-5 Clustering Results Of Rank-Based And K-Means

4.3 本章小结

本章首先说明了传统基于类内距离的聚类方法在基站流量聚类上的不足是缺乏对流量变化的考虑，据此我们提出了 Rank-Based 基站流量模式聚类方法。我们引入了基站 Rank 向量表征短期基站流量变化特征，并对 7110 个基站进行聚类得出了 6 大类流量模式。根据该城市的特点和中国的节日，我们给出这 6 类基站流量模式形成的解释，最后我们将聚类结果与经典的聚类算法 K-means 进行对比分析。

5 基站长期流量变化模式预测

本章主要介绍我们提出的一种基于地理位置和基站地址语义信息的基站模式预测方法。我们评估了作为基站流量模式的两类特征：经纬度和基站宫格名称。当需要预测新建基站的流量模式时，能够使用的信息往往很少，我们选取了两类典型的特征：地理位置特征（经纬度）、语义信息特征（宫格名称），并分析评估其作为预测基站流量模式特征的能力。我们首先可视化了六类流量模式的地理分区情况，得到了其空间分布特点。对语义信息特征，我们从宫格名称分词到词向量的获取，逐步建立了基站的词向量表征。我们将两个特征送入 XGBoost、LR 分类预测模型进行预测，并对结果做出分析。

5.1 模式预测基本思想

对于新建基站，我们想要预测它的流量模式是非常困难的，原因在于新建基站的初始信息非常少。就所获得的数据集来看，对于新建基站我们能够获取的信息只有经纬度和所在的宫格名称。我们想充分利用这两个特征，探索基站流量模式的预测，所以问题描述为：利用新建基站的经纬度特征和基站宫格名称（语义信息）去预测它所从属的流量模式，这里的流量模式为前文聚类得出七类流量模式（包含其他类），属于七分类问题。关于经纬度特征和宫格名称的探讨如下：

(1) 经纬度特征：前文统计分析中我们发现，不同流量模式的基站与所处位置密切相关，例如市中心的基站和郊区的基站类型往往差异明显，市中心的某些区域基站流量会相同，下文将详细展开论述。

(2) 宫格名称：我们发现宫格名称包含某些词语的基站其流量模式会出现类似的情况。例如包含“学校”的基站宫格名称，其流量变化模式往往都为“双 11 型”，下文将详细展开论述。所以基站宫格名称引入模式预测是合理的，同时如何将语义信息转化成一种有效的向量参与建立模型是我们探讨的重点。

5.2 模式地理位置分布

本小节对基站地址经纬度与基站流量模式的关系进行分析。该市 7110 个基站由 Rank-Based 聚类方法划分成七大类（包含 Others 类型）。图中多边形为 Voronoi 图划分的基站区块，表示基站影响范围，我们使用不同颜色填涂基站区块，表示基站流量模式类型的不同，并且将其可视化于地理平面上。如图 5-1 所示。

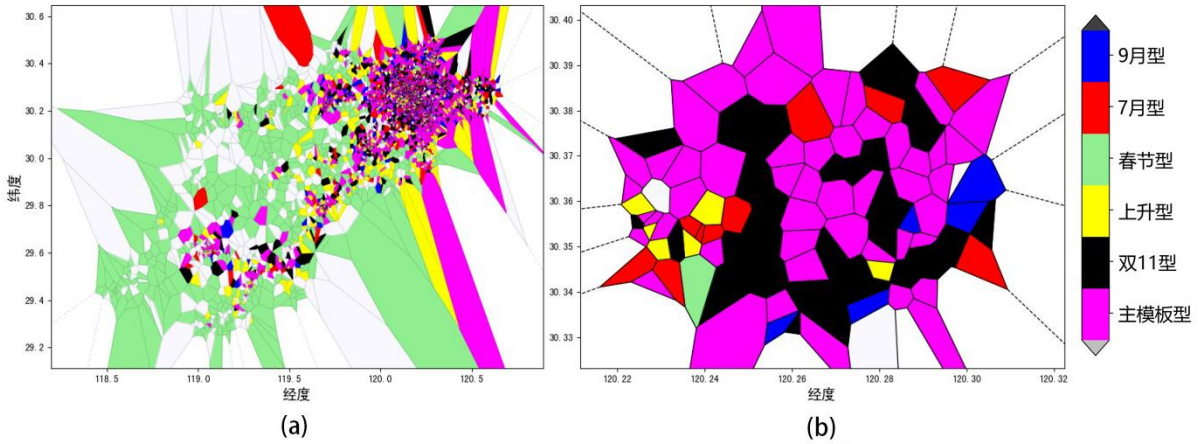

图 5-1 六种流量类型地理分布
Figure 5-1 Geographical Distribution Of Six Base Station Traffic Types

在图 5-1 中，我们绘制了两个视角范围的流量模式分布图，(a)图为全市范围，(b)图为市中心范围。(a)图中比较明显的特点是春节流量型主要分布在该市郊区，其原因与之前分析一致，主要是春节期间人口由市中心向郊区流动，导致该区域 2 月份基站流量突然增加。相比之下，其他流量模式则集中在市中心区域。在(b)图中，市中心存在“小集群”，集群内基站的流量模式相同，相邻集群之间模式截然不同。单个集群规模不大，但是整个集群数目众多，说明市中心区域模式分布关系复杂。

5.3 地址单词与流量模式关系

本节对基站宫格名称与基站流量模式的关系进行分析。我们将基站宫格名称拆分成更细粒度的单词，并分析其模式分布特点，比较不同单词基站流量模式分布的不同，评估单词和模式分布之间的相关性。

5.3.1 统计词频量

我们根据宫格名称的词语特点，汇聚、统计、过滤了经过分词后的词库。基站宫格名称代表该基站地理位置。“宫格名称”具有地理意义且语句较短，包含若干单词，例如

某基站宫格名称为“某市高级中学小区”。我们使用中文分词工具“Jieba 分词”将 7110 个基站的宫格名称进行分词并汇总，去除停用词、去重后共有 3005 个单词，我们对词频绘制了直方图，如图 5-2 所示。

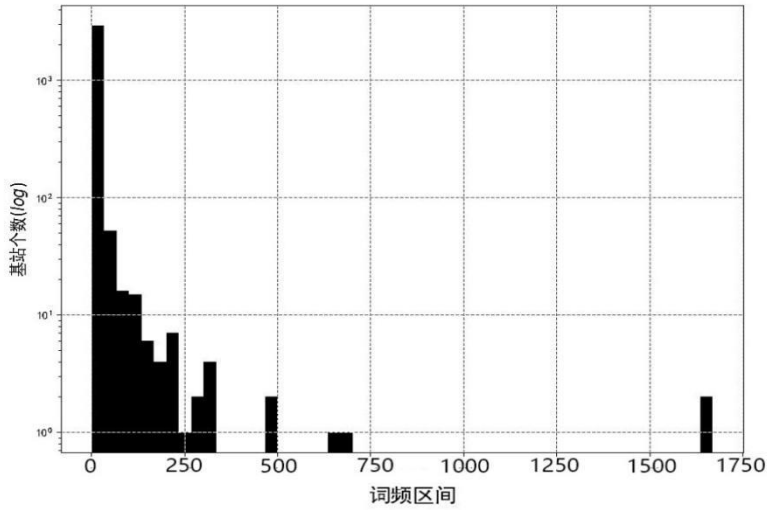

图 5-2 单词词频直方图
Figure 5-2 Word Frequency Histogram

在图 5-2 中，我们发现大部分词语主要集中在低频，需要过滤掉低频单词去除数据噪声。我首先过滤了词频小于 10 的单词，这个阈值将作为超参数，可供后期模型调优。

5.3.2 建立词向量

这里我们将介绍单词词向量的构建过程。对于之前过滤后的单词，我们将统计已经换划分成七类流量模式（包含其他类）的基站在每个单词上的分布。词向量建立的过程，例如单词“小区”的流量模式分布如表 5-1 所示。

表 5-1 某基站的流量类型分布表格
Table 5-1 Traffic Type Distribution Table of A Base Station

单词	1 类型	2 类型	3 类型	4 类型	5 类型	6 类型
小区	10 个	12 个	14 个	40 个	11 个	80 个

那么我们可以得到“小区”的词向量为：[10,12,14,40,11,80,10]

经过归一化后得到：[0.056 , 0.068 , 0.079 , 0.225 , 0.062 , 0.451 , 0.056]

该向量的物理意义是：包含该单词的基站属于某一类型的概率，则根据上述例子包含“小区”的基站其最有可能出现在类型 6 中。我们对所有过滤后的单词进行了统计，得出了所有单词的词向量表。

5.3.3 单词信息熵

我们引入信息熵，量化单词的词向量在某一类型上的集中度。在信息论中，熵是接收的每条消息中包含的信息的平均量，又被称为信息熵[42]、信源熵、平均自信息量。引入的想法是如果某个单词的词向量集中于其中一个或两个类型，那么熵值就会很小，即包含该单词的基站更将有可能集中于某一类型。

根据信息熵的计算公式：

$$H(X) = -\sum p(xi)\log(p(xi)) \quad (i=1,2,\ldots,n) \tag{5-1}$$

我们对所有过滤后的单词进行计算，并且统计了信息熵的分布情况如图 5-3 所示。

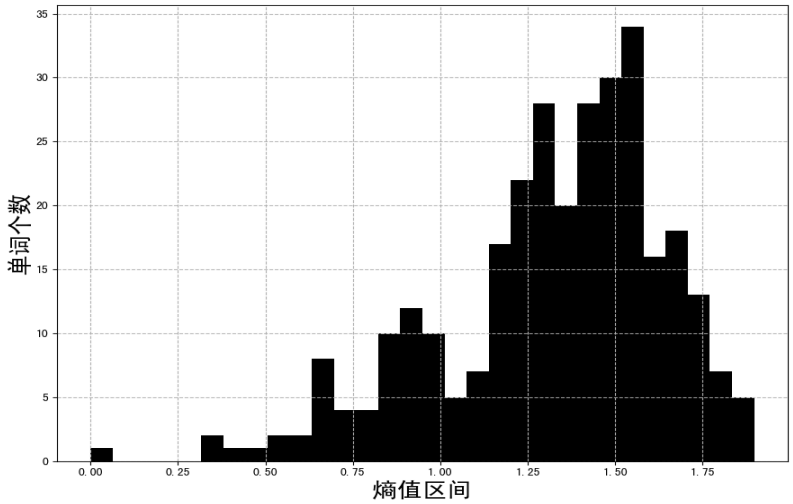

图 5-3 单词熵值直方图
Figure 5-3 Word Entropy Histogram

信息熵整体分布类似正态分布，我们将选定信息熵阈值，作为筛选合适数量词语的超参数。

5.3.4 单词模式区分度

我们选取了三个单词具体说明单词个体其具有的流量模式区分度及其作用。我们首先选取了三个单词：“校区”、“职业”、“潜川镇”，具体信息如下表 5-2 所示。

表 5-2 单词所包含的基站数与信息熵
Table 5-2 Number of Base Stations and Information Entropy Contained in Words

单词	基站数量	信息熵
校区	101	1.366251
楼塔镇	11	1.540306
建德	151	1.421847

同时我们将以上三个单词的词向量以条形图方式绘出，说明其在流量模式区分时的作用，如下图 5-4 所示。

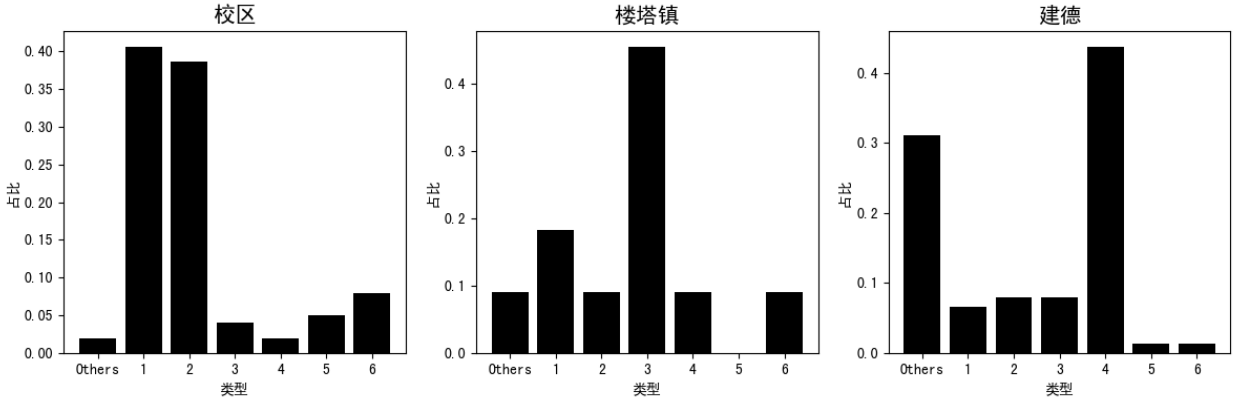

图 5-4 三个单词的词向量条形图
Figure 5-4 Three Word Word Vector Bar Chart

如图 5-4 所示，每个单词的词向量表征其包含该单词的基站归于七类流量模式中各类型的概率（包含 Others 类型），所以在单词“学校”的词向量条形图中能够看出包含“学校”的基站，例如“某市大学附属学校”的基站更偏向于类型 1，同样包含“职业”的基站更加偏向类型 1、2,；包含“潜川镇”的基站更加偏向类型 0、4。单词之间分布集中在不同类型上，使得其具有一定的模式区分度。

我们又对过滤后单词的词向量进行汇总，制作做了七个维度（表示包含“其他”类型在内的七种流量类型）的热力图，如图 5-5 所示。

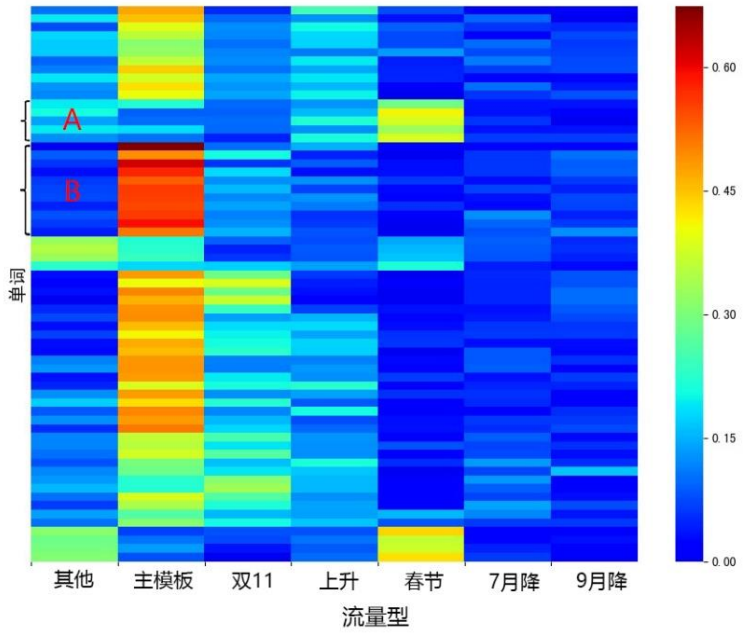

图 5-5 所有单词词向量热力图
Figure 5-5 Heat Map of All Word Words Vector

在图 5-5 中，纵轴表示不同的单词，这里由于数量太多（3005 个）并没有将具体的单词标出，但这里并没有涉及到具体单词。横轴表示 7 种流量类型，颜色越红表示越多该流量类型的基站包含该单词。例如 100 个“春节流量型”的基站都包含“郊区”一词，那么在“郊区”那一行的春节流量型就会越红，反之就越蓝。

区域 A、B 所在的行各自包含了部分单词，从图中我们可以看出在同一区域内的单词其流量型分布类似，但是不同区域的单词分布却截然不同，例如区域 A 的单词集中在春节型、区域 B 的单词主要集中在主模板型，所以这些单词就有了将模式区分的能力。最后一个基站包含了多个单词，如何将这些这些单词的分布向量组合在一起，将在下面小节详细介绍。

5.3.5 获得基站的表征

为了获得包含多个单词的基站宫格名称的词向量表征，我们对单个单词词向量进行组合。在前文我们获得了单词的向量表征，但是一个基站的宫格名称可能包含多个单词。例如某基站宫格名称为“某市某区高级中学”可以拆分成三个单词：“某市”、“某区”、“高级中学”，同时基站所处地理位置不同拆分单词个数也是不确定的。如何获得基站宫格名称的向量表征，我们采用了如下方法：

(1) 向量相加（Sum）：

$$E[\vec{e}_0,\vec{e}_1,\ldots\vec{e}_6]=\left([\vec{a}_0,\vec{a}_1\ldots,\vec{a}_6]+[\vec{b}_0,\vec{b}_1\ldots,\vec{b}_6]+\cdots+[\vec{n}_0,\vec{n}_1\ldots,\vec{n}_6]\right) \tag{5-2}$$

(2) 向量平均（Average）：

$$E[\vec{e}_0,\vec{e}_1,\ldots\vec{e}_6]=\left([\vec{a}_0,\vec{a}_1\ldots,\vec{a}_6]+[\vec{b}_0,\vec{b}_1\ldots,\vec{b}_6]+\cdots+[\vec{n}_0,\vec{n}_1\ldots,\vec{n}_6]\right)/\mathrm{N} \tag{5-3}$$

其中 N 表示基站宫格名称，再去除停用词之后所含有单词数目

5.4 预测新建基站流量模式

本小结我们利用经纬度和基站词向量表征去预测新建基站流量模式类型。我们使用 XGBoost、Logistics Regression 作为分类预测模型，同时在基站词向量表征上我们使用了两种不同处理方式。XGBoost 是采用了 Boosting 集成学习的分类预测模型，它主要通过 CART 弱学习器不断迭代学习拟合残差而提高预测准确性，是一种非线性模型；LR（逻辑回归）则是典型的线性分类模型。

5.4.1 实验操作步骤

(1) 数据处理：我们将数据分成两个部分其中训练集（70%）和测试集（30%），由于

在训练集中各类别存在明显的偏差，我们采用了重采样技术（SMOTE[43]）使得训练集中的各类别数目保持一致。为了保证对比实验的有效性，所有对比实验均基于同样的训练集和测试集。

(2) 模型构建：我们选用了 XGBoost 与 LR 两种类型不同的分类预测模型，LR 模型模型参数为：默认参数，而 XGBoost 模型参数较多，如下表 5-3 所示。

表 5-3 XGBoost 参数设置
Table 5-3 XGBoost Parameter Setting

参数	参数意义
LEARNING_RATE=0.05	学习率
N_ESTIMATORS=80	树的个数：1000 棵树
MAX_DEPTH=3	树的深度
MIN_CHILD_WEIGHT = 1	叶子节点最小权重
GAMMA=0	惩罚项叶子结点个数前的参数
SUBSAMPLE=0.8	随机选择 80%样本建立决策树
COLSAMPLE_BTREE=0.8	随机选择 80%特征建立决策树
SCALE_POS_WEIGHT=1	解决样本个数不平衡的问题

5.4.2 实验结果分析

在分类模型性能评估上我们使用了 F1-score 作为标准，一些多分类问题的机器学习竞赛，也常常将 F1-score 作为最终测评的方法。它是精确率和召回率的调和平均数，最大为 1，最小为 0。我们流量模式预测问题最终转化成七分类问题（包括了六种前文模式挖掘出六种典型流量模式，以及特征不明显的流量模式“其他”类），其性能统计如下表 5-4 所示。

表 5-4 实验结果
Table 5-4 Experimental Result

特征	XGBOOST	LR
经纬度	0.30	0.15
基站词向量表征（Sum）	0.32	0.32
基站词向量表征（AVE）	0.33	0.34
经纬度+基站词向量表征（Sum）	0.34	0.32
经纬度+基站词向量表征（AVE）	**0.35**	**0.35**

如表 5-4 所示，我们首先对单个特征的模式分类能力进行了实验，之后又尝试了特

征之间的组合，从表中我们能够获得如下信息：

(1) 我们提出的基站词向量表征在分类性能上是优于经纬度特征的（在两种模型下均高于经纬度特征），说明这种将语义信息转化成词向量的表征方法是有效的。

(2) 基站词向量特征对 LR 模型的性能提升明显。在仅仅加入基站词向量表征（AVE）特征时 LR 模型的 F1-Score 为 0.34 超过 XGBoost 模型的 0.33。这种提升其意义在于 XGBoost 虽然取得较好性能但是其计算量却远大于 LR，同时计算速度也远慢于 LR。加入基站词向量表征的 LR 模型达到相同的效果，在计算量和计算速度上占有优势。

(3) 特征组合后效果比单一特征效果好。“经纬度+基站词向量表征（Average）”的特征组合在 XGBoost 上 F1-score 为 35%比仅仅加入“经纬度”特征高出 5%，提升(0.35-0.3)/0.3=16.6%；在 LR 上的 F1-score 为 35%比仅仅加入“经纬度”特征高出 20%，提升(0.35-0.15)/0.15=133%。

我们从前文 10 组实验中选取分类性能较优的实验组，具体对分析各类型的分类能力。我们选取的实验组其特征组合为：经纬度+基站词向量表征（AVE）；使用模型为：XGBoost，具体分类情况如表 5-5 所示。

表 5-5 最优结果的详细统计表
Table 5-4 Detailed Statistical Table of Optimal Results

类别	精确率	召回率	F1-Score	类型总数
1	0.48	0.59	**0.53**	796
2	0.33	0.17	0.23	303
3	0.27	0.15	0.19	281
4	0.46	0.58	**0.52**	249
5	0.07	0.14	0.09	132
6	0.05	0.04	0.04	132
其他	0.27	0.17	0.21	240
加权平均	0.35	0.36	**0.35**	2133

在表 5-5 中，我们发现分类器能够较好的分出第 1,4 类，其数量占测试集的(796+249)/2133=49%，所代表的类型为主模板流量型和春节流量型。从图 4-4 中我们我们能够发现这两种模式在“双 11”、“春节”上的反应截然相反，运营商可以根据这两类的模式提出针对性的措施。比如应对春节流量型郊区流量的上升，需要 2 月份在城郊增添临时基站或者增加频段。预测较差的是第 5、6 类，产生的原因可能是这两种基站的数目太少，分类器没能学习到有用的信息将这两类很好的区分。

5.5 基站流量模式的应用研究

本小节着重讨论基站流量模式的多方面应用。前文得到了六种基站流量模式，我们对各类型在时间上的特殊变化结合该市特点、国内节日等做出了解释，也分析了各模式的空间分布特点。基于所有观察和测量，我们将从应用角度分析基站流量模式特点，并给出其应用场景。

5.5.1 基站流量模式互补特性

我们观察各基站流量模式波形时，发现不同模式间有波形互补现象。例如在同一月份，部分基站流量蹿升，另外部分基站流量下降，同时在次月，原本流量蹿升的基站流量迅速下跌，之前流量下降的基站又在该月份流量回升。我们设想这些基站出现在同一个子网中，那么在该子网的月度总流量不会在某月出现极高峰，将缓解基础设施带来压力，优化用户上网体验。同时，互补基站间的资源配置也将节约网络运营商的开支。

在我们获得的六类模式中有两对（四种）较为互补的模式，分别是：春节流量型与双十一流量型，7 月下滑流量型与 9 月下滑流量型，如图 5-6 所示。

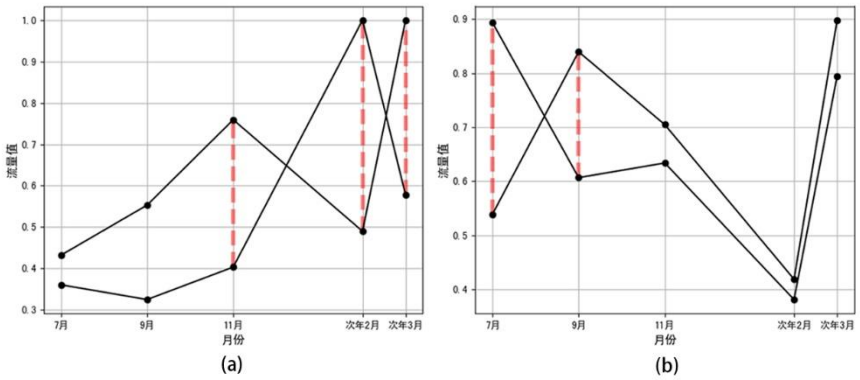

图 5-6 两对较为互补的模式
Figure 5-6 Two Pairs of More Complementary Modes

在图 5-6 中，(a)图为双十一流量型站在 11 月份、次年 2 月、次年 3 月呈现出的流量变化是“涨-跌-涨”，而同期春节流量型的流量变化时“跌-涨-跌”，所以在如上月份内流量可以互补；在(b)图中，7 月下滑流量型站在 7 月、9 月呈现出的流量变化是“跌-涨”，而同期 9 月下滑流量型的流量变化时“涨-跌”，所以在如上月份内流量可以互补。流量类型互补的意义在于：首先如果将具有类型互补的基站放在同一子网中，则该子网不会出现子网总流量在某一个月份出现极高的现象，这也是流量互补本身的特性；其次在于资源的动态调节，对基站的布局建设有重要意义，因为之间相互影响的，某个基站为了应对流量高峰把单一频段设置成多频段，为了减少干扰就需要低流量基站减少辐射频段。

5.5.2 基站流量模式其他应用

该小节我们将讨论互补基站流量模式两种互补状态，即“邻接式”互补、“非邻接式”互补。它们的区别为是否直接相邻。

对于“邻接式”配对的基站：首先在核心网的拓扑中，应当把可配对基站放入在同一个子网中，这样有助于对配对基站的频谱资源的互补利用，功耗上互补调节等。具体的表现为，基站可以在流量低谷时，将频谱资源让给可配对的高流量基站，同时降低自己的功耗，以此缩小基站辐射范围和减少对配对基站的干扰。

对于“非接触式”配对的基站，为了实现资源合理调度可以借助其他设备，例如无人机、临时基站车等，实现远距离资源调度。

5.6 本章小结

本章主要工作是，基于 7110 个真实基站数据和第四章的 Rank-Based 聚类方法的帮助下，对新建基站的流量模式进行预测。我们选取了两个典型的新建基站特征：地理位置特征（经纬度）、语义信息特征（宫格名称），分析评估其作为预测基站流量模式特征的能力，最后对基站流量模式进行预测。具体工作如下：

(1) 模式地理位置分布：我们首先可视化了六类基站流量模式在地理上的分布情况，发现春节流量型特点是主要分布郊区，其他类型密集混杂在市中心区域，分布特点：有集群现象，但单个集群规模不大，整个集群数目众多。

(2) 地址单词与流量模式关系：我们从宫格名称特征的分词入手，通过去除停用词、统计词频分布、过滤低频词到最终由单个单词组建基站的词向量表征。期间我们解释了词向量的创建方法以及引入信息熵、词向量聚类等方式来说明词向量具有一定的模式区分性。最后我们对基站词向量的向量表征给出了两种构成方式：相加、平均。

(3) 通过大量的实验和机器学习模型的调试。我们发现基站词向量表征能够明显提升 LR 线性模型的性能，同时我们将两种特征组合送入模型中，性能都在单个加入时候有所提升，“经纬度+基站词向量表征（Average）”的特征组合在 XGBoost 上 F1-score 为 35%比仅仅加入“经纬度”特征提升了 5%，提升 16.6%；在 LR 上的 F1-score 为 35%比仅仅加入“经纬度”特征提升了 20%，提升 133%。实验结果表明我们的基站词向量表征方法有显著的效果。

(4) 我们发现六类流量模式中有两对（四种）较为互补的模式，分别是：春节流量型与双十一流量型，7 月下滑流量型与 9 月下滑流量型。同时我们对互补流量模式是否处于相邻状态分别做出了其应用场景的讨论

6 总结及展望

6.1 本文工作总结

本文基于国内某运营商在某电商城市部署的超过 7000 个基站，跨度近一年的月度流量数据进行了：基站流量时空分布测量和分析、基站流量模式挖掘、新建基站的流量模式预测。我们发明了一种新的流量模式挖掘方法，它是基于月度总流量值在一年内排序值的多级聚类方法。基于真实数据，我们得到 6 种基站流量模式类型，其中最典型的流量模式占比 38.6%。另外有两种很有特色：一种具有很强的地域特点与“春节返乡”密切相关；一种具有很强的用户属性，和“双 11”电商购物密切相关。预测新建基站流量模式初始信息往往较少，我们创造性地将基站语义标签信息引入基站流量模式预测中，改进了基站流量模式预测的准确性。本文主要工作如下：

(1) 统计分析了该城市基站流量数据的时空分布。在空间上，我们发现全市范围基站流量和流量密度分布不均衡，市中心和非市中心流量差异明显；在时间上，我们总结了基站流量总体增长模式的特点是：基站流量总体呈上升趋势，各基站在关键时间节点（11 月、次年 2 月）的变化有所不同。现象促使我们进基站流量模式挖掘。

(2) 提出了一种新的基站流量演变模式的聚类方法。该方法基于基站月度总流量值在一年内的排序序列进行聚类。在我们数据集上的分析结果表明：该聚类方式对较短时间序列（不存在周期性）的涨跌特点有很好的描述，能够得到比传统聚类方法更容易理解的结果。

(3) 基于提出的聚类方法，我们对运营商的 7 千多个基站的流量演变模式进行了大规模聚类分析，获得了 6 种典型的基站流量演变模式，最主要的一种流量模式涵盖了 38.6%的基站，特点为总体流量呈上升趋势，期间 11 月份流量由高峰直转为次年 2 月的流量低谷。其它还包括：“春节返乡”、 “双 11”电商购物模式等。结合城市特点，我们对所有模式的特点与形成原因做出了解释，这些发现为运营商理解其基站的流量演变情况提供了有益的信息。

(4) 提出了一种基于地理位置和基站地址语义信息的基站模式预测方法，能够对一个新建基站的流量模式进行预测。因为应该新建基站的初始信息往往较少，因此我们创造性地将基站语义标签信息引入基站流量模式预测中，改进了基站流量模式预测的准确性。实验结果表明：基站词向量表征能够明显提升 LR 线性模型的性能，同时我们将两种特征组合送入模型中，性能都比单个特征有所提升，“经纬度+基站词向量表征（Average）”的特征组合在 XGBoost 上 F1-score 为 35%比仅仅加入“经纬度”特征高出

5%，提升(0.35-0.3)/0.3=16.6%；在 LR 上的 F1-score 为 35%比仅仅加入“经纬度”特征高出 20%，提升(0.35-0.15)/0.15=133%，其中对“主模板流量型”和“春节流量型”的预测较为准确。

6.2 未来工作展望

随着 5G 技术商业普及，越来越多新建基站需要对其自身的流量模式进行充分了解，同时也需要针对流量模式特点做出适应性变化。对基站网络流量的模式挖掘只是第一步，如何根据模式特点智能调度基站资源是亟需解决的问题。目前已经有部分工作在对实现实时基站资源智能分配问题做出研究，但是真正的实现与普及仍有一段路要走。相信随着更多移动终端设备的网络互联，基站性能更加提升，数据将会更密集的被采集来分析研究，为实现智能基站的建设做好铺垫工作。

参考文献

[1] Hofer J, Pawaskar S. Impact of the Application Layer Protocol on Energy Consumption, 4G Utilization and Performance[C]//2018 3rd Cloudification of the Internet of Things (CIoT) , 2018:1–7.

[2] Lee D, Zhou S, Zhong X, et al. Spatial modeling of the traffic density in cellular networks[J]. IEEE Wireless Communications, 2014, 21(1):80–88.

[3] Shafiq M Z, Ji L, Liu A X, et al. Characterizing and modeling internet traffic dynamics of cellular devices[J].measurement and modeling of computer systems, 2011, 39(1):305–316.

[4] Zhang K, Cuthbert L G. Traffic pattern prediction in cellular networks[J]. 2008:549–553.

[5] Tang H, Ou L. A temporal analysis of holiday effect on IP backbone traffic[J]. 2017:1042–1046.

[6] Paraskevopoulos P, Dinh T C, Dashdorj Z, et al. Identification and characterization of human behavior patterns from mobile phone data[J]. D4D Challenge session, NetMob, 2013.

[7] Paul U, Subramanian A P, Buddhikot M M, et al. Understanding traffic dynamics in cellular data networks[J]. Proceedings - IEEE INFOCOM, 2011, 8(1):882–890.

[8] Willkomm D, Machiraju S, Bolot J, et al. Primary Users in Cellular Networks: A Large-Scale Measurement Study[C]//2008 3rd IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks. IEEE, 2008:1–11.

[9] Girardin F, Vaccari A, Gerber A, et al. Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate[C]//Intl. Conference on Computers in Urban Planning and Urban Management , 2009.

[10] Zhang M, Xu F, Li Y. Mobile Traffic Data Decomposition for Understanding Human Urban Activities[C]//2016 IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems (MASS) , 2016:1–9.

[11] MacQueen J, et al. Some methods for classification and analysis of multivariate observations[C]//Proceedings of the fifth Berkeley symposium on mathematical statistics and probability , 1967，1:281–297.

[12] Yao N, Cuthbert L G. Prediction of antenna patterns for hotspots in WCDMA Networks[J]. 2006:1–6.

[13] Aamodt A, Plaza E. Case-based reasoning: Foundational issues, methodological variations, and system approaches[J]. AI communications, 1994, 7(1):39–59.

[14] Zang Y, Ni F, Feng Z, et al. Wavelet transform processing for cellular traffic prediction in machine learning networks[C]//2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP) , 2015:458–462.

[15] Elman J L. Finding structure in time[J]. Cognitive science, 1990, 14(2):179–211.

[16] Chui C K. An introduction to wavelets[M]. Elsevier, 2016.

[17] Verenich I, Dumas M, La Rosa M, et al. Complex symbolic sequence clustering and multiple classifiers for predictive process monitoring[C]//International Conference on Business Process Management , 2016:218–229.

[18] Miao D, Sun W, Qin X, et al. MSFS: multiple spatio-temporal scales traffic forecasting in mobile cellular network[C]//2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing

and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech) , 2016:787–794.
[19] Xu F, Lin Y, Huang J, et al. Big data driven mobile traffic understanding and forecasting: A time series approach[J]. IEEE transactions on services computing, 2016, 9(5):796–805.
[20] Greff K, Srivastava R K, Koutnk J, et al. LSTM: A search space odyssey[J]. ÍEEE transactions on neural networks and learning systems, 2017, 28(10):2222–2232.
[21] Jozefowicz R, Vinyals O, Schuster M, et al. Exploring the limits of language modeling[J]. arXiv preprint arXiv:1602.02410, 2016.
[22] Cleveland R B, Cleveland W S, McRae J E, et al. STL: A seasonal-trend decomposition[J]. Journal of official statistics, 1990, 6(1):3–73.
[23] Adámek J, Herrlich H, Strecker G E. Abstract and concrete categories. The joy of cats[J]. 2004.
[24] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining , 2016:785–794.
[25] Kriegel H P, Schubert E, Zimek A. The (black) art of runtime evaluation: Are we comparing algorithms or implementations?[J]. Knowledge and Information Systems, 2017, 52(2):341–378.
[26] Atary A, Bremler-Barr A. Efficient round-trip time monitoring in OpenFlow networks[C]//IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications , 2016:1–9.
[27] Roy A, Acharya T, DasBit S. Quality of service in delay tolerant networks: A survey[J]. Computer Networks, 2018, 130:121–133.
[28] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[C]//Advances in Neural Information Processing Systems , 2017:3146–3154.
[29] Varoquaux G, Raamana P R, Engemann D A, et al. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines[J]. NeuroImage, 2017, 145:166–179.
[30] Wei Y, Yang F, Wainwright M J. Early stopping for kernel boosting algorithms: A general analysis with localized complexities[C]//Advances in Neural Information Processing Systems , 2017:6065–6075.
[31] Kyurkchiev N, Markov S. Sigmoid functions: some approximation and modelling aspects[J]. LAP LAMBERT Academic Publishing, Saarbrucken, 2015.
[32] Bühlmann P, Van De Geer S. Statistics for high-dimensional data: methods, theory and applications[M]. Springer Science & Business Media, 2011.
[33] Martins A, Astudillo R. From softmax to sparsemax: A sparse model of attention and multi-label classification[C]//International Conference on Machine Learning , 2016:1614–1623.
[34] Macqueen J. Some Methods for Classification and Analysis of MultiVariate Observations[C]//Proc of Berkeley Symposium on Mathematical Statistics & Probability , 1965.
[35] Wang W, Yang J, Muntz R, et al. STING: A statistical information grid approach to spatial data mining[C]//VLDB , 1997，97:186–195.
[36] Ester M, Kriegel H, Sander J, et al. A density-based algorithm for discovering clusters in large spatial Databases with Noise[J]. 1996:226–231.
[37] Nakagawa S, Johnson P C, Schielzeth H. The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded[J]. Journal of the Royal Society Interface, 2017, 14(134):20170213.
[38] Rand W M. Objective criteria for the evaluation of clustering methods[J]. Journal of the American Statistical association, 1971, 66(336):846–850.

[39] Suri S, Verbeek K. On the most likely voronoi diagram and nearest neighbor searching[J]. International Journal of Computational Geometry & Applications, 2016, 26(03n04):151–166.
[40] Monti K L. Folded empirical distribution function curves - mountain plots[J]. The American Statistician, 1995, 49(4):342–345.
[41] Ord J K. Families of frequency distributions[J]. 1972.
[42] Jaynes E T. Information theory and statistical mechanics[J]. Physical review, 1957, 106(4):620.
[43] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16:321–357.

作者简历及攻读硕士学位期间取得的研究成果

一、作者简历

苏健，男，1994 年 6 月生。2012 年 6 月至 2016 年 7 月就读于安徽农业大学信息与计算机学院电子信息工程专业，取得工学学士学位。2017 年 9 月至 2019 年 6 月就读于北京交通大学电子与通信工程专业，研究方向是信息安全，取得工学硕士学位。攻读硕士学位期间，主要从事交通时序预测，数据分析与模式挖掘等研究工作。

二、参与科研项目

[1] 贵州省交通厅交通流量预测

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和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京交通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了谢意。

学位论文作者签名：  签字日期：2019 年 5月 31日

学位论文数据集

表 1.1： 数据集页

关键词*	密级*	中图分类号	UDC	论文资助
网络测量，模式提取、预测、机器学习	公开			
学位授予单位名称*		学位授予单位代码*	学位类别*	学位级别*
北京交通大学		10004	工程硕士专业学位	硕士
论文题名*		并列题名		论文语种*
大规模基站网络流量模式挖掘和预测				中文
作者姓名*	苏健		学号*	17125052
培养单位名称*		培养单位代码*	培养单位地址	邮编
北京交通大学		10004	北京市海淀区西直门外上园村 3 号	100044
学科专业*		研究方向*	学制*	学位授予年*
电子与通信工程		信息网络	2	2019
论文提交日期*	2019.06.03			
导师姓名*	陈一帅		职称*	副教授
评阅人	答辩委员会主席*		答辩委员会成员	
	郭宇春			
电子版论文提交格式 文本( ) 图像( ) 视频( ) 音频( ) 多媒体( ) 其他( ) 推荐格式：application/msword；application/pdf				
电子版论文出版(发布)者		电子版论文出版(发布)地		权限声明
论文总页数*	47 页			
共 33 项，其中带*为必填数据，为 21 项。				