学校代码：10286

分类号：

密　级：

U D C：

学　号：

工程硕士学位论文

基于机器学习的无线网络用户业务行为分析

研究生姓名：**李玉**

导 师 姓 名：**潘志文 教授**

申请学位类别 工程硕士　　学位授予单位 **东 南 大 学**

一级学科名称 电子与通信工程　　论文答辩日期 **20　年　月　日**

二级学科名称 通信与信息系统　　学位授予日期 **20　年　月　日**

答辩委员会主席　　评　阅　人

20　年　月　日

东南大学

工程硕士学位论文

基于机器学习的无线网络
用户业务行为分析

专 业 名 称： 电子与通信工程

研究生姓名： 李玉

导 师 姓 名： 潘志文 教授

ANALYSIS OF USER TRAFFIC BEHAVIOR IN WIRELESS NETWORKS BASED ON MACHINE LEARNING

A Thesis Submitted to

Southeast University

For the Academic Degree of Master of Engineering

BY

LI Yu

Supervised by

Prof. PAN Zhiwen

School of Information Science and Engineering

Southeast University

May, 2019

东南大学学位论文独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其它人已经发表或撰写过的研究成果，也不包含为获得东南大学或其它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。

研究生签名：_____日期：_____

东南大学学位论文使用授权声明

东南大学、中国科学技术信息研究所、国家图书馆、《中国学术期刊（光盘版）》电子杂志社有限公司、万方数据电子出版社、北京万方数据股份有限公司有权保留本人所送交学位论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括以电子信息形式刊登）论文的全部内容或中、英文摘要等部分内容。论文的公布（包括以电子信息形式刊登）授权东南大学研究生院办理。

研究生签名：_____导师签名：_____日期：_____

摘要

随着移动通信网络的飞速发展，移动设备数目激增，业务种类繁多，用户的服务质量需求也随之提高。对用户网络业务行为的精准刻画有助于为用户提供定制化的服务，同时为网络优化也提供了有用信息。但移动用户的业务流量数据随时间、空间波动严重，且不同用户个体差异性大，使得传统的时间序列预测算法无法直接适用于个人用户的网络业务流量预测。本文基于机器学习方法，针对用户网络业务流量预测问题进行了分析与研究，具体研究内容包括以下四个方面。

用户流量时域分布规律研究以及用户行为分析。对原始数据进行预处理及聚合求和后，得到了时域的统计分布规律，研究了整个网络的周期性及自相关性特点。然后分别分析研究了用户流量使用特征、用户访问手机应用软件特征以及用户移动性。分析发现用户使用流量行为存在极为明显的长尾效应，同时发现用户移动性对用户业务流量数据及用户应用使用行为有很大的影响。高移动性用户倾向于使用更多流量，并访问更多手机应用软件。

用户业务流量数据聚类。应用因子分析法对用户业务流量时间序列进行特征提取并降维，提取出五个分别代表五个特定时间段的公共因子。应用 k 均值聚类算法对因子分析法降维后的特征向量进行了聚类分析，根据用户使用流量时域特征，挖掘用户之间的相似性，将用户分成六种类型，为运营商提供一种更加合理的专属计费方式，即根据不同类型的用户使用流量的时间分布制定策略。基于因子分析法的 k 均值聚类算法对用户聚类效果更好且特征可解释性更强。

基于用户手机应用软件行为聚类，挖掘用户应用行为与用户业务流量数据之间的相关性。对用户使用手机应用软件的行为进行了重点分析，应用潜在语义分析提取特征并降维，然后用 k 均值对用户进行聚类，将用户分为六种不同类型。分析不同类型用户使用手机应用软件行为与其对应的使用流量的行为，结果发现，用户使用流量的行为与其使用手机应用软件行为紧密联系。因此手机应用软件使用行为对预测用户使用流量很有帮助。

用户业务流量预测算法研究。从单个用户角度出发，只使用用户本身的时间序列流量数据进行预测。针对个人用户业务流量非平稳、数据突变性强等特点，提出一种基于小波变换的 Prophet 与高斯过程回归预测算法。小波变换可将原始用户业务流量时间序列分解为高低频子序列，再根据子序列的特点，分别应用高斯过程回归算法与 Prophet 算法进行预测。使用平均绝对百分误差作为评估指标，对比传统的时间序列预测算法如自回归积分滑动平均模型，提出的方案能将预测误差降低一半。

关键词：移动通信网，用户业务流量，聚类算法，时间序列预测

Abstract

With the rapid development of mobile communication networks, the number of mobile devices has surged, a wider variety of services appears, and the user's service quality is increasingly improving. The accurate portrayal of the user's network traffic behavior can help to improve the customized service for the user and provide useful information for the network optimization. Individual user tends to have large divergence in their behaviors. Due to more serious fluctuation in user's network traffic, existing time-series prediction algorithms cannot be directly applied to the network traffic prediction of individual user. In this thesis, the user network traffic prediction problem is analyzed and studied. The main contributions of the thesis include:

User traffic time domain distribution and user behavior analysis is performed. After preprocessing and aggregating the raw data, the user traffic data time series is obtained, and the periodic and auto-correlation characteristics of the overall network traffic are analyzed. Then for individual user, the characteristics of user traffic usage, the mobile phone application usage and user mobility pattern are analyzed. Results show that there exists a significant long tail effect in the user's traffic behavior, and the user mobility has a great impact on the user's data traffic and the user's application usage behavior. High-mobility mobile users tend to use more traffic and access more mobile apps.

Clustering of user network traffic data is conducted. The explanatory factor analysis method is used to extract the feature of the user's traffic data time series and reduce the dimensionality. Then, five common factors are extracted, which represent five specific time periods respectively. The K-means clustering algorithm is used to cluster the users. And based on the time domain characteristics of the user traffic data, the similarity between users is discovered. The users are finally divided into six types. This provides the possibility of user-specific charging policy based on the prediction result of user data traffic. The K-means clustering based on factor analysis method has better performance and more interpretable features.

Based on the clustering of user mobile application usage, the correlation between user application behavior and user's traffic data is analyzed. Latent semantic analysis is applied to extract the feature and reduce the dimensionality, then K-means clustering is adopted to cluster the user into six different types. Based on analyzing of the traffic data of different types of users, it is found that the behavior of users using traffic data is closely related to their application usage behavior. Therefore, the mobile application usage is helpful for predicting the user traffic.

User traffic prediction algorithm is investigated. A user network traffic prediction model based on Prophet algorithm and Gaussian process regression model is proposed. Wavelet

transform is first employed to decompose the original user traffic data time series into low frequency component and high frequency component. The low frequency component bears the long range dependence of user network traffic, while the high frequency component reveals the gusty and irregular fluctuations of user network traffic. Then Prophet algorithm and Gaussian process regression algorithm are used to predict the low frequency component and high frequency component respectively. Using the mean absolute percentage error as the evaluation metric, compared with the existing time series prediction algorithm, the proposed scheme can lower the prediction error by half.

Keywords: mobile communication network, user data traffic, clustering algorithms, time series prediction

目录

插图目录

表目录

缩略语表

英文缩写	英文全称	中文全称
AR	Auto-Regressive	自回归模型
ARMA	Auto-Regressive Moving Average	自回归滑动平均模型
ARIMA	Auto-Regressive Integrated Moving Average	自回归积分滑动平均模型
BIC	Bayesian Information Criterion	贝叶斯信息量
CDF	Cumulative Distribution Function	累积分布函数
CDR	Call Detail Record	通话行为数据
DBI	Davies-Bouldin Index	戴维森堡丁指数
DBSCAN	Density-Based Spatial Clustering of Applications with Noise,	基于密度的聚类算法
DWT	Discrete Wavelet Transform	离散小波变换
EFA	Explanatory Factor Analysis	探索性因子分析法
GPS	Global Positioning System	全球定位系统
IDWT	Inverse Discrete Wavelet Transform	离散小波逆变换
IMEI	International Mobile Equipment Identity	国际移动设备识别码
KMO	Kaiser-Meyer-Olkin	KMO 统计量
KPI	Key Performance Indicator	关键业务指标
LDA	Linear Discriminant Analysis	线性判别分析
LRD	Long Range Dependence	长程依赖性
LSA	Latent Semantic Analysis	潜在语义分析
LSTM	Long-Short Term Memory	长短期记忆网络
MA	Moving Average	滑动平均
MAE	Mean Absolute Error	平均绝对误差
MAPE	Mean Absolute Percentage Error	平均绝对百分误差
MB	Megabyte	兆字节
MSE	Mean Squared Error	均方误差
PCA	Principal Component Analysis	主成分分析法
POI	Point of Interest	地图兴趣点
QoS	Quality of Service	服务质量
RMSE	Root Mean Squared Error	均方根误差
RNN	Recurrent Neural Network	递归神经网络
STL	Seasonal-Trend decomposition using Loess	基于鲁棒局部加权回归平滑法的时间序列分解方法
SVD	Singular Value Decomposition	奇异值分解
TF-IDF	Term Frequency-Inverse Document Frequency	词频逆文档频率
WLAN	Wireless Local Area Network	无线局域网

第一章 绪论

1.1 研究背景

随着移动通信技术的不断发展，移动用户数目激增，业务种类日益繁多。移动互联网逐渐对人们生活的诸多领域，例如社交、支付、餐饮、出行等产生重大影响。一方面，移动通信网络必须承载和满足的业务传输需求呈现爆发式的增长，这对移动通信网络和网络运营商来说都是极大的挑战。如何提供更快速更稳定更高效的服务，保障用户的体验需求，是未来移动通信技术发展所要面临的一大问题。另一方面，大量移动通信网络数据蕴含重要的网络信息及用户信息，如果加以合理利用，运营商可有针对性地提出基于相关认识的协议设计、频谱分配、计费策略巧节能方案，从而提供更优质的服务、减少运营成本开销、提高利润收入。因此，如何合理有效地挖掘移动通信数据的应用价值成为科研学者的关注重点。

面对移动通信网络未来发展中所面临的挑战，一些科研学者开始尝试寻求在对移动通信网络数据进行分析与挖掘中找到相应的解决方案，如对移动通信网络和用户的特性进行深入分折并如何将机器学习与移动通信网络相结合目前已经成为一个重要的研究方向。机器学习和人工智能算法的引入，有助于实现无线网络的自优化，提高效率，提供最优质的用户体验。个人用户的流量预测对网络资源优化分配及精细化市场计划具有很重要的意义。对未来流量的准确预测有助于提高基于需求的资源分配效率。然而准确预测用户流量是一件非常困难的事情。首先，移动用户在不同的场合不同的时刻需求各异，使得预测流量变得尤为困难。其次，用户之间使用流量行为差异性较大。最后，网络流量同时被诸多外部因素影响，如重大活动，天气等等。因此应用机器学习算法进行移动通信网络数据分析与挖掘的工作具有非常广阔的发展前景与研究意义。

1.2 国内外研究现状

近年来，运营商通过基础设施网络收集到大量移动数据。这些数据往往包含数以万计甚至百万计个用户，覆盖整座城市甚至地区和国家，在时间跨度上达到数周至数月。因此这些蕴含大量宝贵信息的海量数据，逐渐吸引到国内外众多学者，甚至跨学科的学者们的广泛关注。国内外基于运营商数据进行的研究主要在以下几个方面。

用户移动性。移动通信数据能够提供百万用户的移动特征，绝佳地反映了用户的移动轨迹。更重要的是，相比传统的途径，例如通过调查问卷，运营商获得的移动通信数据可以覆盖更多的用户，而且几乎没有额外的运营成本。因此移动通信数据开始被应用在诸多领域来建模人口移动性模型，包括电信网络优化，城市发展规划，智慧交通，人类社会学，疾病传播学等等。通过预测用户未来出现的空间位置，政府部门可以更好地部署规划交通设施和资源调度，从而更好地缓解交通拥堵。例如优步和嘀嘀这样的行程分享平台，凭借精准的用户移动性预测模型估计用户需求，并根据需求进行相应的资源

分配。

文献[1]研究了一个含 4156 个用户的数据集，根据用户访问的基站数目来表征他们的移动性。发现移动性通常比较低，55%的用户只出现在一个位置。然而，移动性的统计分布存在很严重的长尾效应，存在用户一周内访问数百个基站。这种用户移动性的分布不均现象后来也在被文献[2]在一个更大的数据集上验证。该根据运营商在全国 3G 蜂窝网络中收集到的数据，研究网络资源使用和用户行为分析。数据包含百万用户和千个基站，研究用户移动性与时域活动与产生负载的关系，发现 60%的用户是静止的，但是 1%的用户平均一天访问 50 个基站以上，且用户移动轨迹在以天为单位的时间尺度上呈现一定的周期性。

对用户移动性的研究主要基于通话数据（Call Detail Record, CDR）或者无线局域网（Wireless Local Area Network, WLAN）数据。前者仅能捕捉通话过程中用户的移动，而后者与移动通信网络相比，覆盖范围规模又较小。文献[3]证实了根据移动通信数据来评估人口分布的可靠性，结合用户 CDR 数据和土地使用信息，为意大利米兰建立了高精度的人口分布估计模型，在时间和空间上提供实时的人口估计，而无需借助传统的昂贵耗时的人口普查。文献[4]建立一个三步模型，从大量用户移动性数据中提取出常见的时域模式，并发现了用户时域移动模式与其用户行为（例如对手机应用的访问）以及社会经济地位等有一定的相关性。文献[5]研究了 WLAN 网络中个人用户行为特征。文献[6]研究基于移动流量数据的用户移动性分析，识别出三种时域用户移动模式，并且针对这几种用户移动模式，提出一种基于给定位置的手机应用软件使用预测方案。文献[7]应用大数据平台 Hadoop 对大量 4G 网络数据进行相近的分析，通过对比 CDR 数据与 3G 数据对用户移动轨迹的建模，发现 4G 网络流量数据可以提供粒度更细的用户位置与移动信息。文献[8]基于真实 4G 数据，应用基于密度的聚类算法（Density-Based Spatial Clustering of Applications with Noise, DBSCAN），建立用户移动性模型。文献[9]提出一种基于递归神经网络的用户移动轨迹预测算法。

随着智能终端的普及，越来越多的用户不再通过传统的网页浏览方式访问网络，而是借助各种各样的手机应用。但是我们对人们何时何地使用何种软件仍然知之甚少。

文献[10]应用美国一大运营商提供的应用软件使用数据，基于 HTTP 签名对应用市场的各类应用软件进行识别与分类，并深入研究了各类应用在时域与空间与上的流行性、区域性和相关性，发现有些应用软件的在空间覆盖和每日使用模式上有相似性，比如某些应用很可能被一起使用。同时，不同类型的应用软件在时间上有很强的差异性，比如新闻类软件更多地在早上被使用。这些有意义的发现为网络运营商和内容提供商提供了指导作用。文献[11][12]应用数据挖掘技术，从终端角度预测用户即将使用各个应用软件的概率。文献[13]简述了应用软件预测的特征工程与特征选择。文献[14]学习并预测手机应用软件级或服务级流量，主要研究三类应用：即时通信，网页浏览和视频，展示了 alpha-稳定模型建模手机应用软件流量使用的准确度，然后结合字典学习模型得到更精确的预测结果。

文献[15]研究用户移动性与用户使用手机应用软件行为之间的相关性，包括使用流

量大小，访问频率以及应用软件类型等。文献[16]研究了 5342 名用户使用移动应用软件的行为，挖掘不同类型用户访问应用软件的模式，结合用户所在位置的地图兴趣点（Points of Interest, POI）信息提出一种应用软件使用预测算法，来预测用户下一个使用的应用软件。文献[17]研究中国某运营商大规模匿名通信网络数据，对手机移动应用软件使用模式特征进行了描述，并展示移动性，地理特性和用户行为如何影响他们的手机应用软件使用，观察到用户的平均流量消费随移动性增加而增加，即高移动性用户倾向于产生更多的流量，还发现移动性对不同的手机应用软件影响不一，例如，网页浏览类软件随着用户移动性提高而增长。然而，游戏社交类在移动范围较小的用户中使用更频繁。这些发现为根据用户移动性来估计其对手机应用软件使用情况提供了可能性。

文献[18][19][20][21]研究应用软件使用行为预测。其中[18][19]从运营商角度出发，预测指定基站范围内即将被使用的应用软件。文献[18]应用迁移学习理论，结合特定位置附近的 POI 信息，成功预测最可能被使用的应用软件及整体的统计分布。文献[19]首先分析了不同类型应用软件的时域特征，然后根据基站位置附近的地图兴趣点对基站进行聚类分析，并预测不同聚类下基站在一段时间内最流行的几种应用软件类型。这有利于网络运营商了解基站不同应用软件的使用分布，并制定相应的边缘缓存机制。[20][21]则从用户角度，预测用户即将使用的手机应用软件类型。

研究与预测蜂窝网络负载对运营商网络部署、拥塞控制、负载均衡以及费用设计机制可提出指导建议。

文献[22][23][24]对同一组数据进行了长期仔细深入的研究。这组数据来自于 2014 年 8 月中国一大运营商在上海的蜂窝网数据，涵盖 15 万名用户和 9600 个基站。文献[22]把基站负载当作时序信号处理，将其分解为周期分量与随机分量，然后用时间序列预测模型预测规律成分，还证明随机分量的不可预测性。文献[23]从 9600 个基站提取出 5 种不同类型的负载时域模型，还发现每一种类型可以对应不同类型的空间位置，包括居民区、商业区、交通中心、娱乐中心和混合区域。文献[24]利用频谱分析方法，对负载时序信号进行离散傅里叶分析，提取出三个主要频率成分。然后根据提取的频域特征，对基站负载进行聚类分析，得到 5 种不同类型，分别对应 5 种不同的空间位置。与文献[23]相比，基于频谱特征的聚类效果比基于时域特征的效果更好，且算法运行更快。

文献[25]总结了目前对移动通信负载的研究，将其分为主要两类，一类是从个人用户角度出发，研究用户的行为，包括移动性、产生的流量消耗，以及使用的服务类型等。另一类从运营商角度出发，研究基站或扇区内所有用户的聚合负载。

大多数流量预测都是基于历史数据。文献[26][27]提出结合历史流量使用数据与网络其它 KPIs 的利用率。文献[26]先基于互信息用贝叶斯网络判定特征与下一时刻（时间尺度为一小时）流量之间的关系。这一步是为了选出最重要的特征，而去除不相关的或冗余的特征，使模型更简单高效。贝叶斯网络包含 10 个端点，其中一个点为目标值，即下一小时的流量，其它 9 个点代表特征，如当前时刻的流量，前一天同时刻的流量，前一时刻的流量，切换尝试次数，掉话率等。经过特征选择后，再分别用自回归模型，神经网络，高斯过程等机器学习算法来预测流量，并与单纯用时间序列模型来预测的结果

进行比较。结果显示，同样的机器学习算法，结合 KPIs(Key Performance Indicator, KPI)等特征的模型预测准确度更高，均方根误差更小。预测结果进一步应用于拥塞控制。

文献[27]则先对基站进行聚类，将基站划分为 40 个不同类别，提取出每一类的基站的流量使用行为特征。然后对每一类基站提出一个预测模型（时间尺度也是一小时）。目前，预测大量基站的流量仍然是一个挑战，因为对每一个基站建立一个独立的模型是不可行的，运算量过大。而对所有基站用同一个模型会影响预测结果的准确度。而前面聚类的结果显示，同一聚类下的不同基站具有相似的行为。因此，对同一聚类下的所有基站运用一个模型是可行且高效的。最后应用预测结果到拥塞控制机制中，提出预警方案。以及指出可将预测结果应用到节能减耗的方案中。文献[28]应用高斯过程模型，对移动通信网络流量建立贝叶斯空时模型，并预测不同时刻下流量的空间分布。文献[29]提出可用韦布尔分布(Weibull Distribution)来建模无线移动网络负载的空间密度分布。

随着深度学习的发展，神经网络模型渐渐被用于预测流量。文献[30]研究一个 LTE 基站的负载情况。将时间序列预测问题转化为有监督学习，应用循环神经网络预测 LTE 基站负载。文献[31]基于了中国移动的实际网络数据，应用栈式自编码算法模型（Stacked Auto-Encoder）建模空间关系，用 LSTM 进行时域建模，比较了自回归积分滑动平均模型（Auto-Regressive Integrated Moving Average, ARIMA）模型，发现结合 LSTM 和自编码算法模型能取得更好的结果。

近些年来，深度学习领域的快速发展，使得神经网络的应用领域日益广泛。其中长短期记忆网络（Long Short-Term Memory LSTM）尤为适用于时间序列预测问题。LSTM 是一种特殊的循环神经网路（Recurrent Neural Network, RNN），主要是为了解决 RNN 在长序列训练过程中易出现的梯度消失和梯度爆炸问题。文献[32]应用多种基于 RNN 的预测算法，挖掘基站在空间域上的相关性，提高负载预测经度。与普通的 RNN 相比，LSTM 能够学习长期依赖信息，在长序列预测中有更好的表现。文献[33]提出了基于深度神经网络的多任务学习系统，研究了循环神经网络，三维卷积神经网络以及结合卷积神经网络与循环神经网络三种深度学习模型。卷积神经网路可以很好地提取地理空间信息特征，而循环神经网络可以有效提取出时域特征，结合卷积神经网络与循环神经网络，可以实现 70%-80%的预测准确度。文献[34]用深度学习模型 LSTM 结合图形卷积算法，对基站流量进行预测，同时提取到时域与空间域特征。文献[35]通过小波变换得到时域特征，再应用深度学习算法预测流量使用。

基于流量预测的基站节能机制。文献[36]应用机器学习算法预测基站负载，并提出一种基于预测结果的基站休眠算法。从多角度分析基站负载：时域因素，空间域因素，特殊事件的影响，并提出一种多角度集体学习算法来预测基站负载。和当前的机器学习算法相比，该算法在性能上获得了 40%的提升。同时，基于此预测算法提出的基站休眠策略也使基站降低了 10%的能耗。文献[37]提出一种基于负载预测的节能算法。基站在谷时会产生很大的能耗浪费。提前精准预测基站负载，在用户需求较低时关闭一部分基站，就能实现节能减排的效果。实验表明在一天的某些时段可能实现关闭 49%的基站而不降低服务质量（Quality of Service, QoS）。

综上所述，目前国内外学者对基站的负载，用户移动性研究较多。从个人用户角度来研究手机应用的使用情况，对网络运营商和内容提供商都非常必要。如果用户画像及单用户流量需求可被识别及预测，那么运营商就能够获取更多信息来优化网络，提供定制化服务，同时，内容提供商也可以更好的定向目标用户，定制更加精细的市场策略。之前的研究主要在于描述性的统计研究，或者预测即将被使用的手机应用软件。然而，在个人用户应用流量需求方面，研究仍然甚少。因此，本文从单个用户本身出发研究更细粒度的不同用户类型的流量使用情况，并提出一种用户流量预测算法。

1.3 论文研究内容

目前，对移动通信网络负载的研究主要集中在基站端的聚合数据，而对用户终端的流量使用情况少有涉及。本文主要从用户角度研究用户访问手机应用及使用流量情况，并预测用户即将使用的流量大小。

首先从全网用户角度和个人用户角度分析了用户行为时域分布规律，包括用户使用流量及用户使用手机应用软件行为。研究了用户移动性对用户使用流量及用户使用手机应用软件的影响。通过对用户行为时域分布规律的研究，发现用户行为存在极大的差异性。长尾效应在用户使用流量行为极为明显，大部分用户产生极少流量，少量用户产生了大部分流量。

为了研究不同用户的特点，对用户进行聚类分析以得到不同的类别。由于用户使用流量时域特征数目过大，存在一定的冗余性且影响分析，因此选用因子分析法对特征矩阵进行降维处理。然后应用聚类算法可将用户聚类到不同簇中，观察簇内用户平均使用流量情况。进一步研究用户使用应用软件与用户使用流量之间的相关性。将近万种手机应用软件按照功能进行分类，基于用户使用应用软件的行为对其进行聚类，发现聚类结果与基于用户使用流量聚类结果相似，反映出用户使用手机应用软件行为与用户使用流量行为之间的强相关性。

提出一种基于小波分解的 Prophet 与高斯过程回归的网络流量预测算法。由于个人生活的随机性与突发性，用户使用流量行为通常是不连续的、间断的，甚至在部分时刻，用户并不产生流量数据。因此，很难捕捉流量使用的动态细节，并且去预测单个用户的流量使用。考虑到上述问题，先对用户使用流量时间序列应用离散小波变换，将时间序列分解为高频子序列与低频子序列。高频子序列表征原始时间序列的随机性与突发性，而低频子序列则表征时间序列的周期性与长程依赖性。针对两个子序列的特点，分别应用高斯过程回归模型和 Prophet 模型来预测高频子序列与低频子序列。最后通过离散小波逆变换得到最终的预测结果。这是一种有效结合时间序列与机器学习的预测算法，可以有效捕捉个人用户产生流量的突发性特征，实现更好的预测效果。

1.4 论文组织结构

本文共分为六章，各章的主要内容具体如下：

第一章为绪论，介绍移动通信网络数据挖掘的研究背景和意义。总结归纳了国内外近年来在相关方面的技术研究方向、趋势，最后介绍了论文的研究内容与结构安排。

第二章对使用的数据集进行介绍，对原始数据进行预处理，按时间和用户进行聚合。从全网角度和个人用户角度统计分析了用户使用手机应用软件产生分布规律，以及用户的移动性等。

第三章基于用户使用流量行为，对用户网络业务流量时间序列数据进行特征提取与降维处理。然后基于降维后的特征对用户进行聚类分析，将用户分为六类，分析不同类型用户的特点。

第四章基于用户使用手机应用软件的行为对其聚类，发现用户应用软件行为与用户使用流量行为之间强烈的相关性。

第五章提出一种基于小波分解的 Prophet 与高斯过程回归的无线网络用户流量预测算法，分析其预测性能。

第六章为对全文的总结和展望。对研究内容进行回顾，对未来移动通信机器学习进行展望。

第二章 用户流量使用规律研究

2.1 引言

近年来，移动通信网络用户行为分析成为诸多学者研究的课题。文献[38]对收集到的中国某大城市 2G 及 3G 网络的移动流量数据从以下三个方面进行了详尽的分析：数据使用、用户移动模式、手机应用软件使用。流量重度用户及高移动性用户倾向于同时占用大量数据及无线资源，且数据的使用及用户移动性与用户使用手机应用软件行为紧密相关。文献[39]对 3G 蜂窝网络的移动流量数据进行了分析，观察到其中严重的不均匀现象，大部分用户偶尔使用流量，小部分重度流量用户使用了大部分流量。文献重点研究了重度流量用户，发现其中大部分流量消耗由少数应用软件产生，例如在线视频播放软件，社交网站等。

本章主要根据现有数据研究用户网络业务流量的规律与特性。本章第二小节对数据进行一个简要概述，并对数据进行必要的预处理。第三小节从整体移动通信网的角度出发，研究全网用户业务流量的时域分布规律及相关统计数据。第四小节从个人用户角度出发，对用户使用流量，访问手机应用软件，及用户移动性三个方面来观察用户行为时域统计特性。通过对累积分布函数曲线的观察证实整体网络和用户个体流量分布的不均匀现象，同时也发现用户之间流量使用行为差异性较大。并且分析用户移动性与用户使用流量及访问应用软件之间的相关性。

2.2 数据概述及数据预处理

所使用数据来自于中国某大运营商在上海的移动通信网络，数据一共记录了近七万名用户，在为期两个月内访问超过万种手机应用软件并产生流量消耗的行为。数据主要记录了用户使用各种手机应用软件时产生的日志记录。每一条记录包含了用户的匿名国际移动设备识别码（International Mobile Equipment Identity, IMEI），用于在移动电话网络中识别每一部独立的手机等移动通信设备，使用日期，使用应用软件名称，使用时长，使用流量，全球定位系统（Global Positioning System, GPS）经纬度，设备注册省市信息，终端设备型号等。在进行数据分析前，对数据进行必要的预处理，保证数据的合理性与可用性，主要包括以下步骤。

数据聚合。原始数据记录了用户每次使用手机应用软件的时刻、时长以及产生的流量消耗。为了便于分析，以一小时为时间粒度，对每个用户在一小时内产生的流量消耗进行聚合求和。以十月数据为例，每个用户均有 24 小时/天*31 天 = 744 个时隙，得到离散的用户业务流量数据时间序列。

异常用户剔除。首先观察数据可以发现数据的稀疏性。仅有 0.0134%的用户每天都有记录。此外，访问手机应用数据记录比产生流量数据记录多，说明大量访问手机应用行为并未产生数据。观察用户使用流量时间序列，可以发现，存在少部分用户几乎不产

生流量，这些用户对后续分析讨论无法提供有用信息，因此可以剔除这些用户。对每个用户，统计其产生流量的时隙数目，即用户流量时间序列中非零数值的个数。九月数据显示，68650 名用户中有 20130 名用户完全没有产生任何流量消耗。因此在预测流量时不考虑这些用户。

尺度压缩。用户使用流量以字节为单位，通常取值跨度较大，从 0 到数十万。因此，对非零流量数值取以十为底的对数，进行尺度上的压缩。

2.3 全网用户流量使用时域分布统计

下面先从全网用户角度，对用户使用总流量进行分析。以 1 小时为时间尺度聚合求和流量数据。图 2-1、图 2-2 分别为九月、十月的全网使用流量数据。首先可以观察得出全网流量在以天为单位的时间尺度呈现出极为显著的周期性。

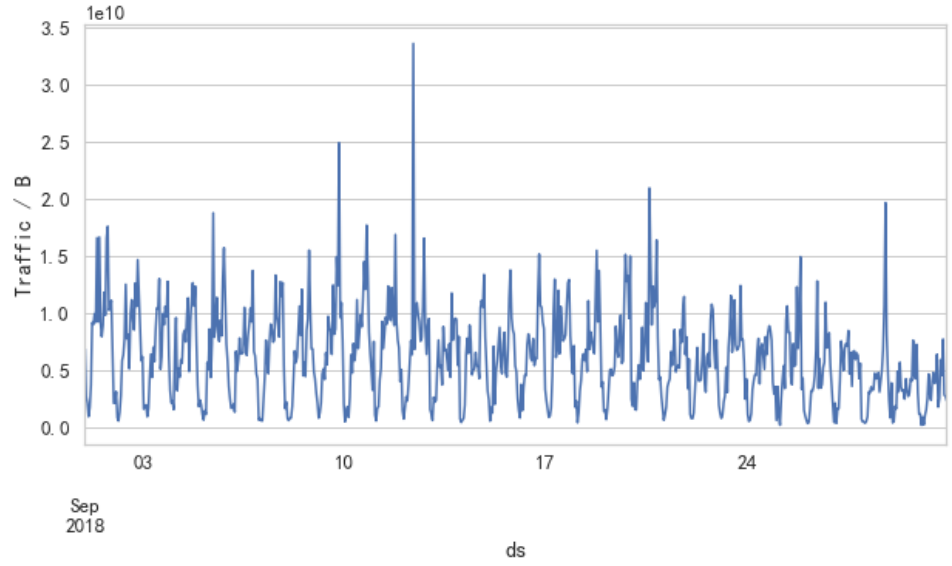


图 2-1 2018 年 9 月全网用户产生流量时域统计

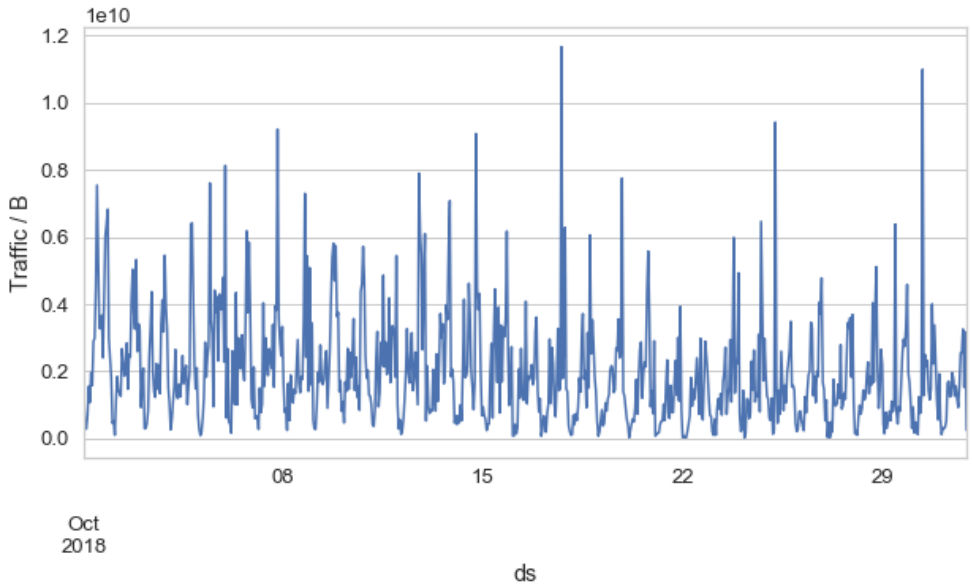


图 2-2 2018 年 10 月全网用户产生流量时域统计

对 9 月和 10 月流量时序信号进行基于鲁棒局部加权回归平滑法的时间序列分解(Seasonal-Trend decomposition procedure using Loess, STL）分解[40]，将时间序列分解为三个主要分项：趋势分量，周期分量及余项。可用公式表示为：

$$y(t)=g(t)+s(t)+r(t) \tag{2.1}$$

其中 $y(t)$ 为全网用户流量数据， $g(t)$ 是趋势项，表示时间序列值非周期性的变化，$s(t)$ 代表周期性变变化，$r(t)$ 代表模型无法捕捉的特殊变化，假设其服从正态分布。考虑以天为时间单位的周期性，分别对 9 月和 10 月全网用户使用流量时间序列进行 STL 分解，得到各分量如下图 2-3、图 2-4 所示。

观察九月与十月的全网流量数据，可以看出九月与十月全网流量整体特征相近，几乎一致。从趋势分量来看，全网流量使用整体上呈现非常显著的下降趋势，可能是由于人们渐渐倾向于使用 WLAN 或者 4G 网络。另外考虑到九月是开学季，可能流量使用较平时多。注意观察十月前四天呈现明显的下降趋势，考虑到十月前一周为中国法定节假日十月黄金周，人们可能选择离开大城市出行旅游或返乡等，因此全网用户产生流量急剧减小。

从周期分量看，全网用户使用流量呈现以天为时间尺度的周期性。观察一天的全网流量使用，可以看出呈现双尖峰特征，一早一晚，其中晚间的高峰值更大。人们普遍白天上班，夜晚休息时间使用流量更多，到午夜凌晨又逐渐减小趋近于 0，这符合大部分人的生活工作习惯。

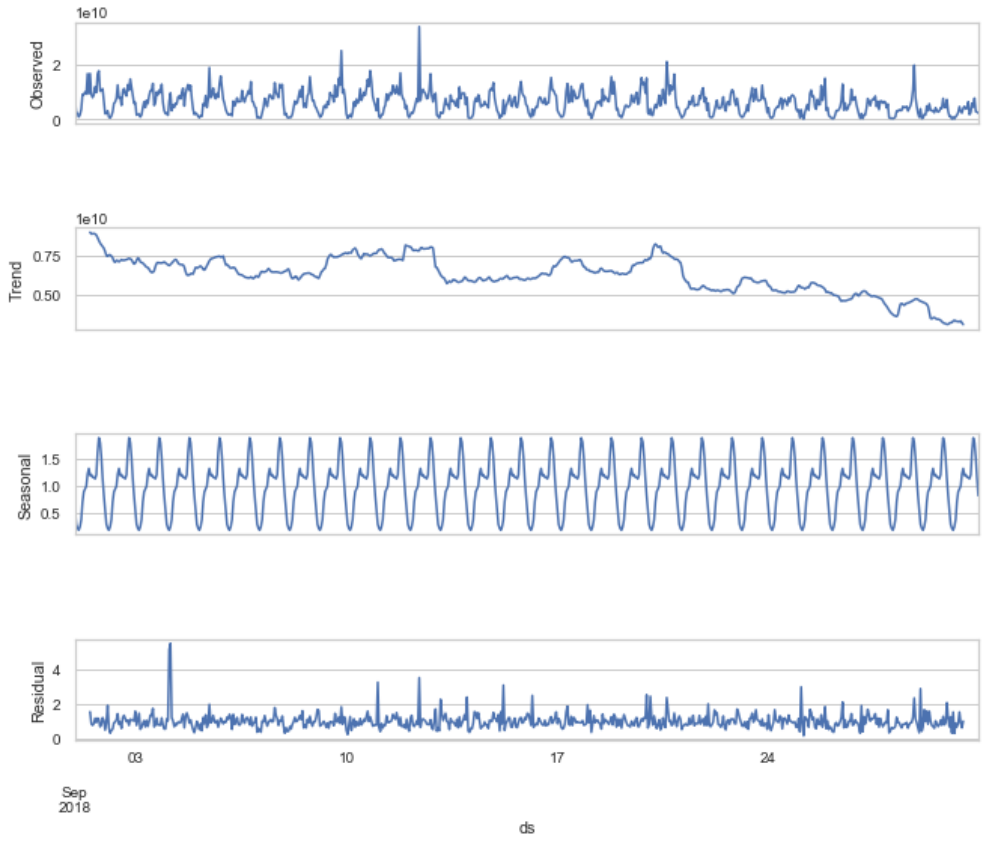


图 2-3 2018 年 9 月全网用户产生流量时域 STL 分解

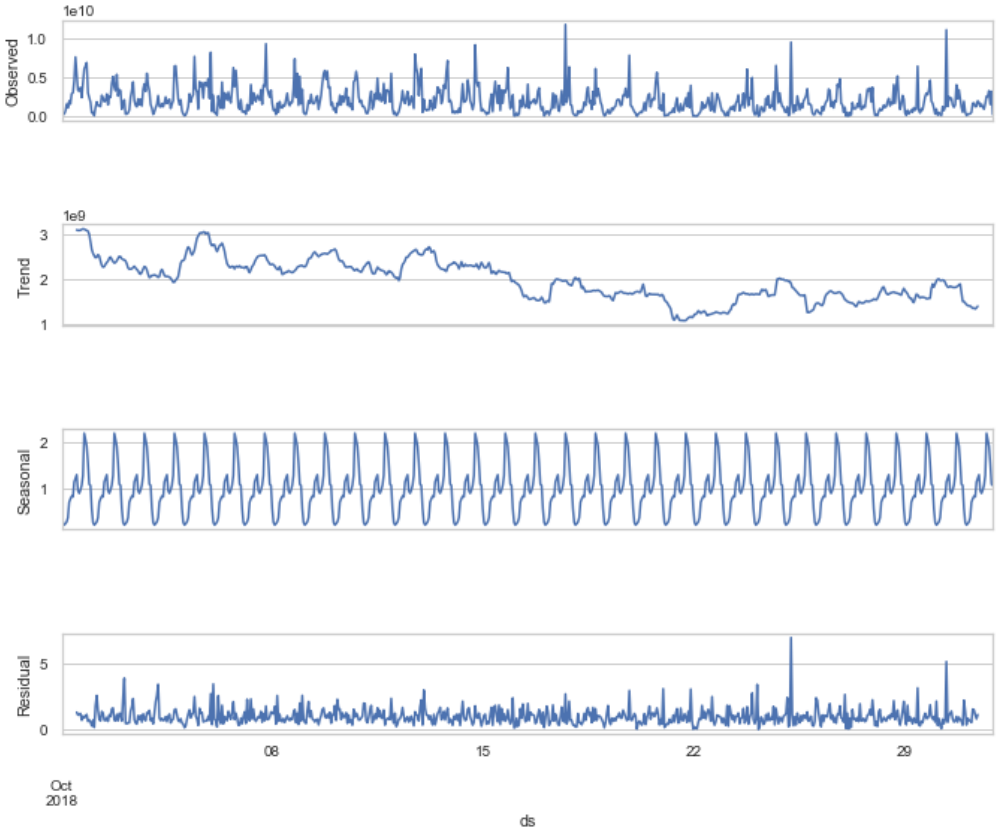


图 2-4 2018 年 10 月全网用户产生流量时域 STL 分解

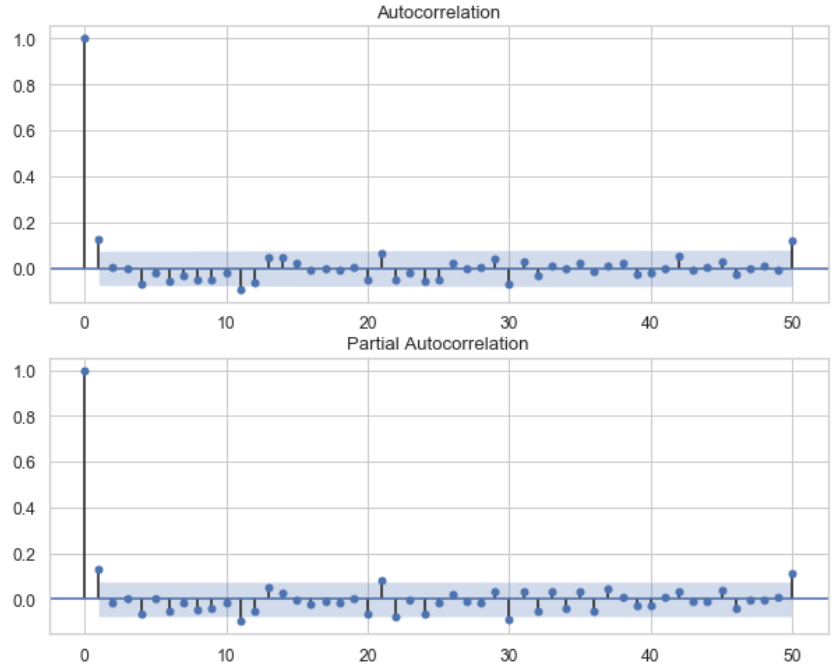


图 2-5 2018 年 10 月全网用户产生流量 STL 余项自相关与部分自相关图

如图 2-5 所示，对 10 月时间序列分解得到的余项求自相关与部分自相关，可看出相关值都很小，基本均处于置信区间内，即余项基本上是白噪音信号。因此可验证时间序列 STL 分解的效果很好。

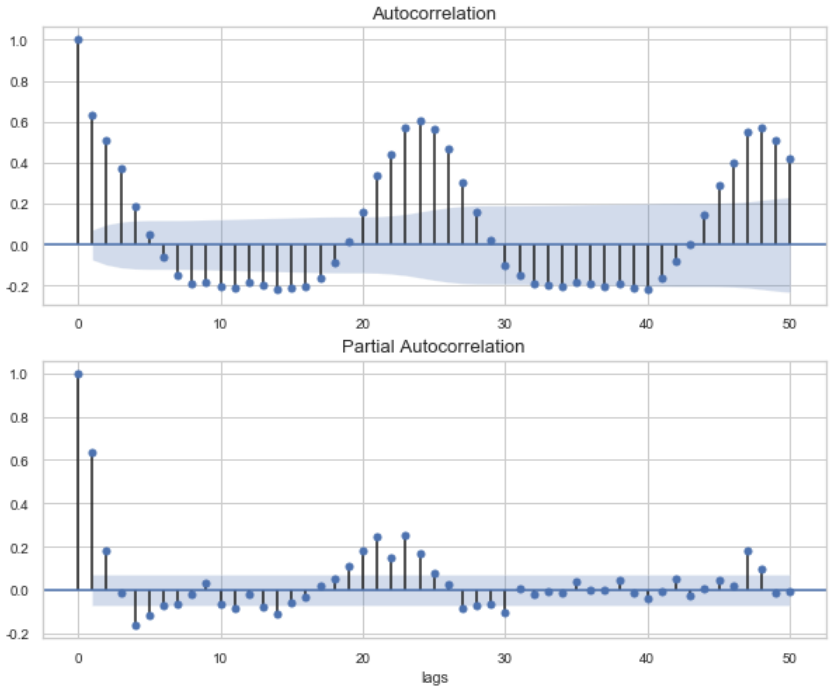


图 2-6 2018 年 9 月全网用户产生流量自相关与部分自相关函数图

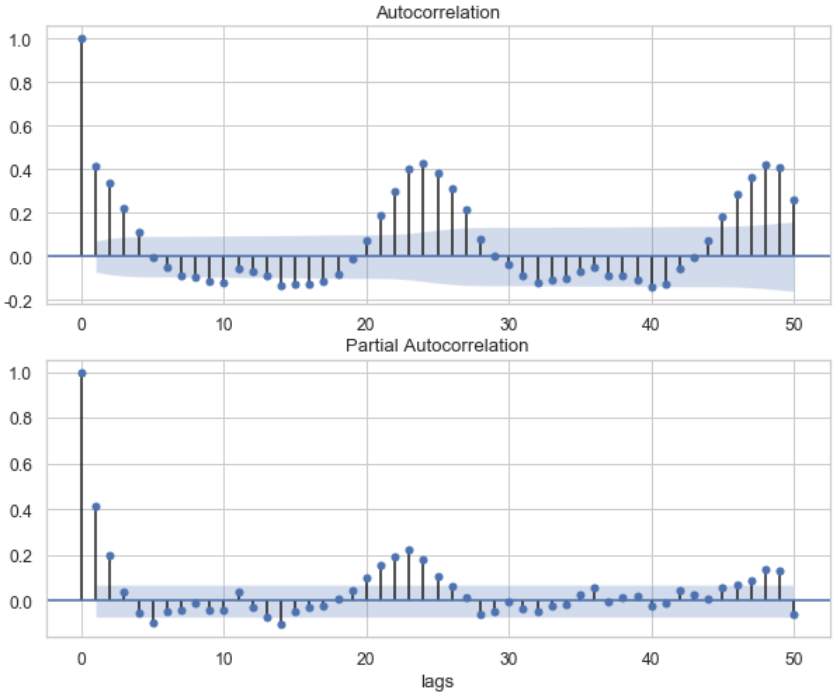


图 2-7 2018 年 10 月全网用户产生流量自相关与部分自相关函数图

如上所示图 2-6、图 2-7 分别为九月、十月全网用户流量自相关与部分自相关函数图，图中阴影区域为 95%置信区间，在置信区间内，可以认为序列是平稳的，超出置信区间的自相关系数反映出强相关性。可以看出全网流量不是平稳的，整体呈现出强烈的周期性。仔细观察可看出，相关性最高的时刻间隔为 24。这是因为数据是以小时为采样间隔，而用户使用流量以天的时间尺度呈现强烈的周期性。即每隔 24 小时，用户使用流量的行为重复性最强。

如下图 2-8 所示，为十月原始数据中全网用户在每个时刻使用流量的次数。首先可以发现整体上依然呈现明显的下降趋势，人们使用流量的频次越来越少。其次仔细观察

还可以发现，2018 年 10 月 6-7 日，10 月 13-14 日、10 月 20-21 日以及 10 月 27-28 日用户使用流量的次数比其它日期较低，正是因为这几个日期是周末。即相对工作日而言，周末用户使用流量次数较少。这与人们的工作生活习惯也是息息相关的。周末，人们通常长时间在家中休息，可以使用 WLAN，因此较少使用蜂窝网络。

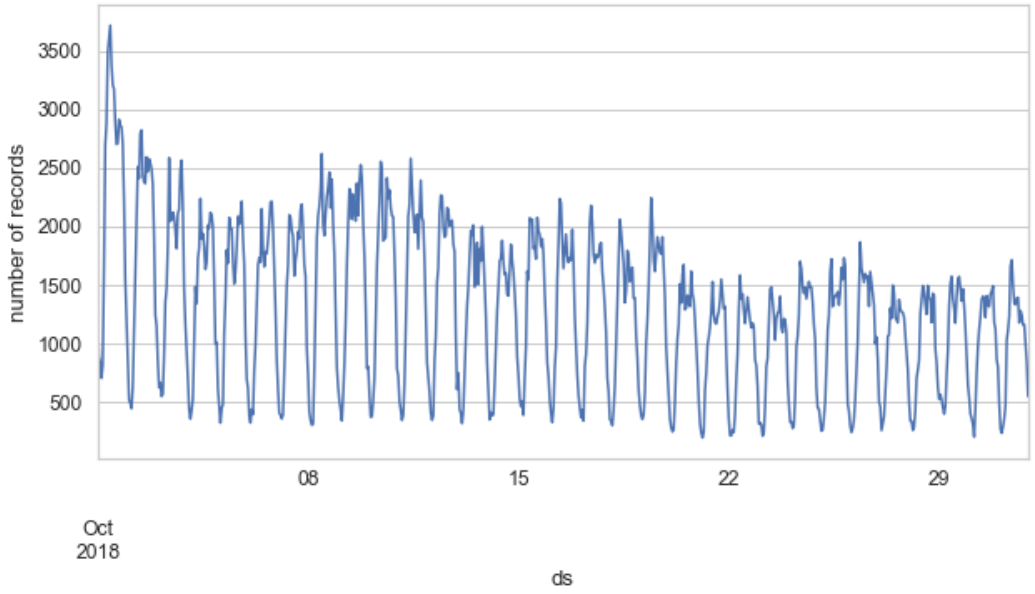


图 2-8 2018 年 10 月全网用户使用流量次数

从全网角度考虑聚类求和，所有用户使用流量的时长时域分布如下图 2-9 所示。统计所有用户每小时使用流量时长之和，总体趋势仍然是下降的。以天为时间尺度呈现一定的周期性，但是周末与工作日的差别并不十分明显。结合图 2-8、图 2-9 可以看出，周末人们使用流量的次数相对来说比工作日少，但是使用流量的时长差别不大。也就是说，虽然周末时，人们使用流量的次数变少了，但是每次使用流量的时间会变长，而工作日人们使用流量的习惯则是短时间多次。这显然也比较符合人们的生活工作习惯。

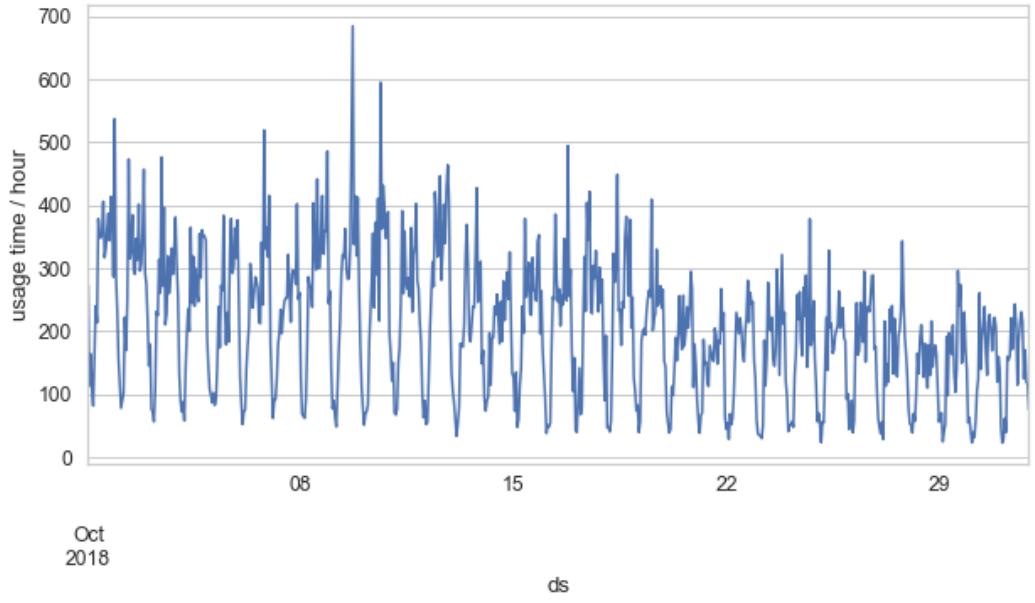


图 2-9 2018 年 10 月全网用户使用流量时长

2.4 个人用户流量使用时域分布及统计

上一节主要从全网用户角度，分析了流量使用的时域特征。本节将从个人用户角度，分析用户流量使用的时域特征与统计信息，同时研究单个用户使用手机应用软件的行为特征，以及用户移动性对用户产生流量的影响。

2.4.1 个人用户业务流量

使用直方图对用户使用流量次数的统计特性进行分析。如图 2-10 为 2018 年 9 月所有用户使用流量次数统计直方图，可以非常直观地看出绝大部分用户每月产生流量次数低于 100 次，大部分集中在 60 次以内。其中将近 20%的用户甚至在九月没有产生流量。计算得出，有流量使用记录的所有用户在九月产生流量使用记录次数平均值为 19 次，中位数为 4 次，属于右偏分布。由以上统计数据可以看出用户有流量产生的记录非常少。这符合著名的长尾理论所指出的，绝大部分访问网络记录都是由少数用户产生的，相应地，绝大部分用户只有很少的上网记录。

如图 2-11 图 2-12 所示，使用直方图及累积分布函数（Cumulative Distribution Function, CDF）分析用户平均每天使用流量。注意预先对数据进行了取对数处理。可以发现绝大部分用户平均每日使用流量集中在 1KB 到 1GB 之间，尤其集中在 0.1MB 至 100MB 区间内。观察用户平均每日使用流量 CDF 图，可以发现，60%的用户每日产生低于 1MB 的流量。

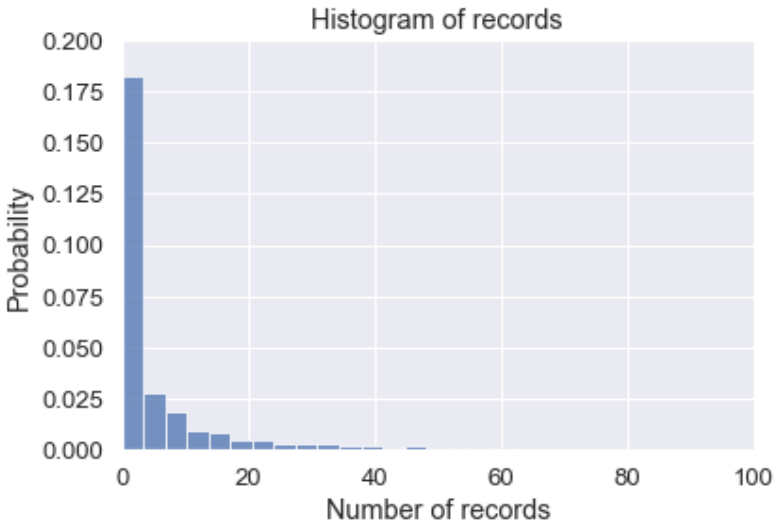


图 2-10 2018 年 9 月全网用户使用流量次数直方图

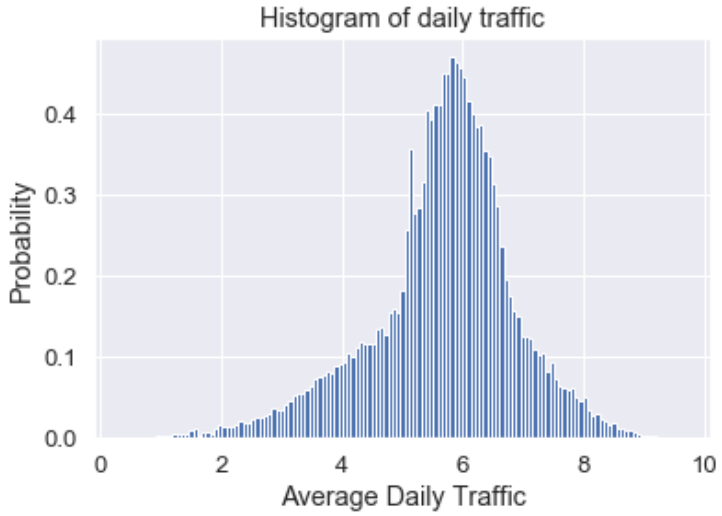


图 2-11 2018 年 9 月用户平均每日使用流量直方图

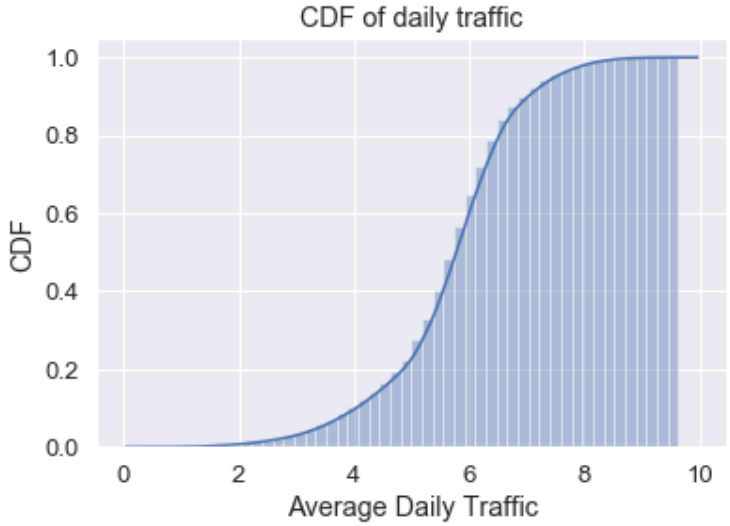


图 2-12 2018 年 9 月用户平均每日使用流量 CDF 图

2.4.2 个人用户访问手机应用软件行为

本文所用的数据一共记录了近七万名用户，在为期两个月内访问超过万种手机应用软件并产生流量使用的行为，因此下面分析用户访问手机应用软件行为特征。如图 2-13 所示为 2018 年 9 月使用次数最多的 50 个手机应用软件，可以看到除了微信、QQ、今日头条、支付宝、快手以及拼多多等耳熟能详的手机应用软件之外，还有手机营业厅、系统桌面、联系人、网络位置以及输入法等常用的手机自带系统工具。

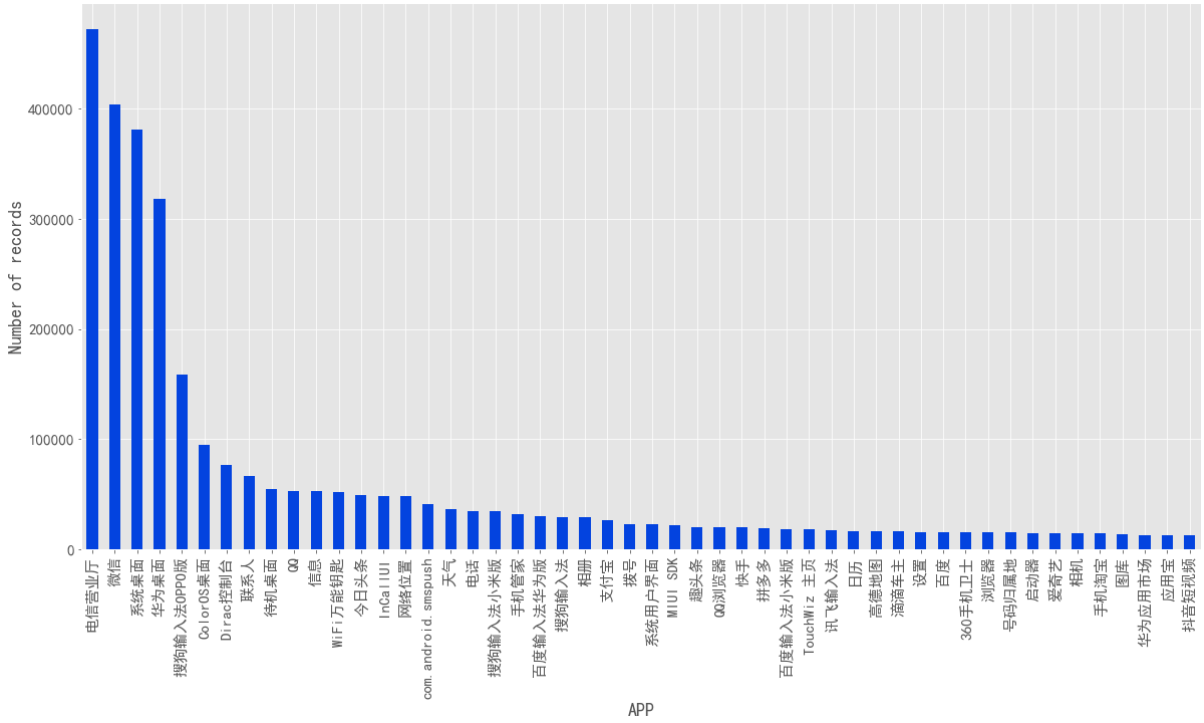


图 2-13 2018 年 9 月使用次数最多的 50 个手机应用软件

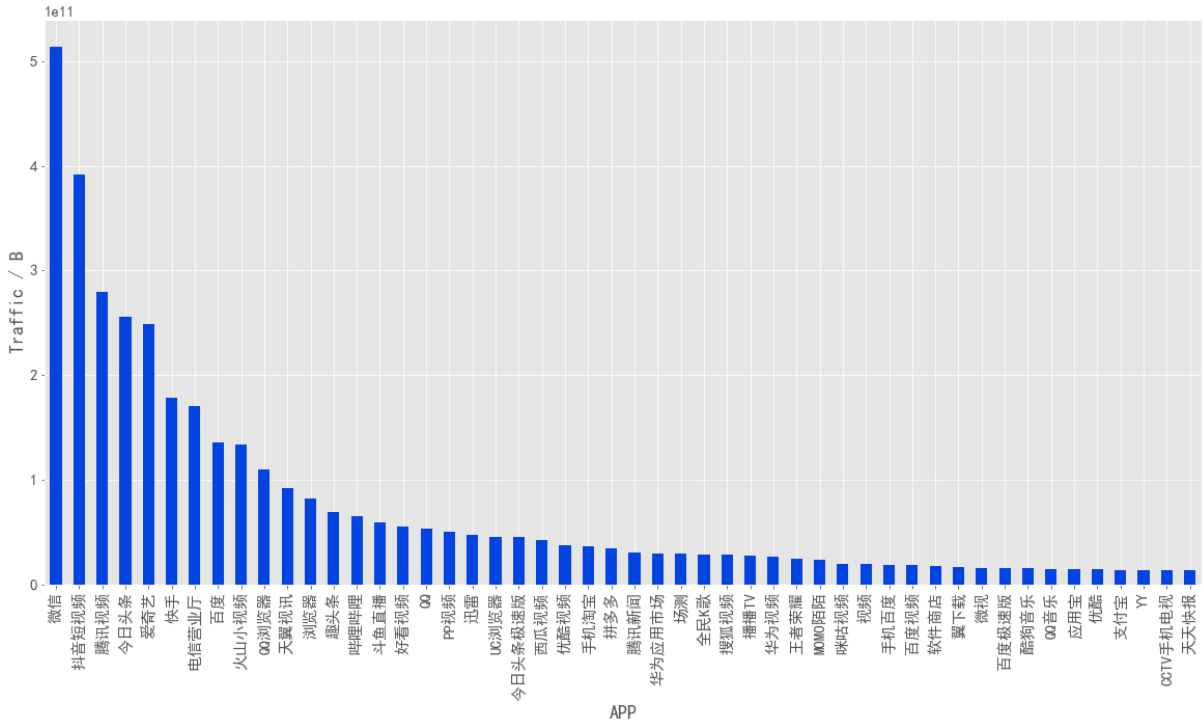


图 2-14 2018 年 9 月使用流量最多的 50 个手机应用软件

如上图 2-14 所示为 2018 年 9 月期间使用流量最多的 50 个手机应用应用软件，排名前十从高到低分别为微信、抖音短视频、腾讯视频、今日头条、爱奇艺、快手、电信营业厅、百度、火山小视频、QQ 浏览器。前十里有 5 个应用软件属于视频播放类软件，两个属于浏览搜索类手机应用软件，符合视频类软件产生流量大的特征。事实上，产生流量最多的前 50 个手机应用软件占万种手机应用软件产生流量总数的 83%。这也在另一个角度反映了长尾效应。

首先分析平均每天访问手机应用软件时长。如图 2-15 所示，大部分用户每日使用手机应用软件时长在 6 小时以内，70%的用户使用手机应用软件的时长甚至在一小时以内。计算得到，用户平均每日使用手机应用软件时长平均值为 64 分钟，中位数为 9 分钟。

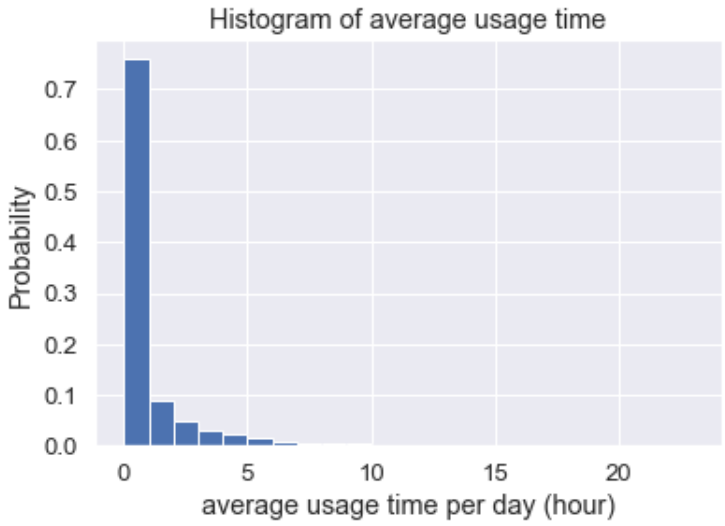


图 2-15 2018 年 9 月用户平均每日访问手机应用软件时长直方图

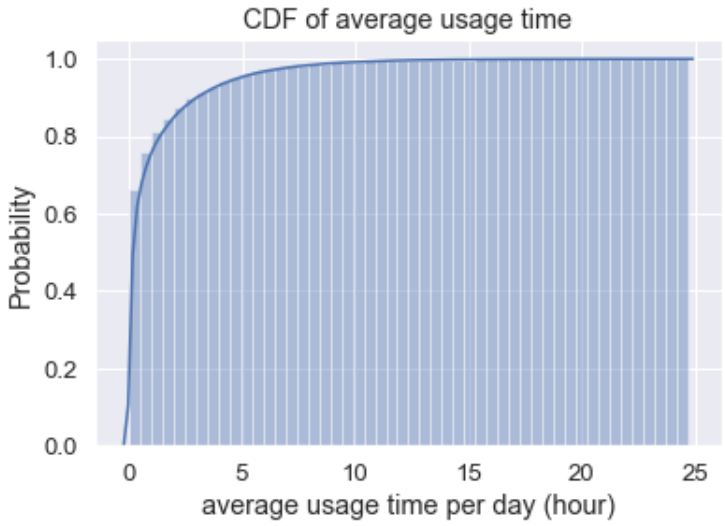


图 2-16 2018 年 9 月用户平均每日访问手机应用软件时长 CDF 图

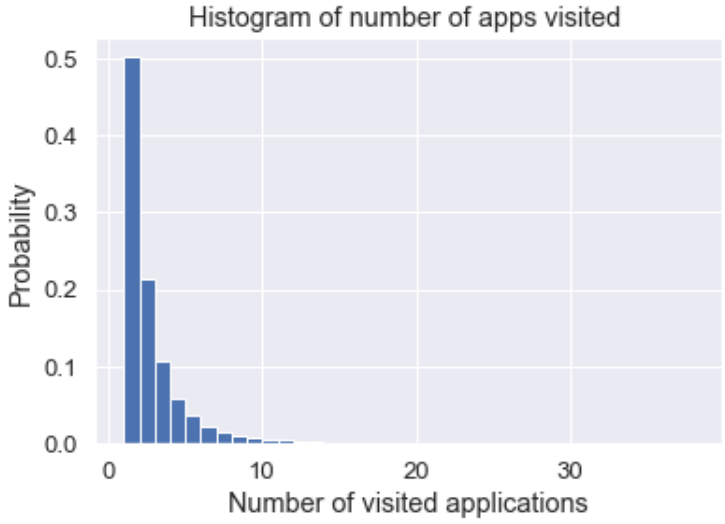


图 2-17 2018 年 9 月用户平均每日访问手机应用软件个数直方图

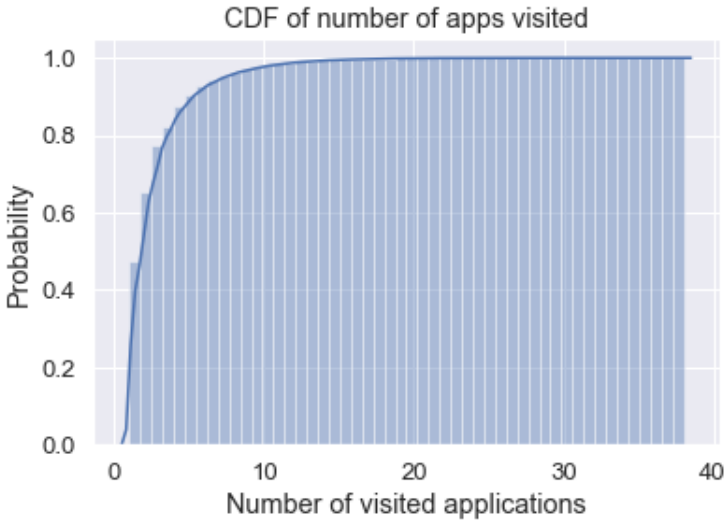


图 2-18 2018 年 9 月用户平均每日访问手机应用软件个数 CDF 图

再分析用户平均每天访问手机应用软件的数目，如图 2-17 所示，大部分用户平均每日访问手机应用软件个数在 15 个以内。图 2-18 显示 80%的用户每天使用 3 个手机应用软件，仅 10%的用户每天使用超过 5 个手机应用。

2.4.3 用户移动性

下面分析用户移动性特征。考虑到仅十月数据含有 GPS 信息，因此以下分析使用 2018 年 10 月数据。一共有近 7000 名用户含有 GPS 信息，用户位置均在中国上海。文献[21]中用重要地点，行程距离，活动半径，一定时间内访问的地点数目来衡量用户的移动性。如下图 2-19 所示，所有用户平均每天访问的地点在 25 个以内。其中 80%的用户每日访问地点少于 3 个。90%的用户每日访问地点少于 4 个，只有 1%的用户每日访问地点在八个或以上。

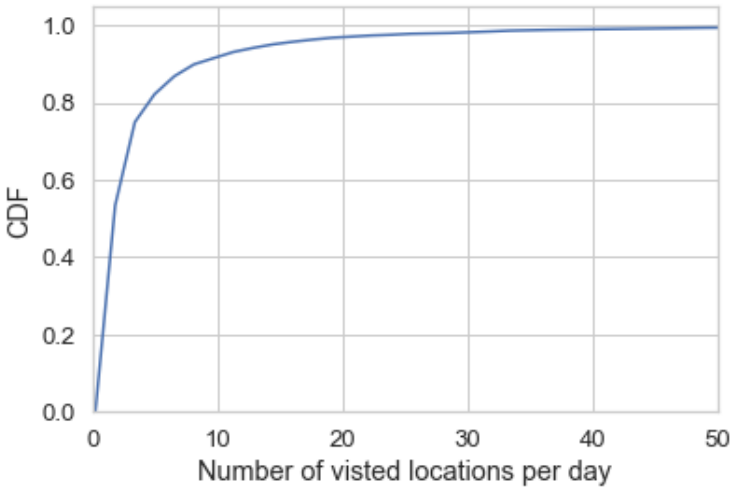


图 2-19 2018 年 10 月用户平均每日访问地点数目 CDF 图

定义访问地点数目最多的 20%用户为高移动性用户，其余 80%为低移动性用户。首先分析两类用户产生流量的行为差异性。如图 2-20 所示，可以看出高移动性用户平均使用流量几乎所有时刻都比低移动性用户多。计算得出该 20%高移动性用户使用总流量占

所有用户使用流量总和的 90%。

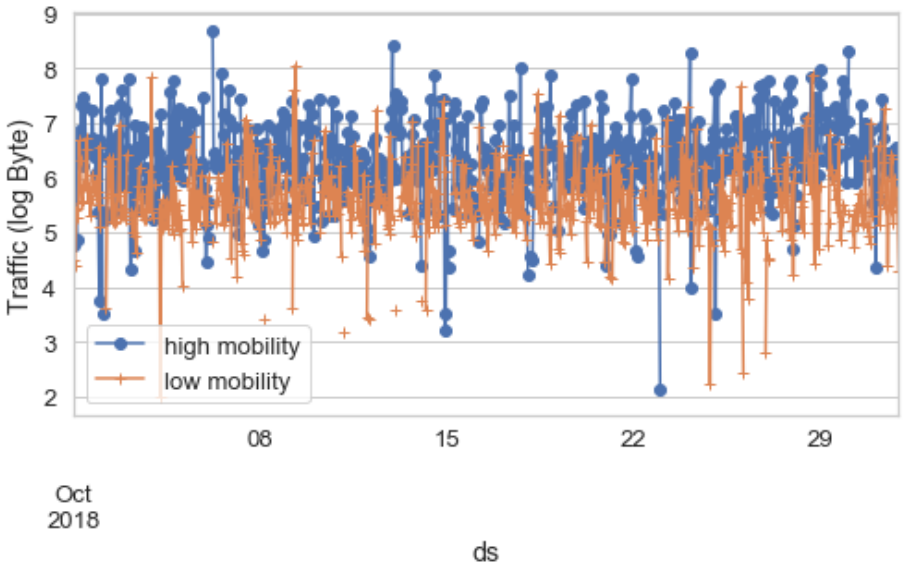


图 2-20 2018 年 10 月不同移动性用户平均产生流量时域图

如下图 2-21 所示，不同移动性用户平均每天产生流量的累积分布函数图。与图 2-20 一致，从图中也可以看出，高移动性用户普遍使用流量较低移动性用户多。超过 80%的低移动性用户产生每日产生流量低于 1MB，而只有将近 50%的高移动性用户产生流量低于 1MB。以上分析发现都揭示了用户移动性与用户使用流量之间强烈的相关性。

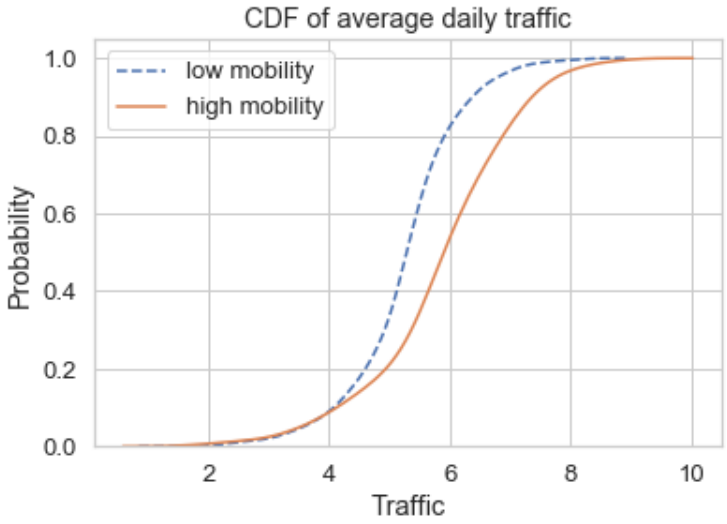


图 2-21 2018 年 10 月不同移动性用户平均产生流量 CDF 图

其次对不同移动性的用户，分析其使用手机应用软件行为的差异性。如下图 2-22 所示，为高移动性用户和低移动性用户平均每日访问手机应用软件数目累积分布函数图。从图中可以看出，高移动性用户每日使用的手机应用软件数目比低移动性用户多。将近 20%的高移动性用户每日使用 5 个及以上的手机应用软件，而这个比例在低移动性用户只有不到 5%。

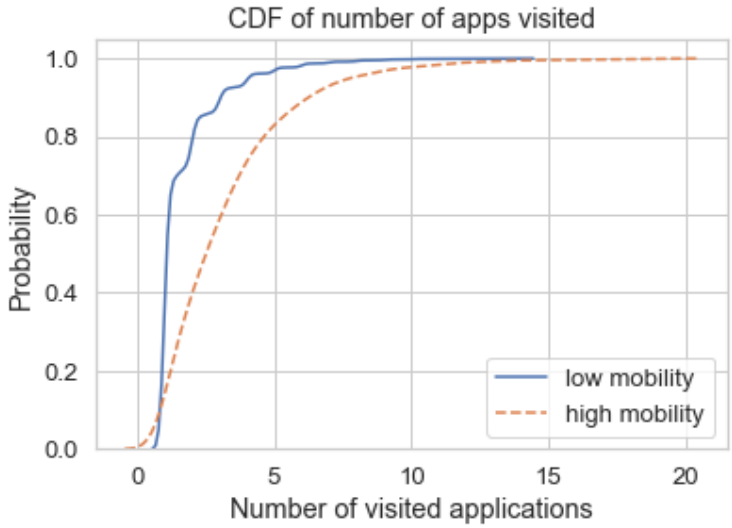


图 2-22 2018 年 10 月不同移动性用户平均每日使用应用软件数目 CDF 图

由上面的分析可以看出，不同用户的移动性差别较大。不仅如此。用户的移动性也会影响其使用流量的行为和访问手机应用软件的行为。

2.5 本章小结

本章首先对所用数据进行了概述，并简要地说明了数据预处理方法。然后分别从全网用户角度和个人用户角度对用户使用流量时间序列进行了多角度深入的分析。

首先，对全网用户使用流量时间序列进行 STL 分解，发现了其中的强烈周期性与显著的下降趋势。同时通过分析时间序列的自相关函数，更加验证了全网用户使用流量时间序列以天为时间尺度的强周期性。通过对比工作日与周末的流量使用差异，发现周末时，人们使用流量的次数变少了，但是每次使用流量的时间会变长，而工作日人们使用流量的习惯则是短时间多次。这显然也比较符合人们的生活工作习惯。

再从个人用户角度，分别分析研究了用户流量使用特征，用户访问手机应用软件特征。分析发现用户使用流量行为存在极为明显的长尾效应，即绝大部分访问网络记录都是由少数用户产生的，相应地，绝大部分用户只有很少的上网记录。发现 60%的用户每日产生低于 1MB 的流量。产生流量最多的前 50 个手机应用软件占万种手机应用软件产生流量总数的 83%。70%的用户使用手机应用软件的时长在一小时以内。

最后，本章分析了用户的移动性特征，发现 80%的用户每日访问地点少于 3 个。为了研究用户移动性特征与用户使用流量及访问手机应用软件行为之间的相关性，本章将访问地点数目最多的前 20%用户，定义为高移动性用户，其余 80%定义为低移动性用户，研究两种用户的行为差异性。结果表明高移动性用户倾向于使用更多流量，并访问更多手机应用软件。

第三章 用户使用流量时序特征提取及用户聚类

3.1 引言

前一章主要从整体移动通信网的角度出发研究了用户流量时域分布规律。同时分析了用户访问手机应用软件的行为习惯，例如，每日使用手机应用软件的次数，时长，所产生的流量等，并且结合用户的移动性进行了分析与研究。在研究个人用户使用流量大小及时长的时域分布规律时也发现用户之间的差异性很大。因此，除了用户差异性之外，本章考虑发掘用户之间的相似性，将用户聚类到不同类别。

目前文献研究中对基站负载的聚类算法较多，而对用户的聚类研究较少。文献[23][24][27][49]研究基于基站的聚类。文献[23][28]基于基站负载时域特征进行聚类。文献[24]先对时间序列进行离散傅里叶变换，然后基于频率特征对基站进行聚类。文献[49]则采用因子分析法对基站负载时间序列进行降维，再基于通过因子分析法得到的公共因子对基站进行聚类。文献[50]将因子分析法应用到个人用户流量聚类分析中。

本章第二小节提取出用户使用流量的时域特征，再应用因子分析法对特征向量进行降维处理，得到五个公共因子，对应提取出来的五个重要时段。本章第三小节对降维后的特征向量应用 k 均值聚类算法，将用户聚类到不同类别，得到六种不同类型的用户。分析不同类型用户使用流量的行为，并对其进行语义标注。

3.2 用户流量时序特征提取及压缩

由于前一章节的分析发现 2018 年 9 月与 2018 年 10 月数据特征相似，因此本章选用 2018 年 9 月数据进行研究分析。要研究单个用户流量使用的时域分布，首先将用户使用流量数据按照一小时采样间隔重新采样，一共会衍生出 24*30 = 720 维特征，即用户在 30 天每一小时产生的流量数据。由于个人用户流量使用存在以周为时间尺度的周期性，因此对九月四周数据进行求平均，得到 24*7=168 维特征向量，即用户平均一周每小时产生的流量数据。特征数过大会使得计算量加大，难以进行数据可视化分析，模型的可解释性变差。同时并非所有的特征都对模型有贡献，特征的冗余会影响算法模型的性能。因此，首先要对原始时间序列数据进行降维处理。

3.2.1 降维方法

机器学习领域的降维方法有两种类型：特征选择和特征提取。特征选择旨在从原始特征集中选取一组最有统计意义的特征。而特征提取则是对原始特征进行一些变换，使其具有明显的物理意义、统计意义或核的特征。特征提取和特征选择均可以使特征的维度降低，减少冗余，减少存储数据和输入数据的带宽，并发现最有意义的特征和变量，帮助对数据进行进一步分析和挖掘。特别地，将高维数据映射到二维或三维特征空间中还可以实现数据的可视化，对数据有更直观的了解和分析。本章主要使用特征提取方法。

目前机器学习领域已有很多较为成熟的降维和特征提取的方法，比如小波分解(Wavelet Decomposition)[41]，因子分析法(Factor Analysis, FA)[42][43]，主成分分析法(Principal Component Analysis, PCA)[44]，奇异值分解(Singular Value Decomposition, SVD)[44]，线性判别分析(Linear Discriminant Analysis, LDA)[45]等。最为常用的主成分分析中，主成分分析法旨在寻找最大化数据方差的正交线性组合。也就是说，主成分是原始变量的线性组合，将变量以不同的系数组合起来，得到若干个复合变量，然后从中选择最能体现整体的复合变量作为主成分。主成分分析法在保留主要信息的基础上，达到简化和降维的目的。主成分的数量相对于原始数量更少，主成分保留了原始变量的大部分信息，主成分之间相互独立。缺点是主成分分析法没有明确的模型，可解释性较差。因此在这里使用因子分析法对原始特征向量进行降维处理。

3.2.2 因子分析法

因子分析则与主成分分析法不同，因子分析法将变量拆分开，分为公共因子与特殊因子。通过研究众多变量之间的内部依赖关系，探求观测数据的基本结构，并用少数几个假想变量（因子）来表示原始数据。因子能够反映众多原始变量的主要信息。因子个数远远少于原始变量个数，因子并非原始变量的简单取舍，而是一种新的综合。因子之间没有线性关系，因子具有明确解释性，可以最大限度地发挥专业分析的作用。因子分析法优点是新变量具有实际的意义，能解释原始变量间的内在联系。

假设可观测随机变量 $\boldsymbol{x}=(x_1,\cdots,x_p)$ 维度为 p ,其均值为 0，方差为 $\boldsymbol{\Sigma}$，即 $\mathrm{E}(\boldsymbol{x})=0$，$\boldsymbol{\Sigma}=\mathrm{cov}(\boldsymbol{x})$。设 $\boldsymbol{f}=(f_1,\cdots,f_m)$ 是不可观测的随机变量，维度为 m，且 $m<p$。$\mathrm{E}(\boldsymbol{f})=0$，$\boldsymbol{\Phi}=\mathrm{cov}(\boldsymbol{f})$。$\boldsymbol{\Lambda}$ 为 $p\times m$ 维矩阵。又设错误项 $\boldsymbol{\varepsilon}=(\varepsilon_1,\cdots,\varepsilon_m)$ 与 $\boldsymbol{f}$ 互不相关，且其期望为 0，方差为对角阵，即 $\mathrm{cov}(\boldsymbol{\varepsilon},\boldsymbol{f})=0$，$\mathrm{E}(\boldsymbol{\varepsilon})=0$，$\boldsymbol{\Psi}=\mathrm{cov}(\boldsymbol{\varepsilon})$。如果 $\boldsymbol{x}$ 可由 $\boldsymbol{f}$ 线性表达，加上错误项 $\boldsymbol{\varepsilon}$ ，即满足以下的模型：

$$\boldsymbol{x}=\boldsymbol{\Lambda}\boldsymbol{f}+\boldsymbol{\varepsilon} \tag{3.1}$$

则称模型为正交因子模型。其中 $\boldsymbol{f}$ 是隐含特征矢量（latent feature vector），$f_1,\cdots,f_m$ 称为 $\boldsymbol{x}$ 的公共因子。$\varepsilon_1,\cdots,\varepsilon_m$ 称为 $\boldsymbol{x}$ 的特殊因子。公共因子 $f_1,\cdots,f_m$ 一般对 $\boldsymbol{x}$ 的每一个分量 x_i 都有作用，而特殊因子 ε_i 只对 x_i 起作用。而且各个特殊因子之间以及特殊因子与所有公共因子之间都是互不相关的。模型中的 $\boldsymbol{\Lambda}_{p\times m}$ 是待估的系数矩阵，称为因子载荷矩阵（loading matrix）。$\boldsymbol{\Lambda}_{ij}$ 称为第 i 个变量在第 j 个因子上的载荷（简称为因子载荷）。$\boldsymbol{\Lambda}$ 中元素刻画变量 x_i 与 f_j 之间的相关性 $\boldsymbol{\Lambda}_{ij}$ 越大，表明特征向量中的第 i 维特征与公共因子 j 有越强的相关性。一般用极大似然估计来求得因子载荷矩阵和隐含特征向量。

3.2.3 特征降维

因子分析的目的是从原有众多变量中综合出少量具有代表意义的因子变量，这必定有一个潜在的前提要求，即原有变量之间应具有较强的相关关系。因此，在因子分析时，

首先需要对原有特征矢量进行相关分析。文献[46]提出 KMO（Kaiser-Meyer-Olkin, KMO）检验统计量，来验证因子分析法的合理性。KMO 检验统计量是用于比较变量间简单相关系数和偏相关系数的指标，主要应用于多元统计的因子分析。KMO 统计量计算公式如下：

$$KMO = \frac{\sum r_{ij}^2}{\sum r_{ij}^2 + \sum u_{ij}^2} \tag{3.2}$$

其中，$R=[r_{ij}]$为相关矩阵，$U=[u_{ij}]$为偏相关矩阵。不难发现，KMO 统计量取值在 0 和 1 之间。当所有变量间的简单相关系数平方和远远大于偏相关系数平方和时，KMO 值接近 1，意味着变量间的相关性越强，原有变量越适合作因子分析。当所有变量间的简单相关系数平方和接近 0 时，KMO 值接近 0，意味着变量间的相关性越弱，原有变量越不适合作因子分析。文献[46]给出了常用的 KMO 度量标准，0.9 以上表示数据非常适合进行因子分析，0.5 以下表示极不适合。本文中 KMO 指数超过 0.976，表示非常适合进行因子分析。此外，因子分析法通常要求样本数量远远大于特征维度，而所用数据样本数量是特征维度的400倍以上，因此也满足因子分析法的要求，这也是大数据的优点之一。

本章选用因子分析法检测一天内的时刻间的相关性，从而降低特征矢量维度。用探索性因子分析法（Explanatory Factor Analysis, EFA）将相关的时刻合并起来，得到更有代表性的时间段。因子载荷矩阵表征了用户在不同时刻的流量使用之间的相关性，帮助合理地组合时间。公共因子的数目是事先给定的，由原始特征向量数据的方差矩阵的特征值确定。由于一个公共因子会产生一个大的特征值，所以公共因子的数目可由较大的特征值的数目确定。选定公共因子数目为 5。

对 2018 年九月用户使用流量数据时域特征进行因子分析法降维，通过因子分析法可从特征向量矩阵中提取出五个公共因子。因子载荷矩阵如下图 3-1 所示。图中横轴表示一天 24 小时，纵轴表示一周七天，即时域 168 维特征。因子载荷矩阵体现出每一维特征与公共因子的相关性。从上到下，从左至右分别为公共因子一至五。从图中可以看出，公共因子一，主要与工作时间的用户流量使用行为强烈相关，主要时段包括工作日的早七点至晚七点。与公共因子二相关的时隙则主要集中在午夜至凌晨时段，从晚上十一点至次日凌晨三点。公共因子三相关的时段主要在早间，从早上三点至六点。公共因子四包含的时段主要是周末的白天休息时段。公共因子五主要是晚间，从晚上七点至午夜。如表 3-1 所示为公共因子语义标记及代表的具体时段。

表 3-1 公共因子语义标记

公共因子	标记	时段
0	工作时间	周一至周五的早七点至晚七点
1	夜半	晚上十一点至早晨三点
2	凌晨	早上三点至六点
3	周末的白天时段	周六周日早七点至晚八点
4	晚间	晚上六点至十点

应用因子分析法后，特征矩阵从原始的 168 维降至 5 维，大大降低了特征的冗余度，使得后续的聚类算法运算速度更快，聚类效果更好。

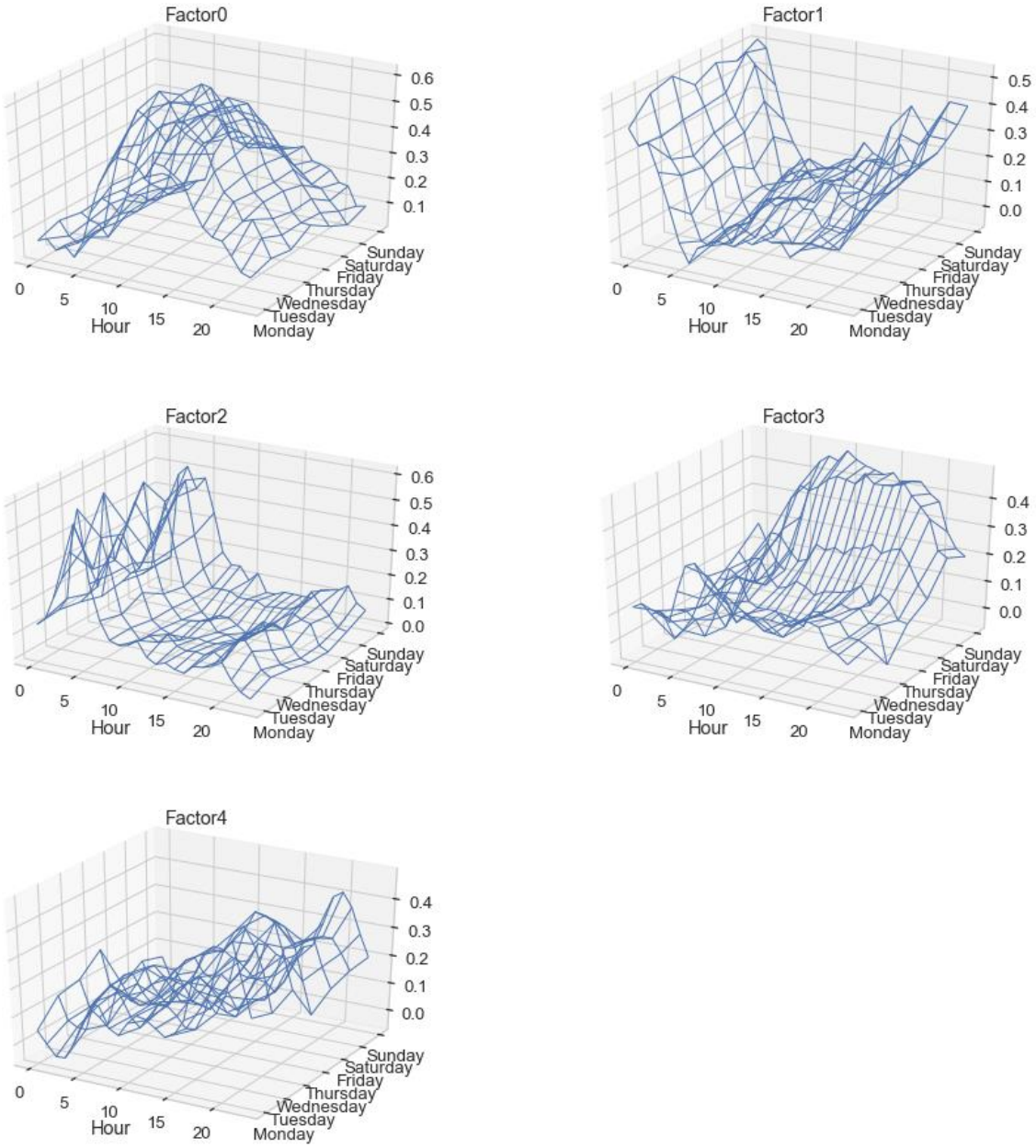


图 3-1 2018 年 9 月数据特征向量因子载荷矩阵

3.3 用户聚类及用户类型识别

聚类（Clustering）分析属于机器学习中无监督学习（Unsupervised Learning）的一种。与有监督学习不同，如常见的分类问题，具备标注好分类类别的训练数据，通过这些数据训练出一个模型来对新的数据的分类进行预测。在无监督学习中，数据样本的标记信息时位置的，目标时通过对无标记训练样本的学习来揭示数据的内在性质和规律，为进一步的数据分析提高基础。聚类分析是其中研究最多，应用最广泛的一种。

聚类试图将数据集中的样本划分为若干个通常是不相交的子集，每个子集称为一个簇（cluster）。通过这样的划分，在同一个子集中的成员对象往往都有相似的一些属性。但值得注意的是，聚类过程仅能自动形成簇，并不能生成对每个簇对应的概念语义。

假定样本集合 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$ 包含 m 个无标记样本，每个样本 $\boldsymbol{x}_i=(x_{i1},x_{i2},\ldots,x_{in})$ 是一个 n 维向量，则聚类算法将样本集 D 划分为 k 个不相交的簇 $\{C_l\mid l=1,2,\ldots,k\}$，其中 $C_{l'}\bigcap_{l'\neq l}C_l=\Theta$ 且 $D=\bigcup_{l=1}^{k}C_l$。相应地，可用 $\lambda_j\in\{1,2,\ldots,k\}$ 表示样本 x_j 的簇标记，即 $x_j\in C_{\lambda_j}$。于是，聚类的结果可用包含 m 个元素的簇标记向量 $\lambda=(\lambda_1,\lambda_2,\ldots,\lambda_m)$ 表示。

3.3.1 K 均值聚类

K 均值算法[47]是最为常用的聚类算法之一。该算法使用两个样本之间的距离作为相似度度量。距离越大，相似度越小。K 均值算法旨在寻找使得所有样本点到其所属簇中心距离之和最小的聚类方案。给定样本集合 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$， k 均值算法针对聚类所得簇划分 $C=\{C_1,C_2,\ldots,C_k\}$ 最小化平方误差：

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}\left\|\boldsymbol{x}-\boldsymbol{\mu}_i\right\|^2 \tag{3.3}$$

其中

$$\boldsymbol{\mu}_i=\frac{1}{|C_i|}\sum_{\boldsymbol{x}\in C_i}\boldsymbol{x} \tag{3.4}$$

是簇 C_i 的中心（均值向量）。

最小化式(3.3)并不容易，找到最优解需考察样本集所有可能的簇划分，这是一个 NP 难问题。因此，一般 k 均值算法采用贪心策略，通过迭代优化来近似求解。算法流程如下算法 3.1 所示。第 1 行随机选择簇初始中心点。第 2-7 行分配各个样本点到距离其最近的簇。第 8-10 行，根据样本归属结果，更新簇中心点。若迭代更新后聚类结果保持不变，或已达到最大迭代次数，则输出簇划分结果及各样本点所属情况。

算法 3.1 K 均值聚类

输入：聚类簇数 k，样本数目 n，特征数据 $\mathbf{X}$

输出：k 个簇 $(C_1,C_2,\ldots,C_k)$，包括聚类中心及各样本点所属情况

1. 初始化。随机选择 k 个样本点作为聚类中心
2. 分配样本点到各个簇
3. For i=1 to n（对每个样本点）：
4. For j = 1 to k（对每个簇中心）：
5. 计算每个样本点与每个簇中心的距离 $dist(\boldsymbol{x}_i,\boldsymbol{c}_j)$
6. End for
7. 根据距离最近的中心点确定样本点 $\boldsymbol{x}_i$ 的标记为 $\lambda_i=\arg\min dist(\boldsymbol{x}_i,\boldsymbol{c}_j)$
8. 将 $\boldsymbol{x}_i$ 分配至距离其最近的簇 C_{λ_i}
9. End for
10. 更新簇中心
11. For j = 1 to k（对每个簇中心）：
12. 根据第二步对样本点的归属情况，更新每个簇的中心

$$\boldsymbol{c}_j=\frac{1}{|C_j|}\sum_{\boldsymbol{x}\in C_j}\boldsymbol{x}$$

13. End for
14. 重复上述第二、三步直到聚类中心不再移动或变化幅度小于阈值或达到最大迭代次数

预先给定的聚类簇数目与初始簇中心的选择对算法最终效果有很大影响。因此，如何选取合适的 k 值非常重要。聚类簇数目的选择依赖于聚类算法评价指标。选取戴维森堡丁指数（Davies-Bouldin Index, DBI）[48]作为聚类算法性能度量。其计算公式如下：

$$DBI = \frac{1}{k}\sum_{i=1}^{k}\max_{j=1,j\neq i}^{k}\frac{\overline{C_i}+\overline{C_j}}{M_{ij}}, \tag{3.5}$$

这里，$\overline{C_i}$ 是聚类内所有点到中心点的平均距离。M_{ij} 是两个聚类之间的距离。$\overline{C_i}$ 和 M_{ij} 的计算公式如下：

$$\bar{C}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\left\|\boldsymbol{c}_i - \boldsymbol{x}_j\right\|_2, M_{ij} = \left\|\boldsymbol{c}_i - \boldsymbol{c}_j\right\|_2 \tag{3.6}$$

其中 N_i 是聚类 i 中的用户数目，$\boldsymbol{c}_i$ 是聚类 i 的中心点，$\boldsymbol{x}_j$ 是用户 j 的特征矢量。

聚类是将样本集划分为若干互不相交的子集，即样本簇。同一簇内的样本尽可能彼此相似，不同簇的样本尽可能不同。即聚类结果的簇内相似度越高且簇间相似度越低，则聚类结果越好。戴维森堡丁指数正式反映了聚类间距离与聚类内部距离的比例关系。即同一聚类间用户距离越小，不同聚类距离越大，戴维森堡丁指数越小。基于戴维森堡丁指数，最终选定簇数目为 6。

3.3.2 用户类型分析

K 均值聚类必须先对数据进行归一化，该算法对数据大小尺度及异常点很敏感。用 k 均值聚类算法对用户进行聚类分析，最终得到 6 个聚类。如图 3-2 所示，被分到聚类 0，聚类 1，聚类 2，聚类 3，聚类 4，聚类 5 得用户比例分别为 13.1%，9.0%，9.9%，54.0%，5.4%，8.6%。

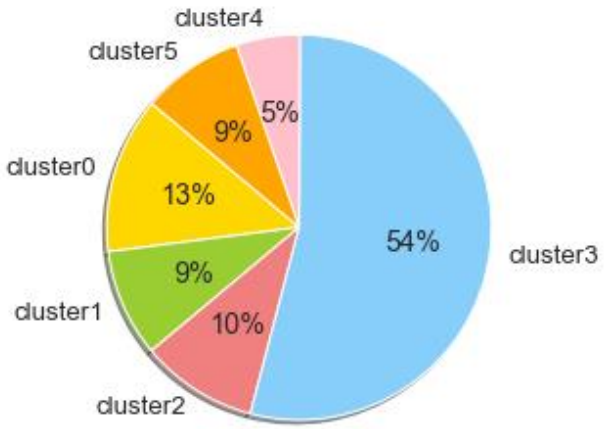


图 3-2 2018 年 9 月基于流量使用用户聚类比例

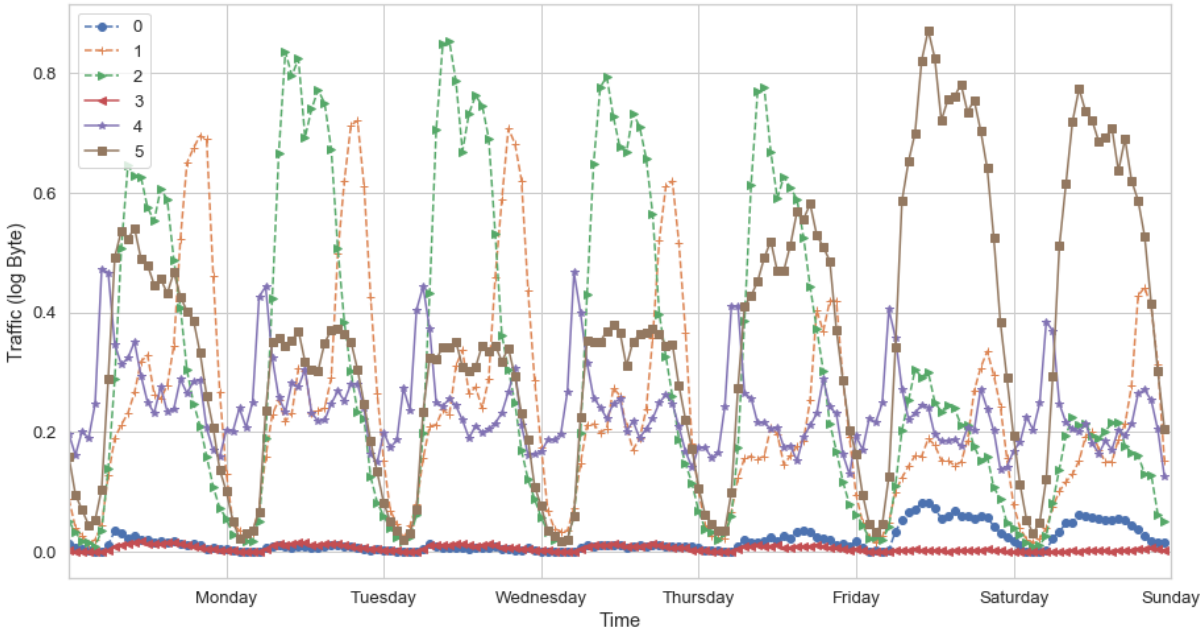


图 3-3 2018 年 9 月用户聚类结果

下面对每个用户聚类进行分析。对每个聚类内的所有用户求其平均使用流量，得到时域分布如图 3-3 所示。图 3-4 为六个聚类内所有用户的平均流量使用。

聚类 0 整体产生流量较小，属于轻度流量用户。且其周末较工作日产生流量更多。

聚类 1 产生流量很多，属于重度流量用户。此外，用户每日产生流量呈现双尖峰，尤以晚上八点左右最为明显，表明用户在晚间休息时段较白天工作时间产生更多流量。用户周末产生流量比工作日略少。

聚类 2 为流量重度用户，其主要特征为工作日产生流量多，而周末产生流量少。

聚类 3 所占比重最多，和聚类 0 相似，聚类 3 内的用户产生的平均流量很少，属轻度流量用户。与聚类 0 不同的是，聚类 3 用户周末产生流量较工作日少。值得注意的是，在第二章对个人用户流量使用时域统计特征分析时，发现 60%的用户每日产生流量低于 1MB。这里用户聚类结果也验证了之前的分析，即大部分用户每日产生流量很少。

聚类 4 产生流量较多，尤其集中在早间，且产生流量时段较长，几乎长期保持使用流量大于零的状态。

聚类 5 用户产生流量较多，且周末比工作日产生流量多。注意到受周末影响，周一和周五该聚类内用户产生流量也较多。

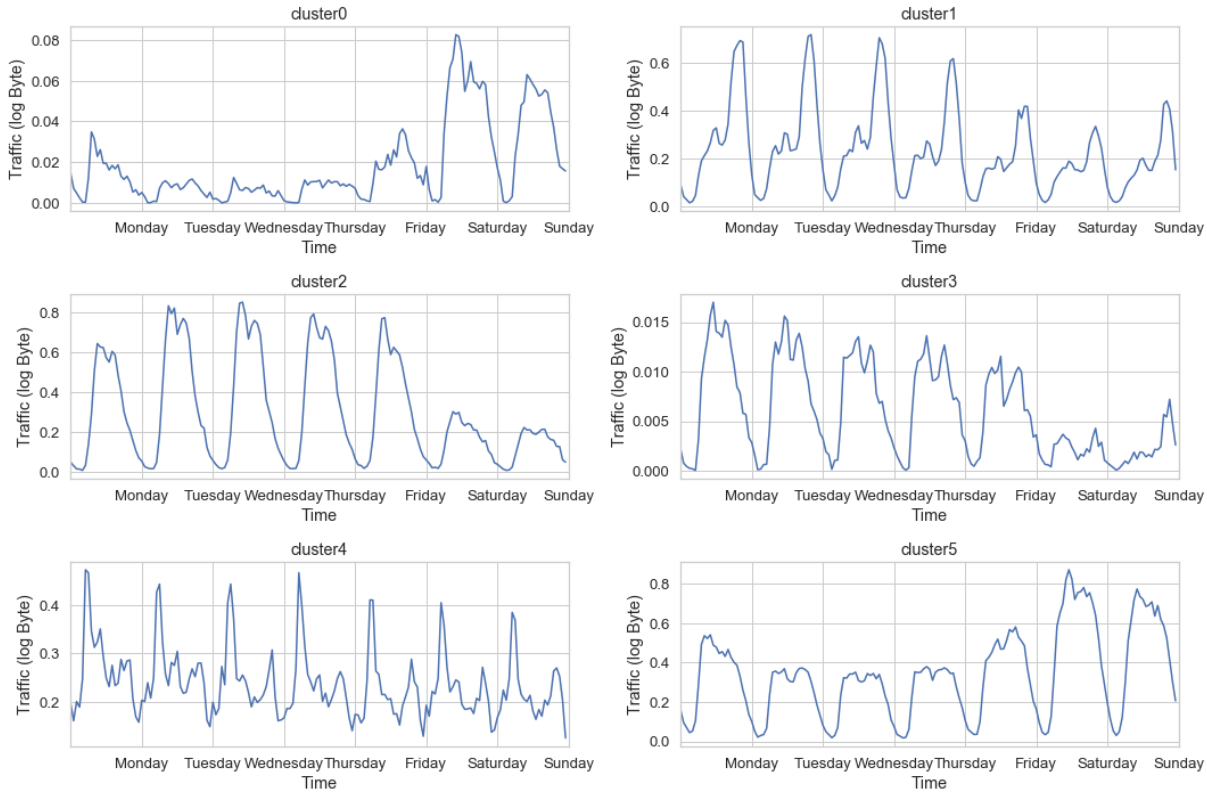


图 3-4 2018 年 9 月六种聚类用户流量时域图

根据上述六种类型的用户使用流量时域特征，对各聚类进行语义标记如下表 3-2 所示。

表 3-2 用户聚类类型标记及占比

类型	占比	语义标记
聚类 0	13%	周末轻度流量用户
聚类 1	9%	晚间重度流量用户
聚类 2	10%	工作日重度流量用户
聚类 3	54%	工作日轻度流量用户
聚类 4	5%	早间重度流量用户
聚类 5	9%	周末重度流量用户

分析不同聚类用户产生总流量，发现占比最多的两个轻度流量用户类型产生了最少的流量。如图 3-5 所示，聚类 0 及聚类 3 分别为周末轻度流量用户及工作日轻度流量用户，这两个聚类在用户中占比分别为 13%及 54%。但是观察图 3-5，这两种用户产生总流量分别为 2.4%及 3.3%，仅占全网用户使用总流量的 5.7%。这进一步验证了大部分用户产生少了流量的现象，即长尾效应。

对用户使用流量行为时域特征的分析，有助于运营商了解每类用户产生流量的峰时，对应的提供更好的资源分配策略。将聚类 0 与聚类 3 中的用户定义为轻度流量用户，其余四类为重度流量用户。图 3-6 与图 3-7 分别为重度流量用户与轻度流量用户日均使用流量 CDF 图，可以发现 80%的轻度流量用户日均使用流量在 1 以下（对流量取对数之后的值），相对应的，80%的重度流量用户日均在 10 以下。可以发现，聚类分析很好地

区分了重度流量用户与轻度流量用户，方便进一步地研究两类用户的行为，更好地制定个性化的流量套餐业务。文献[51]研究了基于用户流量预测结果的用户业务套餐更新策略。

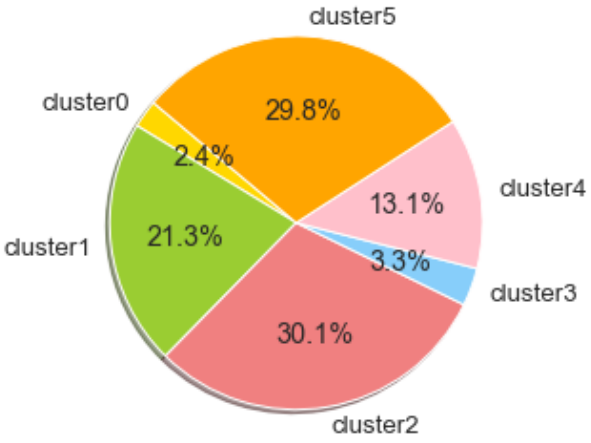


图 3-5 2018 年 9 月六种聚类用户使用总流量饼图

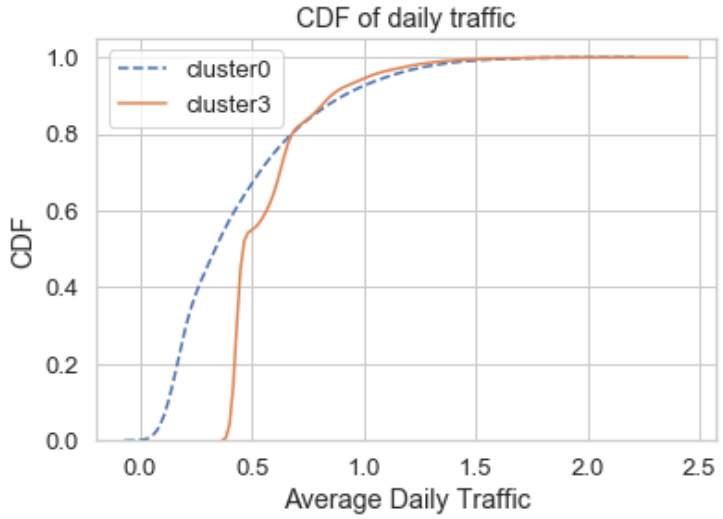


图 3-6 2018 年 9 月轻度流量用户日均流量 CDF 图

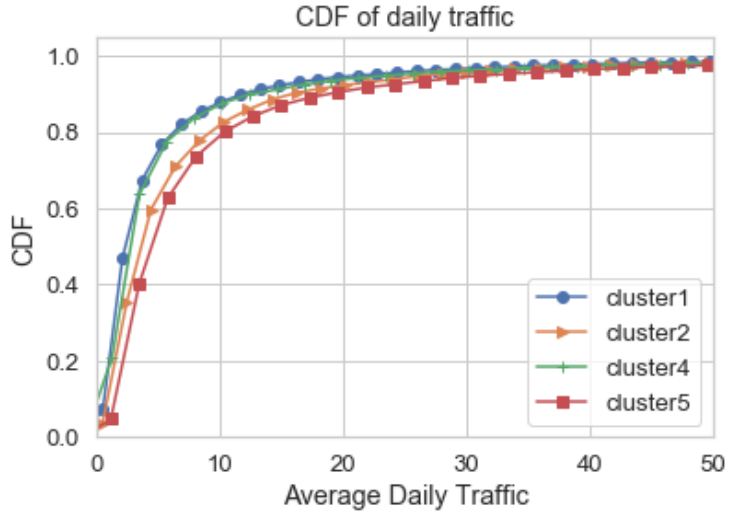


图 3-7 2018 年 9 月重度流量用户日均流量 CDF 图

3.3.3 性能分析及比较

因子分析法将原始特征向量维度从 168 维减小到 5 维，大大降低了特征维度。在使用因子分析法对特征向量降维后，再运用 k 均值聚类算法，可大大提高算法运行速度，同时算法效果也有所提升。如下表所是，比较基于因子分析法的聚类算法与其它方法的性能差异。可以看出，因子分析法降维后，聚类算法运行速度比不降维直接运用聚类算法快 7-8 倍。此外，基于因子分析法的聚类效果也更好，可以看出，与基于时间序列的聚类算法相比，簇用户比例更加均匀，对噪声异常点鲁棒性更强。

表 3-3 不同聚类方法性能比较

方法	运行时间	DBI 指数	簇用户比例
基于因子分析法的 k 均值聚类	4.350 秒	1.1348	13%,9%,10%,54%,5%,9%
基于时间序列的 k 均值聚类	37.174 秒	3.8979	2.1%,19%,2.2%,74%,1.3%,1.4%
基于主元分析法的 k 均值聚类	6.181 秒	1.1507	56%,8%,8%,4%,19%,5%

3.4 本章小结

本章主要对单个用户流量时间序列的特征进行了提取及压缩，并对用户进行了聚类分析，根据其使用流量时域特征，挖掘用户之间的相似性，将用户分成六种类型。

首先本章提取出用户使用流量时域特征，以一小时为时间尺度，分析用户一周的流量使用特征，得到 168 维原始特征数据。特征数过大会使得计算量加大，同时还会有特征的冗余，影响算法模型的性能。

因此，应用因子分析法对原始特征向量进行特征提取并降维，提取出五个公共因子。这些公共因子分别代表五个特定时间段。公共因子一，主要与工作时间的用户流量使用行为强烈相关，主要时段包括工作日的早七点至晚七点。与公共因子二相关的时隙则主要集中在午夜至凌晨时段，从晚上十一点至次日凌晨三点。公共因子三相关的时段主要在早间，从早上三点至六点。公共因子四包含的时段主要是周末的白天休息时段。公共因子五主要是晚间，从晚上七点至午夜。

最后本章应用 k 均值聚类算法对因子分析法降维后的特征向量进行了聚类分析，将用户分为六种不同类型，分别为：工作日重度流量用户，周末轻度流量用户，早间重度流量用户，晚间重度流量用户，工作日轻度流量用户，周末重度流量用户。注意轻度流量用户占比总和达到 60%，这与上一章分析中发现的长尾效应相吻合，进一步说明了大部分用户产生流量很少这个现象。

本章的研究旨在为运营商的资源配置、智能计费方案以及定制套餐业务提供新的思路。

第四章 用户应用软件使用特征提取与聚类

4.1 引言

应用指移动用户利用无线网络传输得到的数据根据不同的用户需求进行的具体操作，例如在线视频、社交软件等。用户应用软件行为蕴含非常丰富的信息，对刻画用户画像有很大的帮助。文献[52]研究了四万名用户的应用软件使用数据，并将研究范围缩小到使用最频繁的 500 种应用软件中。发现 99%以上的用户使用不同的应用软件集合，即用户在使用应用软件行为习惯上有极大的差异性。文献[53]从用户分类的角度研究应用软件数据。分析了 10 万名安卓用户为期一个月的应用软件使用记录，根据他们的历史行为，从中发现了 382 种不同用户类型。数据还包含用户其它个人信息，例如年龄收入等，以此建立更加详细的用户个人画像，并结合手机应用软件使用，为他们定义描述性的标签，例如，夜晚通信者，晚间学习者，年轻父母，爱车人士等。最后根据研究结果，给学者、手机设备商、网络运营商及手机应用软件开发商提供了启发性意见。文献[54]中应用机器学习分支自然语言处理中的主体模型（Topic model），基于无线网络用户业务流量行为中的访问网站记录，对用户进行聚类。然后再结合聚类结果与用户其他个人信息，如年龄收入等，挖掘聚类结果中隐藏的实际意义。

本章主要研究内容为用户使用手机应用软件行为分析。并基于用户行为对用户进行聚类，提取出不同类型的用户访问应用软件的模式，并观察用户访问应用软件与其使用数据流量之间的相关性。本章第二小节首先对手机应用软件按照提供的服务功能进行类别划分。应用自然语言处理领域内的相关技术与算法，建立自动分类模型。第三章节提取用户应用软件使用特征，再使用潜在语义分析进行特征降维，最后用 k 均值算法对用户进行聚类，得到六种不同类型的用户。针对每种类型的用户，分析其使用流量行为之间的差异。

4.2 应用软件类型划分

所用数据涵盖了上万种手机应用软件。由于应用软件数目较大，为了便于分析数据，首先将手机应用软件根据功能分成 26 类。分类如表 4-1 所示。

由于大部分手机应用软件仅通过名称就能判断其所属类型，因此首先根据关键词等信息对 4863 个应用软件进行手动分类，然后将手动分类结果作为训练数据，用机器学习方法，对剩余 5387 种应用软件进行自动分类。这属于机器学习领域的一个分支自然语言处理的研究课题，旨在对文本进行分类。分类问题属于机器学习里的有监督学习。与一般的分类问题不同，如何从文本里挖掘出有意义的特征是自然语言处理的重要环节。

4.2.1 文本特征提取

中文自然语言处理的第一步通常是分词。中文分词是中文自然语言处理的一个基础

性工作，指的是将一个汉字序列切分成一个个单独的词。本章节先调用 Python 库“结巴”对手机应用软件的名称进行中文分词。“结巴”的分词算法是基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 ，然后采用了动态规划查找最大概率路径，找出基于词频的最大切分组合。

经过分词后，对于每个手机应用软件，我们都得到一个列表，列表含有应用软件名称的词组，例如原手机应用名称“搜狗输入法小米版”经过分词后，会被切割成“搜狗”、“输入法”、“小米”、“版”等词，原手机应用名称“快狗打车车主（原 58 速运司机版）”经过分词后，被切割为“快狗”、“打车”、“车主”、“（”、“原”、“58”、“速运”、“司机”、“版”、“)”等词。可以看出分词算法效果很好，可以正确处理标点符号及数字等特殊符号，并合理地将有意义的词语分割提取出来。分词之后得到的词语列表构成了基本的文本信息。

表 4-1 手机应用软件分类

序号	类型	手机应用软件	序号	类型	手机应用软件
0	视频	爱奇艺，暴风影音	13	支付	支付宝
1	新闻	今日头条	14	教育	沪江英语
2	游戏	开心斗地主，炉石传说	15	金融	同花顺
3	社交	微信，陌陌，QQ	16	餐饮	美团，大众点评
4	地图	百度地图，高德地图	17	育儿	宝宝树
5	阅读	起点文学	18	天气	墨迹天气
6	办公	邮箱，印象笔记	19	时钟日历	万年历
7	音乐	网易云音乐，QQ 音乐	20	浏览搜索	搜狗，浏览器
8	电商	京东，淘宝	21	健康	丁香医生
9	应用商店	腾讯应用宝	22	美妆	美图秀秀
10	服务	电信营业厅	23	手机主题	主题
11	旅游	携程，去哪儿	24	系统工具	更新，垃圾清理大师
12	出行	嘀嘀打车，优步	25	其它	

4.2.2 文本特征降维

经过分词得到的词语列表非常广泛，即维度很高，因此往往需要应用特征降维方法来降低特征维度。潜在语义分析（Latent Semantic Analysis, LSA）[55]是一种常用的文本特征挖掘及降维的方法。因此，本章节应用潜在语义分析来进一步提取文本特征。

第一步，首先要得到文本信息的词频-逆文件频率（Term Frequency-Inverse Document Frequency, TF-IDF）矩阵。TF-IDF 是一种用于信息检索的常用加权技术，它定量分析了单个字词对于一个文件集合或语料库中的一份文件的重要性。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。即一个词语在一篇中出现次数越多, 同时在所有文档中出现次数越少, 越能够反映该的特征。

首先计算词频（Term Frequency, TF），最简单的选择是二元法，即倘若该词出现在

中，则词频为 1，否则为 0。计数法则是简单地计算某一个词出现在中的次数。其它的计算方法如表 4-2 所示。一般较为常用的词频方法是将该词出现的次数除以的总词数，进行归一化处理。

表 4-2 词频计算方法

加权算法	词频权重
二元法	0 ,1
计数法	$f_{t,d}$
词频	$f_{t,d} / \sum_{t' \in d} f_{t',d}$
对数分布	$\log(1+ f_{t,d})$

其次计算逆文档频率（IDF）。逆文档频率表征一个词能提供多少信息，也即它是否在所有文本中是独特的。例如中文中的词“的”，“我们”，这些词在文本中出现的频率很高，但是他们无法提供充分的信息，因为他们在所有文本中出现的次数都很高。逆文档频率计算公式如下所示

$$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{4.1}$$

其中 N 是文集里的数目总和，$N = |D|$，$|\{d \in D : t \in d\}|$ 是给有定词出现的总数。如果这个词没有出现在任何中，会造成 0 被除的情况，因此一般会在分母加上 1 调整，得到公式如下：

$$idf(t,D) = \log \frac{N}{1+|\{d \in D : t \in d\}|} \tag{4.2}$$

结合表 4-2 及公式 4.2，可得到 TF-IDF 计算公式如下：

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \tag{4.3}$$

如果某个词在一篇文档中出现的频率高，且在其它文档中出现的频率低，则表示该词具有很好的类别区分能力，则赋予更高权重，即其 TF-IDF 值更大。由于文集通常较大，含有很多的词，因此得到的 TF-IDF 矩阵通常维度很高，矩阵行数为文集里数目，矩阵列数为文集的总词数。在本章节中，矩阵行数为手机应用软件数目，矩阵列数即所有手机应用软件名称分词之后得到的词的总数。本章节，对手机应用软件数据求得的 TF-IDF 矩阵列数为 7179，维数非常大。此外，还应注意到原始特征矩阵即 TF-IDF 矩阵具有很强的稀疏性，因为每个手机应用软件名称实际上只由少数几个词组成。因此需要对 TF-IDF 矩阵进行降维处理，同时也是去噪过程。

潜在语义分析就是对 TF-IDF 矩阵进行奇异值分解（Singular Value Decomposition, SVD），从而降低特征矩阵维度，更方便后续的聚类处理。假设 $\mathbf{X}$ 是 TF-IDF 矩阵，其元素 $\mathbf{X}_{ij}$ 代表词语 i 在文档 j 中 TF-IDF 值，$\mathbf{X}$ 如下式子所示

$$\begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix}$$

可以看出，每一行代表一个词的向量，该向量描述了该词和所有文档的关系。相似地，每一列代表一个文档向量，该向量描述了该文档与所有词的关系。

两个词向量之间的点乘可以描述两个词之间的相似度，矩阵 $\mathbf{XX^T}$ 就体现了所有词之间的相似度。同样地，两个文档向量之间的点乘可以描述两个文档之间的相似度，矩阵 $\mathbf{X^TX}$ 反映了所有文档之间的相似度。线性代数中的奇异值分解，即存在正交阵 $\mathbf{U,V}$ 和对角阵 $\mathbf{\Sigma}$ 使得：

$$\mathbf{X = U\Sigma V^T} \tag{4.4}$$

其中 $\mathbf{X}$ 是 $m\times n$ 的矩阵，$\mathbf{U}$ 是 $m\times m$ 的正交阵，$\mathbf{V}$ 是 $n\times n$ 的正交阵，$\mathbf{\Sigma}$ 是 $m\times n$ 的对角阵。那么有：

$$\begin{aligned} \mathbf{XX^T} &= \mathbf{(U\Sigma V^T)(U\Sigma V^T)^T} \\ &= \mathbf{(U\Sigma V^T)(V\Sigma^T U^T)} \\ &= \mathbf{U\Sigma\Sigma^T U^T} \end{aligned} \tag{4.5}$$

$$\begin{aligned} \mathbf{X^TX} &= \mathbf{(U\Sigma V^T)^T(U\Sigma V^T)} \\ &= \mathbf{(V\Sigma^T U^T)(U\Sigma V^T)} \\ &= \mathbf{V\Sigma\Sigma^T V^T} \end{aligned} \tag{4.6}$$

因为 $\mathbf{\Sigma\Sigma^T}$ 与 $\mathbf{\Sigma^T\Sigma}$ 是对角矩阵，因此 $\mathbf{U}$ 是由 $\mathbf{XX^T}$ 的特征向量组成的矩阵，同理 $\mathbf{V}$ 是由 $\mathbf{X^TX}$ 的特征向量组成的矩阵。这些特征向量对应的特征值即为 $\mathbf{\Sigma\Sigma^T}$ 中的元素。$\mathbf{U}$ 中的每个特征向量被称为左奇异向量，$\mathbf{V}$ 中的每个特征向量被称为右奇异向量。$\mathbf{\Sigma}$ 被称为奇异值矩阵，其对角元素即为奇异值。奇异值矩阵中的奇异值是按从大到小排列，且奇异值下降很快。选取奇异值较大的几个值，其对应的奇异向量，得到对原始向量的近似。

4.2.3 分类算法结果分析

应用潜在语义分析对 TF-IDF 矩阵进行降维处理后，得到特征维度为 1000 的特征矩阵，用随机森林算法对数据进行分类。随机森林是一种基于决策树的集成算法，通过组合多个弱分类器，最终结果通过投票或取均值，使得整体模型的结果具有较高的精确度与泛化能力，避免出现过拟合现象。最终在验证集上随机森林算法准确率可以达到 79%。然后再把预测准确度较低、可靠性较差的样本分类结果列为其它类。

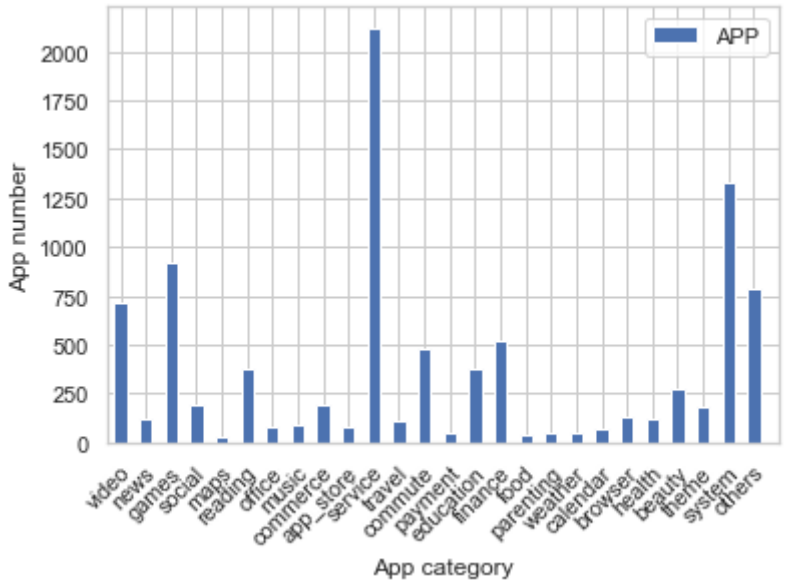


图 4-1 2018 年 9 月各类手机应用软件数目

分类后得到每一类手机应用软件内手机应用软件数目如上图 4-1 所示。可以看出服务类、游戏类、系统应用类应用软件数目较多，导航类、支付类、餐饮类以及母婴类应用软件数目较少。图 4-2 为每类手机应用软件使用用户数，可以看出服务类、社交类、浏览搜索类以及视频类使用的用户较多，母婴类、旅游类以及健康类手机应用软件使用用户较少，这是比较合理的，因为母婴类、旅游类以及健康类手机用户软件的目标用户群体较小。

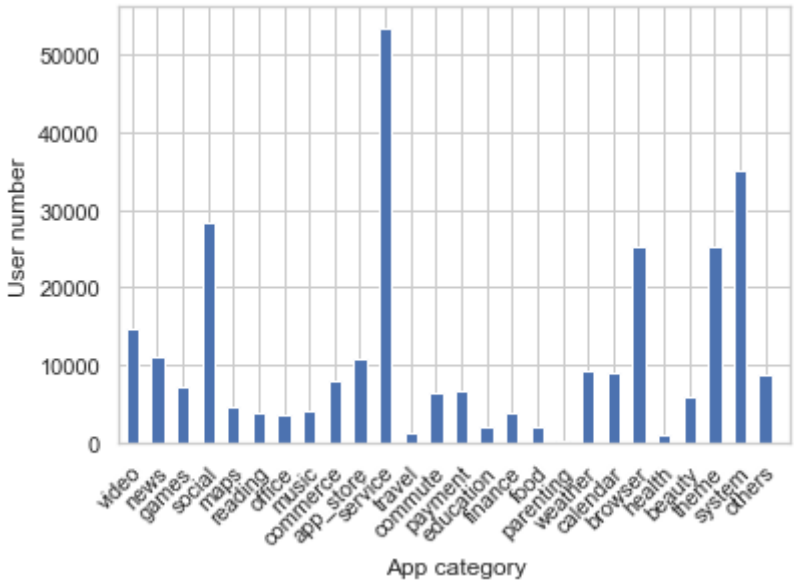


图 4-2 2018 年 9 月各类手机应用软件使用用户数

如图 4-3、图 4-4、图 4-5 分别为所有用户每一类手机应用软件使用总时长、使用总次数、以及消耗总流量。从使用时长来看，可以看出社交类、浏览搜索类、服务类、手机主题以及系统工具使用时长较长，母婴、旅行、健康类以及餐饮类手机应用使用时长较短。从使用总次数来看，服务类，浏览搜索类，手机主题及系统工具类使用次数较多，旅游，育儿，健康类手机应用软件使用次数较少。从消耗总流量来看，消耗流量最多的

几类手机应用是服务类，社交类，浏览类，视频及新闻类。消耗流量最少的手机应用类型是母婴类、日历、旅行以及健康类。

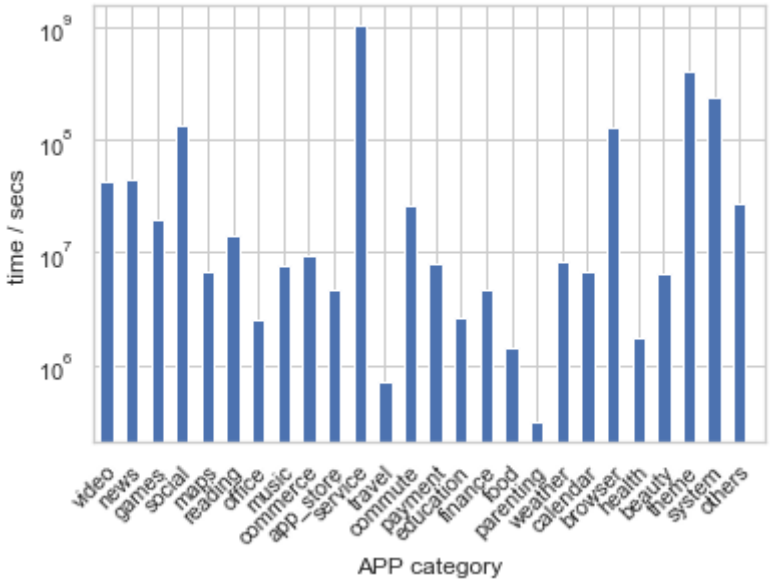


图 4-3 2018 年 9 月各类手机应用软件使用总时长

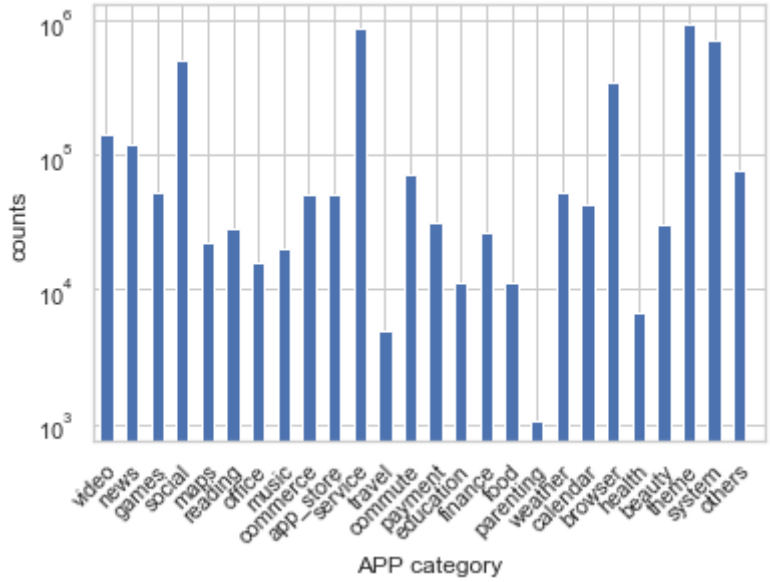


图 4-4 2018 年 9 月各类手机应用软件使用总次数

结合这三张图，可以看出社交类、游戏类、视频类及浏览搜索类，不论是使用时长，使用次数，还是使用流量，都属于较多的类型。注意到主题类、系统应用类虽使用次数及时长都较多，但产生流量并不多。这也符合常理，因为一般系统应用类手机软件较少产生流量。

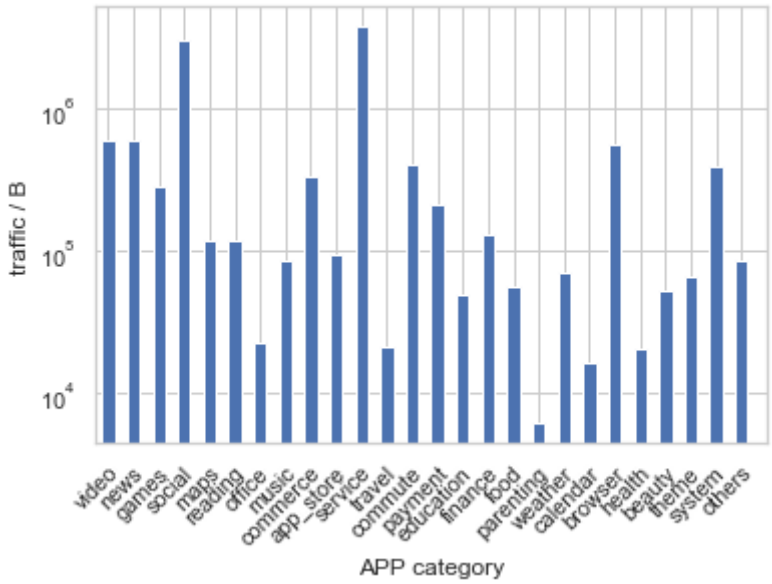


图 4-5 2018 年 9 月各类手机应用软件使用总流量

4.3 用户应用软件使用类型识别

经过上小节对应用软件类别的划分，我们将近万种手机应用软件划分为 26 类。本小节首先提取出用户使用手机应用软件的特征，然后应用聚类算法对用户进行聚类分析，观察不同类别用户使用手机应用软件的行为，最后进一步分析其使用流量的行为与手机应用软件之间的相关性。

4.3.1 特征提取

首先统计每个用户一天平均每小时使用各种类型手机应用软件的次数，得到 24*26=624 维的计数矩阵。正如之前所说，特征维度过大影响聚类算法的运行速度与效果，存在较大的冗余性以及噪声。因此，本小节继续应用上一章节提到的潜在语义分析对原始特征矩阵进行降维去噪处理。

根据计数矩阵计算相应的 TF-IDF 矩阵，其中词频在这里表示为某用户使用某一类手机应用软件的次数，而逆文档频率则表示为使用某类手机应用软件的用户数占所有用户数的比例。最后应用奇异值分解对得到的 TF-IDF 矩阵进行降维处理，最终得到维度为 50 的特征矩阵。

4.3.2 用户聚类

应用 k 均值聚类，对上述最终得到的特征矩阵进聚类分析。根据 DBI 指数，确定聚类簇数目为 6。得到簇用户比例如图 4-6 所示，被分到聚类 0，聚类 1，聚类 2，聚类 3，聚类 4，聚类 5 得用户比例分别为 7.7%，10.3%，26.1%，13.7%，22.2%，19.9%。

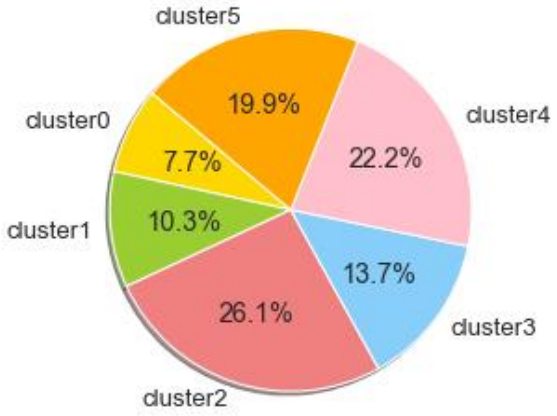


图 4-6 2018 年 9 月基于应用软件使用聚类用户比例

下面分析每个簇内用户使用手机应用软件的情况。对簇内用户求平均，得到每个类型用户一天平均每小时使用各类手机应用软件的次数，如图 4-7 所示，横轴表示一天 24 小时，纵轴表示共 26 类手机应用软件。

从图中可以很显然地看出，聚类 0 中的用户主要在早上时段使用手机应用软件，尤其集中在早上五点到七点，主要使用手机应用软件类型有视频类（0）、新闻类（1）、社交类（3）、服务类（10）、浏览搜索类（20）。

聚类 1 的用户主要使用手机应用软件集中在夜间，尤其在晚上八点至凌晨时段，使用手机应用软件类型与聚类 0 相似。

聚类 2 用户主要在日间使用手机应用软件，类型主要为系统工具类（24）、服务类（10）以及浏览搜索类（20）。聚类 2 用户使用应用软件类型较多，但整体次数较少，且系统工具类软件一般产生流量消耗较少。

聚类 3 用户使用手机应用软件主要集中在服务类，主要尖峰时刻在中午 12 点左右。

聚类 4 可以观察到明显的双尖峰现象，两个尖峰时刻分别在上午八点至十一点，下午五点至八点，主要手机应用软件类型也是服务类。

聚类 5 中用户主要使用手机应用软件类型为手机主题类（23）、系统工具类（24）以及社交类（3）。聚类 5 用户使用手机应用类型较多，且次数也较多。

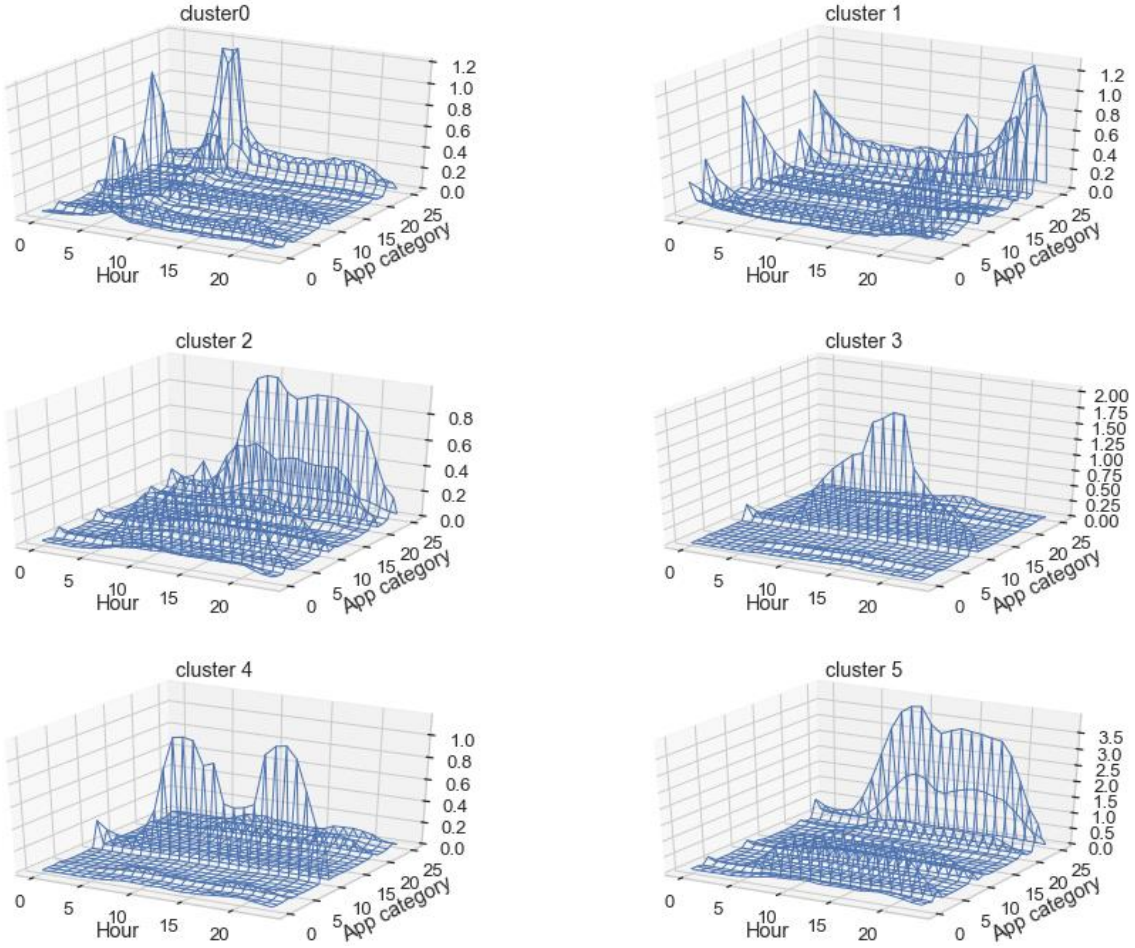


图 4-7 2018 年 9 月不同聚类用户使用应用软件使用次数

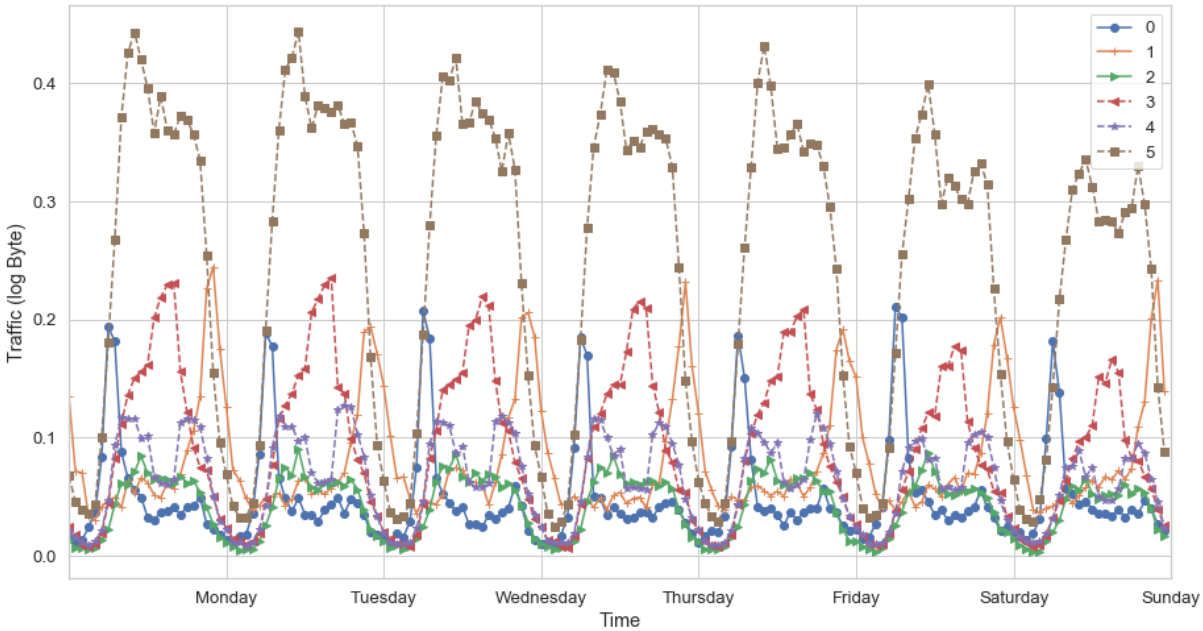


图 4-8 2018 年 9 月不同聚类用户使用流量时间分布

在分析用户使用手机应用软件行为之后，接下继续分析用户使用手机应用软件行为与用户使用流量之间的相关性。对每一个聚类，分别计算其簇内所有用户一周内平均每小时使用的流量，得到其使用流量模式如上图 4-8 所示。

下图 4-9 分别为每一个聚类用户平均使用流量的时域分布图。从图中可以看出，聚类 0 有一个明显的尖峰在早上时段，符合图 4-7 中用户主要在早间时段使用应用软件的行为。且聚类 0 中的用户使用流量较多。聚类 1 中用户使用流量的尖峰出现在晚间，属于晚间用户。聚类 2 中的用户使用流量很少。聚类 3 使用流量较多，且尖峰在白天时段。聚类 4 用户使用流量很少，呈现双尖峰特征。聚类 5 使用流量最多，且时间段持续最长。结合图 4-7，聚类 5 用户使用手机应用软件次数也是最多的。注意到聚类 2 及聚类 4 用户占比和达到近 50%，进一步说明了，大部分用户产生流量很少。聚类结果充分说明，用户不同的手机应用软件使用行为会直接影响其使用流量的行为。

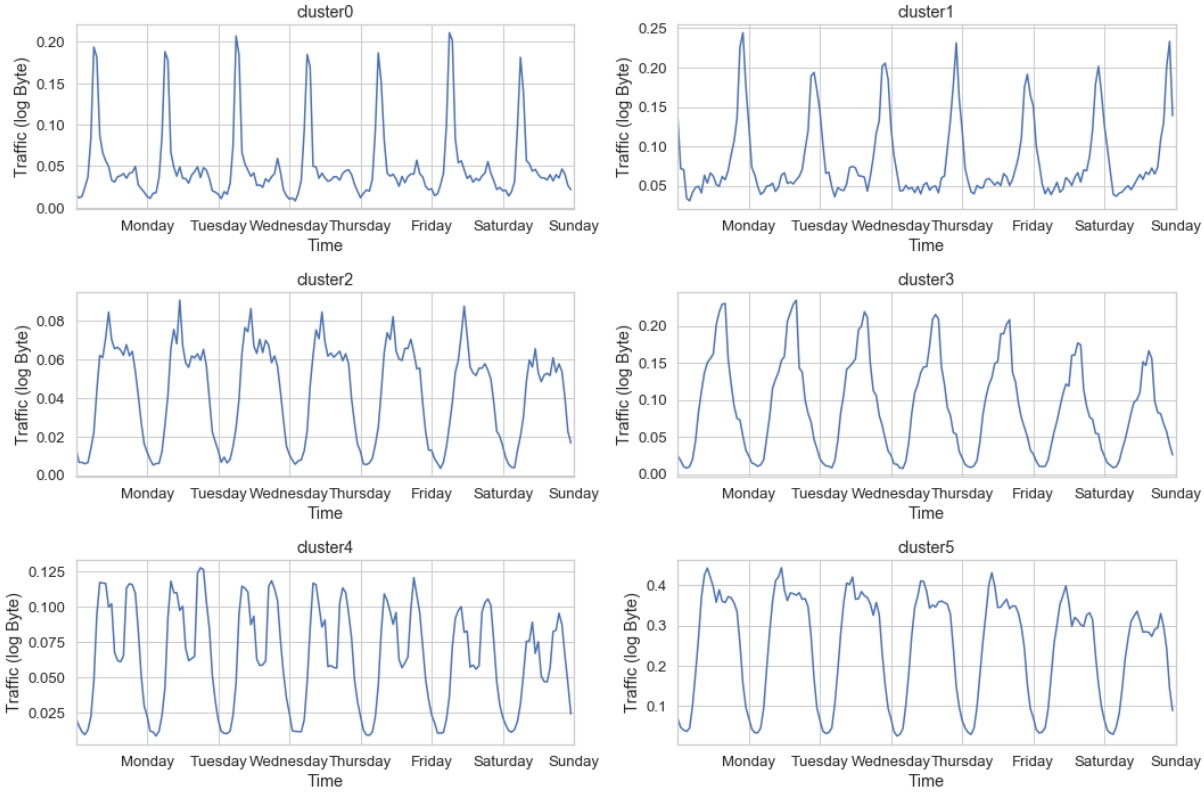


图 4-9 2018 年 9 月不同聚类用户使用流量时间分布

4.4 本章小结

本章主要对用户使用手机应用软件的行为进行了重点分析，并基于用户使用手机应用软件行为对用户进行聚类分析，将用户分为六种不同类型。然后继续分析了用户使用手机应用软件行为与用户使用流量行为之间的相关性。

首先本章对近万种手机应用软件进行了分类，采用了机器学习中的自然语言处理领域的相关算法。应用中文分词算法对应用名称分词，计算其 TF-IDF 矩阵，并应用潜在语义分析对特征矩阵进行降维处理。根据常识对最常用的部分应用软件进行人工标注，作为训练数据，然后应用随机森林分类算法对应用进行分类，最终将万种手机应用分为 26 类。在验证集上算法准确率可达到近 80%。

对应用软件进行分类后，计算用户每小时使用各类型应用软件的次数，同样根据计

数矩阵计算其 TF-IDF 矩阵，并应用潜在语义分析进行降维处理。然后用 k 均值对用户进行聚类分析，得到六种不同用户类型。分析不同类型用户使用手机应用软件行为，结合其使用流量的行为，结果发现，基于用户使用手机应用软件行为的聚类结果，和基于用户使用流量的聚类结果有相似性。用户使用流量的行为与其使用手机应用软件行为紧密联系。因此手机应用软件使用行为对预测用户使用流量很有帮助。

第五章 基于小波变换的 Prophet 与高斯过程用户流量预测

5.1 引言

用户的流量使用可以很自然的当作一个时间序列，从前面的分析也可以看出，用户的流量使用在时序上呈现一定的周期性，相关性和趋势性。预测个人用户使用流量，可以很自然地作为一个时间序列预测问题。其解决方案可分为两类，基于统计学习的方法，和基于机器学习的方法。基于统计学习，指的是根据流量的统计及概率分布特征，建模并预测用户流量。比如文献[14]中的 alpha-稳定模型，文献[56]中的泊松过程，文献[57]中的马尔可夫模型，文献[58][59]中使用的 ARIMA 模型。这些模型大多是线性模型。传统上，学者们大多应用 ARIMA 模型来预测基站的负载或者用户的流量，并且往往将 ARIMA 模型作为基本算法。ARIMA 的应用可参见文献[58][59][60][61]。文献[50]从用户个人角度研究应用层面流量使用模式，提取出数种个人用户的流量模式。并利用用户访问手机应用软件数据集，预测单用户流量需求。应用基于小波分解的滑动平均 (Auto-Regressive Moving Average, ARMA)算法，来预测单用户流量需求。首先对个人用户流量使用时间序列进行离散小波变换，得到多个子序列，再应用 ARMA 模型分别预测多个子序列，最后应用离散小波逆变换，得到最终的预测结果。在实际数据上应用算法，结果验证比基准算法的预测精确度提升了 7 到 8 倍。文献[62]结合传统时间序列预测模型 ARMA 与机器学习中的聚类算法，合理引入聚类得到的用户簇中心，从而提升 ARMA 模型预测准确度。

然而，实际应用中线性模型往往不能很好地建模流量业务。随着大量网络数据的获取，以及机器学习方法的发展进步，基于机器学习的预测模型使用越来越广泛。常用的模型如线性回归，支持向量机、高斯过程回归[26][27]等模型被用来预测流量。文献[35]对无线网格骨干网络的流量预测进行了研究，提出一种基于小波分解的深度信念网络结合高斯过程的预测模型。该文献首先对网络流量进行离散小波变换，提取出其中的高频分量与低频分量。其中高频分量显示了网络流量的突发特征与无规律的波动性，而低频分量则显示了网络流量时域上的长期依赖。因此，文献作者提出用深度信念网络建模及预测低频成分，而用高斯过程去建模及预测高频成分，最后结合两个模型的预测结果，重构出网络流量的预测。文献[63]对比研究三种机器学习方法预测基站负载的模型性能，发现支持向量机模型在预测多维度的网络负载时效果比多层感知机及系数衰减的多层感知机好，而系数衰减的多层感知机在预测单维度的网络负载时效果最好。

本章提出一种基于小波分解的 Prophet 与高斯过程回归模型的用户流量预测算法。针对用户网络流量时间序列的非平稳性、时变性等复杂特性，采用小波变换对用户网络流量时间序列进行预处理分析。经过小波变换后得到高频子序列与低频子序列，其中高频子序列反映了时间序列的突变性与无规律的波动性特征，而低频子序列则反映了时间序列的周期性与长程依赖性（Long-range Dependence, LRD）。针对高频子序列与低频子

序列的特点，应用 Prophet 模型预测低频子序列，用高斯过程回归模型预测高频子序列。最后再进行离散小波逆变换，重构得到最终的网络流量预测结果。

5.2 基于小波变换的 Prophet 与高斯过程预测算法

定义用户 i 使用流量的时间序列为 $\boldsymbol{x}^i=\left[x_1^i,x_2^i,\ldots,x_t^i\right]$，其中 $x_t^i=\log\left(1+c_t^i\right)$，$c_t^i$ 为用户在时隙 t 中产生的流量消耗。时间序列预测的目标是预测该时间序列下一时隙的值，即用户 i 在 $t+1$ 时隙产生的流量消耗 x_{t+1}^i。

针对个人用户流量非平稳不连续等特点，本章提出一种基于小波变换（Discrete Wavelet Transform, DWT）的 Prophet 与高斯过程回归模型预测算法。经过小波变换后，可以得到时间序列的低频子序列与高频子序列，其中低频子序列描述了时间序列的周期性与长程依赖性，而高频子序列则反映了时间序列的突变性与无规律的波动性特征。针对高低频子序列的各自特点，分别应用 Prophet 模型与高斯过程回归模型预测低频子序列与高频子序列。最后将高低频子序列预测值通过离散小波逆变换（Inverse Discrete Wavelet Transform, IDWT），得到最终的用户流量预测结果。算法流程如图 5-1 所示。

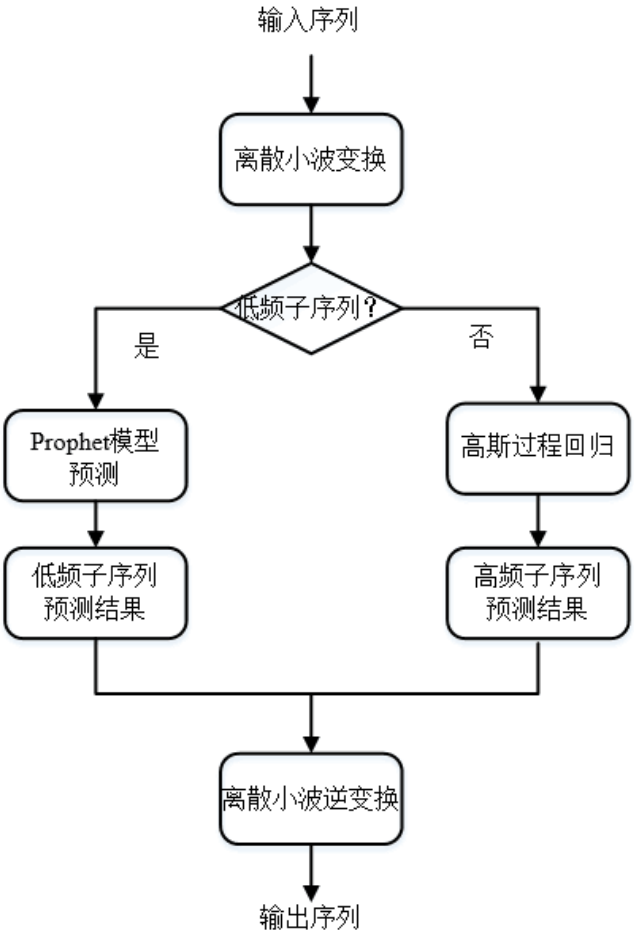


图 5-1 基于小波变换的 Prophet 与高斯过程预测算法流程图

下面首先介绍小波变换的原理，然后分别深入探讨 Prophet 模型与高斯过程回归模型，最后列出基于小波变换的 Prophet 与高斯过程预测算法的具体过程。

5.2.1 基于小波变换的时间序列分解

个人用户流量使用时间序列 $x_i(t)=\boldsymbol{x}^i$ 可由尺度函数 $\varphi(t)$ 与小波函数 $\psi(t)$ 进行级数展开，如下所示：

$$x_i(t)=\sum_{n=1}^{\frac{N}{2}}c_i(n)\varphi\left(\frac{t}{2}-n\right)+\sum_{n=1}^{\frac{N}{2}}d_i(n)\psi\left(\frac{t}{2}-n\right) \tag{5.14}$$

其中 $c_i(n)$ 通常称为近似系数或尺度系数，$d_i(n)$ 称为细节系数或小波系数。展开系数计算如下：

$$c_i(n)=\frac{1}{\sqrt{2}}\sum_{t=1}^{N}x_i(t)\varphi\left(\frac{t}{2}-n\right) \tag{5.15}$$

$$d_i(n)=\frac{1}{\sqrt{2}}\sum_{t=1}^{N}x_i(t)\psi\left(\frac{t}{2}-n\right) \tag{5.16}$$

离散小波变换在信号处理领域一直都是使用一组带通滤波器将信号分解为不同频率分量，即将信号 $x[n]$ 送到带通滤波器 $g[n]$ 以及 $h[n]$ 中。如图 5-1 所示，信号通过低通滤波器之后，可以将输入信号的高频部分过滤掉而输出低频部分。高通滤波器则与之相反，滤掉低频部分而输出高频部分。对于许多信号，低频成分相当重要，它常常蕴含着信号的特征，而高频成分则给出信号的细节或差别。

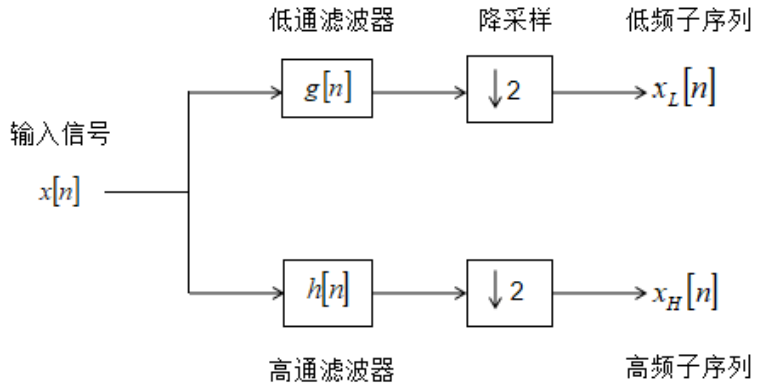


图 5-1 一维离散小波变换

与传统的傅里叶变换相比，小波变换尤其适用于非平稳的、不连续的且有尖峰的时间序列。傅里叶变换仅能提取出时间序列中的频率成分，但是看不出来信号频域随着时间变换的情况。而小波变换不仅可以知道时间序列中频率的成分，还可以知道这些频率出现在时域上的具体位置。即傅里叶变换只能得到一个频谱，而小波变换却可以得到一个时频谱。

5.2.2 基于 Prophet 的低频子序列预测

对于小波变换得到的低频子序列，应用 Prophet 模型进行预测。Prophet 模型由美国公司 Facebook 提出，该模型结合时间序列分解和机器学习的拟合，文献[64]简要地阐述了算法原理。

Prophet 采用时间序列加法分解模型，将低频子序列$c_i(n)$分解为三个主要分项：趋势，季节性，节假日。可用公式表示为：

$$c_i(n)=g(n)+s(n)+h(n)+\varepsilon_n \tag{5.1}$$

其中$g(n)$是趋势项，表示时间序列值非周期性的变化，$s(n)$代表时间序列值周期性变变化，$h(n)$代表特殊节假日对时间序列值的影响。错误项ε_n代表模型难以拟合的特殊变化，假设其服从正态分布。Prophet 算法分别拟合上述三项并求和得到最终的预测结果，模型中各项的参数由序列值$c_i(n)$决定。

首先对于趋势项，Prophet 采用非线性饱和增长模型和分段线性模型来拟合。非线性模型基于逻辑回归函数实现，其基本形式如下所示：

$$g(n)=\frac{B}{1+\exp\left(-k\left(n-m\right)\right)} \tag{5.2}$$

其中B为承载能力，指时间序列曲线的最大渐进值，例如总市场规模，总人口数等。通常这个值由市场规模的数据或者专业领域知识来决定。k表示曲线的增长速率，m为时间偏移量参数。这个函数实际上类似于人口增长函数，随着时间n的增加，$g(n)$越趋近于上限B，且k越大，增长速度就越快。

然而，在现实的时间序列中，曲线的走势不会一直保持不变，因此参数B,k,m不可能都是常数，而是随时间的迁移而改变的参数。因此，在 Prophet 算法中，这三个参数被定义为随时间而变化的函数，即$B=B(n),k=k(n),m=m(n)$。进一步地，为拟合真实数据复杂的时变特性，Prophet 模型引入了趋势变点的概念，在这些变点处，曲线的增长速率发生改变。假设存在S个变点分别发生在s_j时刻，$1\le j\le S$。定义调整量向量$\boldsymbol{\delta}\in R^S$，$\delta_j$表示在时刻$s_j$处速率的调整量，则任意时刻$n$的增长速率为基本速率$k$与此前所有变点处速率变化量之和，即$k+\sum_{j:t>s_j}\delta_j$。定义向量$\boldsymbol{a}(n)\in\{0,1\}^S$，其中

$$a_j(n)=\begin{cases}1, & \text{if } n\ge s_j\\ 0, & \text{otherwise}\end{cases} \tag{5.3}$$

那么，时刻n的增长速率为$k+\boldsymbol{a}(n)^T\boldsymbol{\delta}$。增长速率改变后，偏移量参数也随之改变，趋势变点$s_j$对时间偏移量参数的调整量的计算公式如下：

$$\gamma_j=\left(s_j-m-\sum_{l<j}y_l\right)\cdot\left(1-\frac{k+\sum_{l<j}\delta_l}{k+\sum_{l\le j}\delta_l}\right) \tag{5.4}$$

则$\boldsymbol{\gamma}=(\gamma_1,\gamma_2,\ldots,\gamma_S)^T$。得到分段非线性饱和增长模型如下：

$$g(n)=\frac{B(n)}{1+\exp\left(-\left(k+\boldsymbol{a}(n)^T\boldsymbol{\delta}\right)\left(n-\left(m+\boldsymbol{a}(n)^T\boldsymbol{\gamma}\right)\right)\right)} \tag{5.5}$$

时间序列通常会以天、周、月或年等季节性的变化而呈现周期性的变化。Prophet 算法用傅里叶级数来模拟时间序列的周期性。任意一个平滑的周期性曲线可以建模为：

$$s(n)=\boldsymbol{e}(n)\boldsymbol{\beta}=\sum_{l=1}^{L}\left(a_l\cos\left(\frac{2\pi nl}{P}\right)+b_l\sin\left(\frac{2\pi nl}{P}\right)\right) \tag{5.6}$$

其中 $\boldsymbol{e}(n)=\left[\cos\left(\frac{2\pi(1)n}{P}\right),\ldots,\sin\left(\frac{2\pi(L)n}{P}\right)\right]$，$P$ 代表目标序列的周期，$\boldsymbol{\beta}=[a_1,b_1,\cdots,a_L,b_L]^T$ 为模型要估计的参数，满足高斯分布。$2L$ 为设定的近似项个数，用于控制滤波程度。例如星期趋势，将变量 P 设置为 7，对应 L 取值通常为 3。

对于节假日及特殊事件的影响，往往是时间序列需要考虑的一个重点。因为节假日及特殊事件的发生会对时间序列原有的规律产生较大的影响。不同的节假日日期不同，影响事件跨度及力度都有所差异，因此 Prophet 将不同的节假日看成相互独立的模型。对于第 i 个节假日而言，D_i 表示该节假日产生影响的时间段。定义一个指示性函数，表示时刻 n 是否处于节假日 i 的影响时段内，并且为每个节假日设置一个参数 κ_i 来表示节假日的影响范围。假设存在 M 个节假日，节假日项可以拟合为：

$$h(n)=Z(n)\kappa=\sum_{i=1}^{M}\kappa_i\cdot\mathbf{1}_{\{n\in D_i\}} \tag{5.7}$$

其中，$Z(n)=\left[\mathbf{1}_{\{n\in D_1\}},\ldots,\mathbf{1}_{\{n\in D_M\}}\right]$，$\kappa_i\in N\left(0,\upsilon_i^2\right)$。

把周期项与节假日项结合到矩阵 $\mathbf{X}_c=\left\{\left[\boldsymbol{e}(n)\quad\boldsymbol{z}(n)\right]\right\}_{n=1}^{N/2}$ 中，将变点指示函数结合到矩阵 $\mathbf{A}=\left\{\boldsymbol{a}(n)\right\}_{n=1}^{N/2}$ 中。由于上述三项中模型参数先验概率均已给定，因此 Prophet 模型使用最大后验概率估计来获取参数。即

$$\boldsymbol{\lambda}^{MAP}=\arg\min\left(-\log p\left(\boldsymbol{c}_i\mid\mathbf{X},\boldsymbol{\lambda}\right)-\log p\left(\boldsymbol{\lambda}\right)\right) \tag{5.8}$$

其中，$\boldsymbol{c}_i=\left\{c_i(n)\right\}_{n=1}^{N/2}$ 是低频子序列，$\boldsymbol{\lambda}=\left(k,m,\boldsymbol{\delta},\boldsymbol{\beta},\boldsymbol{\kappa}\right)$，$p\left(\boldsymbol{c}_i\mid\mathbf{X},\boldsymbol{\lambda}\right)=N\left(\boldsymbol{\mu},\boldsymbol{\varepsilon}\right)$，且

$$\boldsymbol{\mu}=\frac{B}{\left(1+\exp\left(-\left(k+\mathbf{A}\boldsymbol{\delta}\right)\cdot\left(n-\left(m+\mathbf{A}\boldsymbol{\gamma}\right)\right)\right)\right)}+\mathbf{X}_c\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\kappa}\end{bmatrix} \tag{5.9}$$

得到参数后，预测低频子序列如下

$$\hat{c}_i\left(n+1\right)=g\left(n+1\right)+s\left(n+1\right)+h\left(n+1\right) \tag{5.10}$$

5.2.3 基于高斯过程的高频子序列预测

对于高频子序列，应用高斯过程回归算法进行预测。随着机器学习的发展，越来越多的机器学习算法也被用于时间序列的回归预测中。回归是机器学习中有监督学习的一个重要问题，其目标是学习一个模型，是模型能够对认定给定的输入，对其相应的输出做出一个良好的预测。回归模型的输入样例 $\boldsymbol{x}$ 的特征向量为 $\boldsymbol{x}=(x^{(1)},x^{(2)},\ldots,x^{(n)})^T$，其中

$x^{(i)}$表示$\boldsymbol{x}$的第i个特征。对应每一个输入样例，都有输出y。

下面简述高斯过程回归模型的算法原理[65]。首先由高频子序列构建训练数据输入样本集合$\mathbf{X}_d=\{\boldsymbol{x}_n\}_{n=1}^{N/2}$以及对应的输出样本集合$\boldsymbol{y}=\{d_i(1),d_i(2),\ldots,d_i(N/2)\}^T$，其中$\boldsymbol{x}_n=\left[d_i(n-3),d_i(n-2),d_i(n-1)\right]$为输入样本特征。

高斯过程回归模型可以表示为：

$$y_i=\mathrm{f}(\boldsymbol{x}_i)+\varepsilon \tag{5.11}$$

其中ε为与样本数据独立的高斯白噪声，服从均值为 0，方差为σ_n^2的高斯分布，记作$\varepsilon\sim N(0,\sigma_n^2)$。函数空间$\mathrm{f}(\boldsymbol{x}_1),\mathrm{f}(\boldsymbol{x}_2),\ldots,\mathrm{f}(\boldsymbol{x}_n)$构成随机变量的一个集合，且服从联合高斯分布。即

$$p(\boldsymbol{f}\mid\mathbf{X}_d)=N(\mathbf{0},\mathbf{K}) \tag{5.12}$$

$\mathbf{K}$为高斯分布的方差矩阵，其值可由方差函数确定。选取核函数为平方指数协方差函数（Squared exponential co-variance function），如下所示：

$$\mathbf{K}_{ij}=\mathrm{k}_{\mathrm{SE}}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\frac{1}{2\theta^2}(\boldsymbol{x}_i-\boldsymbol{x}_j)^T(\boldsymbol{x}_i-\boldsymbol{x}_j)\right) \tag{5.13}$$

θ为待估计参数，可用最大似然法获得最优超参数θ^{ML}，

$$\theta^{ML}=\operatorname{argmin}\left(-\log p(\boldsymbol{y}\mid\mathbf{X}_d,\theta)\right) \tag{5.14}$$

其中训练样本的对数似然函数$p(\boldsymbol{y}\mid\mathbf{X}_d,\theta)=N(\mathbf{0},\mathbf{K}_T)$，其中$\mathbf{K}_T=\mathbf{K}+\sigma_n^2\mathbf{I}$。

参数确定后，可通过后验概率函数$p(y_*\mid\boldsymbol{x}_*,\theta)$计算得到测试数据$\boldsymbol{x}_*=\left[d_i(n-2),d_i(n-1),d_i(n)\right]$的样本输出值，如下所示：

$$p(y_*\mid\boldsymbol{x}_*,\theta)=N(\mu_*,\sigma_*^2) \tag{5.16}$$

其中$\mu_*=\boldsymbol{k}_*\mathbf{K}_T^{-1}\boldsymbol{y}$为测试样本输出值的均值，其中$\boldsymbol{k}_*=\left(k(\boldsymbol{x}_*,\boldsymbol{x}_1),\ldots,k(\boldsymbol{x}_*,\boldsymbol{x}_{N/2})\right)^T$。一般选其作为测试样本输出值的估计值，即

$$\hat{d}_i(n+1)=\mu_*。 \tag{5.17}$$

5.2.4 流量预测

得到低频子序列与高频子序列得预测结果之后，应用离散小波逆变换即可得到用户使用流量最终预测结果。

如下对算法流程进行简述。

算法 5.1 基于小波变换的 Prophet 与高斯过程回归预测算法

输入：用户历史流量数据时间序列$x(t),t=1,2,\ldots,T$

输出：用户下一时刻使用流量$x(T+1)$

1. 对时间序列 $x(t)$ 进行小波变换，得到高频子序列 $h(t)$ 与低频子序列 $l(t)$：
2. $\left[h(t), l(t)\right] \leftarrow DWT\left(x(t)\right)$
3. 应用高斯过程回归模型预测高频子序列，得到 $\tilde{h}$
4. 应用 Prophet 预测低频子序列，得到 $\tilde{l}$
5. $x(T+1) \leftarrow IDWT\left(\tilde{h}, \tilde{l}\right)$

5.3 算法结果分析与讨论

5.3.1 算法性能评估指标

预测算法性能评估指标有很多种，下面主要介绍常用的几种评估指标，分别为均方误差（Mean Squared Error, MSE）、均方根误差（Root Mean Squared Error, RMSE）、平均绝对误差（Mean Absolute Error, MAE）以及平均绝对百分误差（Mean Absolute Percentage Error, MAPE）。

各个评估指标计算公式如下表 5-1，其中 y 为测试集的实际值，$\hat{y}$ 为回归模型在测试集上的预测值，测试集一共有 m 个样本点。MSE 为测试集的真实值与预测值之差的平方和取平均。RMSE 是 MSE 取平方根后的值。MAE 表示预测误差的大小，而 MAPE 反映了误差与真实值之间的比例。

表 5-1 常用预测算法评估指标

评估指标	计算公式
MSE	$\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$
RMSE	$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$
MAE	$\frac{1}{m}\sum_{i=1}^{m}\left\|y_i - \hat{y}_i\right\|$
MAPE	$\frac{100\%}{m}\sum_{i=1}^{m}\left\|\frac{y_i - \hat{y}_i}{y_i}\right\|$

5.3.2 算法性能测试方案

随机选取一个用户，如图 5-2 所示为其九月一周使用流量数据，注意已对流量值取对数。可以观察到用户使用流量存在较为明显的周期性，但突发性也较强。例如在 2018 年 9 月 3 日、9 月 5 日等日期用户几乎完全没有流量产生，这可能是个人生活原因导致的。图 5-3 为该用户在 2018 年 9 月使用流量时间序列自相关图与部分自相关图。从图中也可以看出，该用户使用流量呈现明显的以天为时间尺度的周期性。用户当前时刻使用流量的数值与前后七小时使用流量相关性较强，此外还与 24 小时、48 小时及 72 小时之前使用流量相关性强，均超过 0.5。

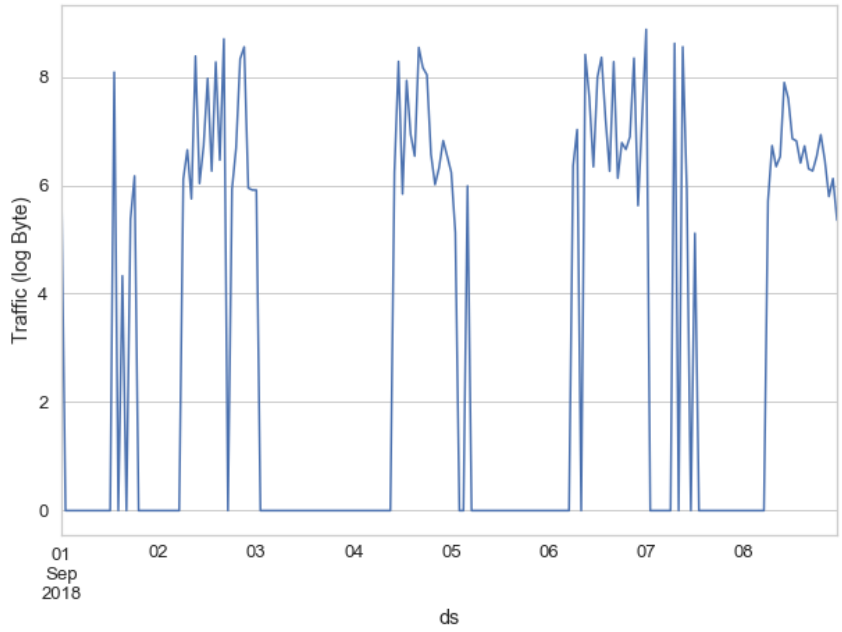


图 5-2 用户在 2018 年 9 月一周使用流量数据

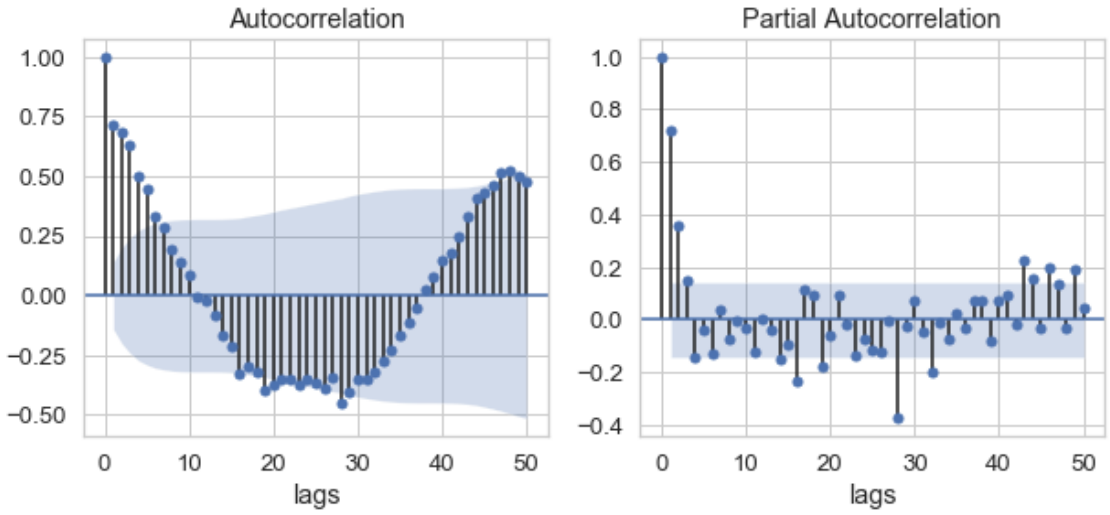


图 5-3 用户在 2018 年 9 月使用流量数据相关图与自相关图

根据以上分析，选取过去前 7 小时的历史数据，两天前同一时刻的历史数据及三天前同一时刻的历史数据共 9 维，作为数据特征。选取 2018 年 9 月 1 日至 9 月 7 日共为期一周的数据作为训练数据，预测 2018 年 9 月 8 日该用户每小时使用流量的情况。

首先对用户使用流量进行小波分解，选择阶数为 4 的 Daubechies 小波 DB4 作为小波基，分解层数为 1 层。得到高频子序列和低频子序列如下图 5-4 所示。观察可以看出，低频子序列呈现一定的周期性，而高频子序列则更显杂乱无序，随机性更强。

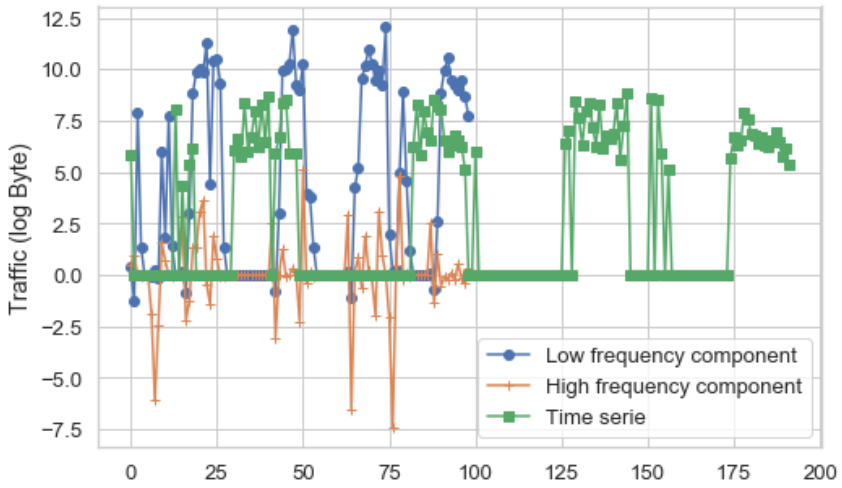


图 5-4 小波变换高频子序列与低频子序列

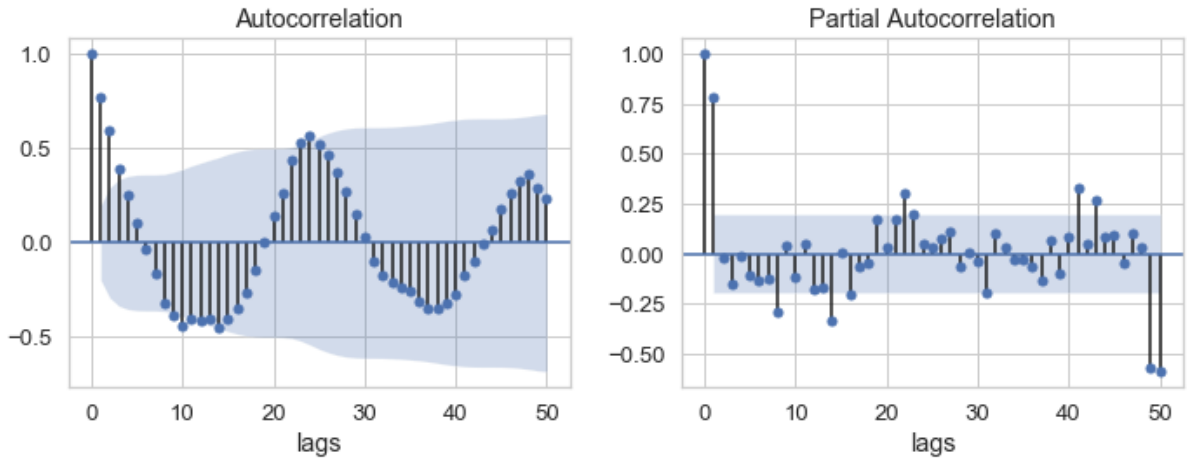


图 5-5 小波变换低频子序列自相关图与部分自相关图

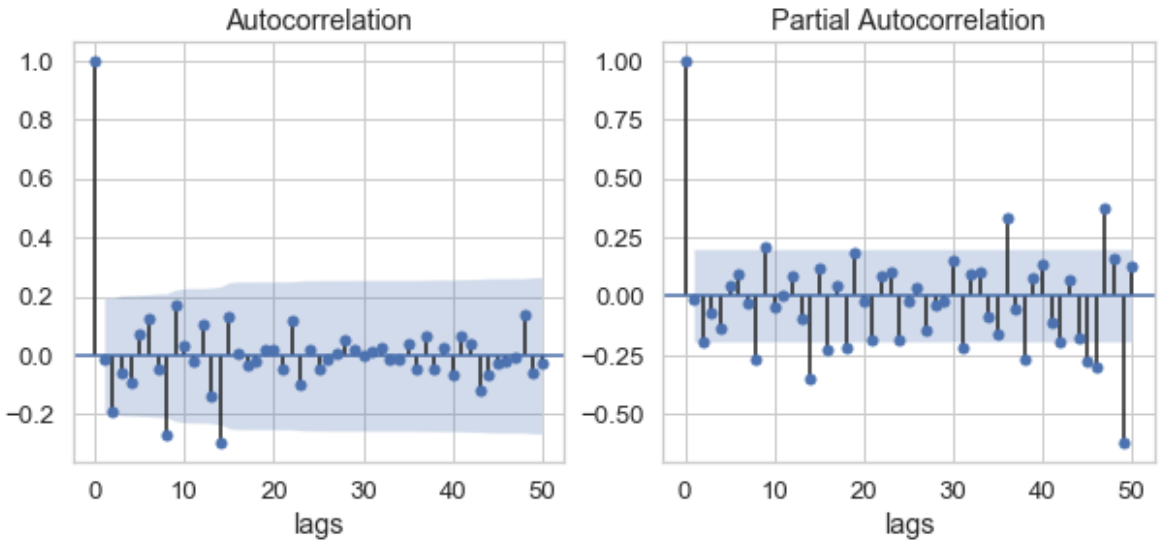


图 5-6 小波变换高频子序列自相关图与部分自相关图

分别观察低频子序列和高频子序列的自相关与部分自相关，如图 5-5 图 5-6 所示。低频子序列的自相关函数呈现以 24 为周期的周期性。而高频子序列则呈现近似随机噪声的特征。针对两个子序列的时域特征，对低频子序列，应用 Prophet 算法建模并预测。

而对于高频子序列，则应用高斯过程回归模型去预测。

预测结果如图 5-7 所示。可以看出预测结果很好。尤其注意到该算法能够有效地捕捉到用户突发性的流量需求，如 2018 年 9 月 7 日的上午时段。

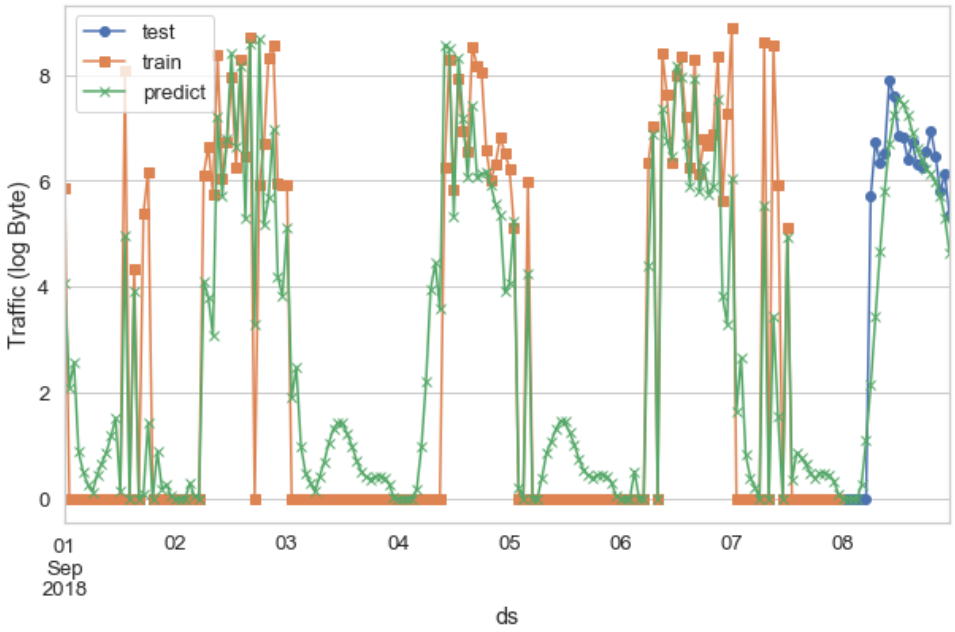


图 5-7 Prophet-高斯过程算法用户流量预测结果

文献[50]中提出基于小波变换的 ARIMA 模型预测用户流量，与提出算法性能进行比较，如下表 5-2 所示。可以看出本章提出的算法模型在预测性能上比基于小波分解的 ARIMA 算法更好，RMSE 是 ARIMA 算法的三分之一，MAPE 也降低了一半左右。此外，和 ARIMA 算法相比，本章提出的算法模型更加简单。ARIMA 模型需要手动选择参数，往往比较耗时，复杂度高，而 Prophet 和高斯过程都可以自动拟合参数，复杂度更低，运算速度更快。

表 5-2 不同预测算法性能比较

算法	RMSE	MAPE
基于小波变换的 Prophet-高斯过程回归	1.182	34.7%
基于小波变换的 ARIMA 预测模型	3.502	53.7%
ARIMA	3.741	65.6%

下面主要考虑两个参数对预测算法的性能影响：预测长度以及预测精度。如下图 5-8 所示，为不同预测长度对预测算法性能的影响。横轴表示预测未来一天到 6 天内每小时用户使用流量的情况。可以看出，本章提出的预测算法随着预测时间长度的增长性能逐渐趋于稳定，并且始终比 ARIMA 算法及基于小波变换的 ARIMA 算法好很多。观察可以发现，预测长度为 24 小时，本章提出的 Prophet 与高斯过程预测算法性能最好。

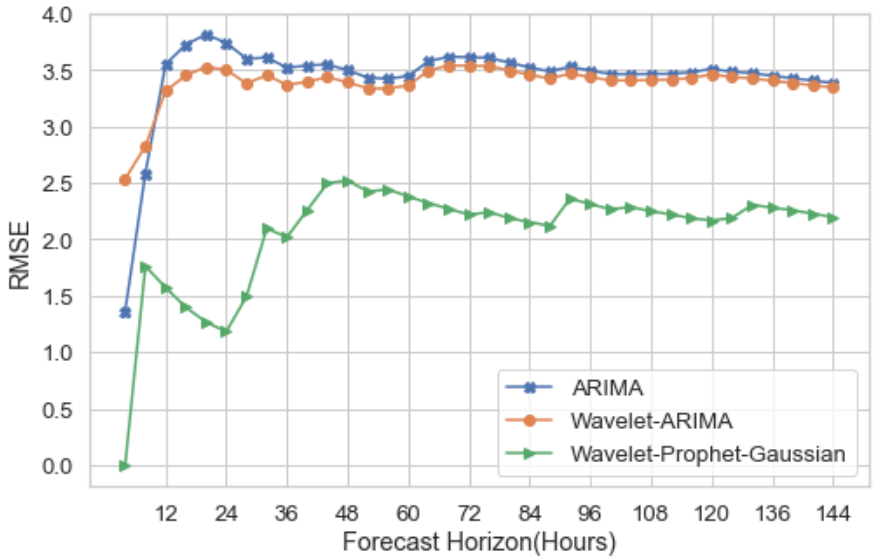


图 5-8 不同预测时间长度算法性能比较

下面考虑不同预测精度下算法的性能。固定训练数据长度为一周数据，测试数据时间长度为一天。分别预测用户未来一天每 15 分钟、每 30 分钟或者每 60 分钟使用流量值。如图 5-9、图 5-10 所示，为不同算法在不同预测精度下的预测 RMSE 与 MAPE。观察可以看出，随着预测精度要求的降低，预测性能越来越好。本章提出的基于小波变换的 Prophet 与高斯过程算法在所有预测精度下预测性能都最佳。在 15 分钟的预测精度下，RMSE 值也仅为 2.6，MAPE 为 59.2%，远远低于 ARIMA 算法以及基于小波变换的 ARIMA 算法。

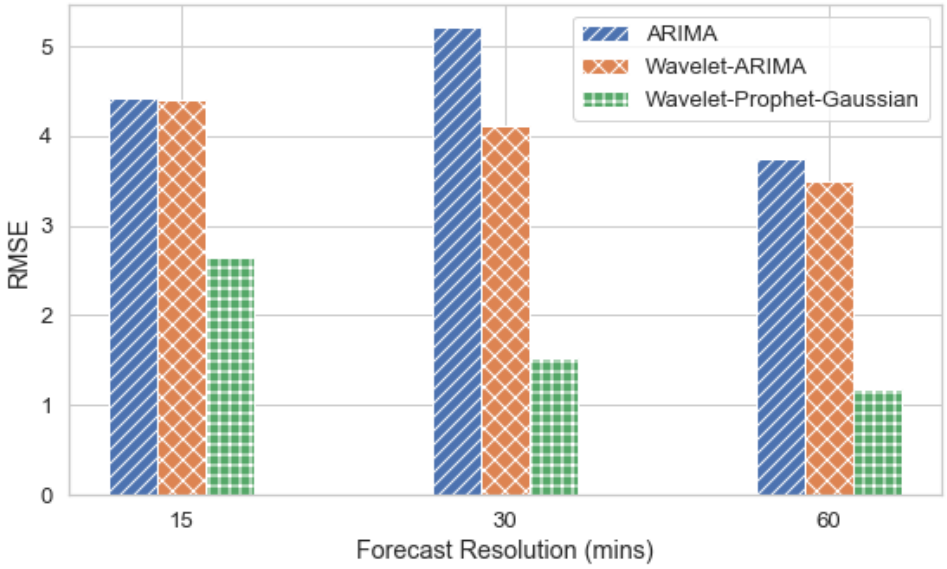


图 5-9 不同预测精度算法性能 RMSE 比较

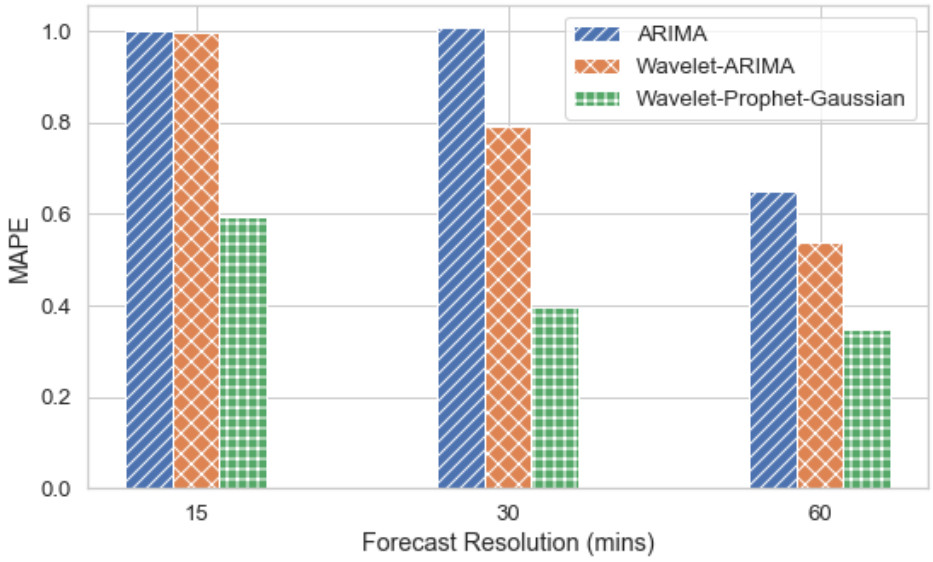


图 5-10 不同预测精度算法性能 MAPE 比较

随机选取一千名用户，应用本章提出的算法预测所有用户未来一天 24 小时使用流量值。得到 RMSE 的分布累积函数如图 5-11 所示。可以看出，将近 80%的用户预测 RMSE 都在 1 以下，说明该算法可以很好地适用于不同用户的流量预测。对比基于小波变换的 ARIMA，60%的用户预测 RMSE 在 1 以下。

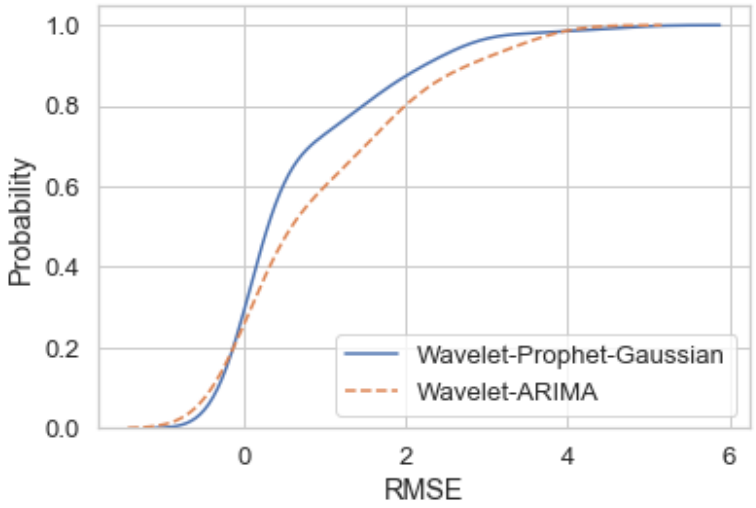


图 5-11 不同用户预测 RMSE 的 CDF 图

5.4 本章小结

本章针对个人用户流量使用时间序列的特点，提出一种基于小波变换的 Prophet 结合高斯过程回归预测算法。由于个人用户流量使用的突发性与不连续性，传统的时间序列预测方法不能很好地应用，因此本章首先对用户流量使用时间序列进行离散小波变换，分解得到低频子序列与高频子序列，再分别应用 Prophet 模型与高斯过程回归模型拟合并预测。

小波变换不仅可以知道时间序列中频率的成分，还可以知道这些频率出现在时域上

的具体位置。即小波变换可以得到一个时频谱。因此和传统的傅里叶变换相比，小波变换尤其适用于不连续的有尖峰的信号。离散时间序列通过小波变换后，得到高频子序列与低频子序列。其中高频子序列反映了时间序列的突变性与无规律的波动性特征，而低频子序列则反映了时间序列的周期性与长期依赖特性。对于高频子序列，应用高斯过程回归模型拟合并预测。对于低频子序列，应用 Prophet 模型预测。

随机选择一个用户，以其一周的流量使用时间序列作为训练数据，预测未来一天每小时产生的流量消耗，本章提出的预测算法预测误差在 34%。对比文献[50]中提出的基于小波分解的 ARIMA 算法，本章提出的算法预测误差降低了一半左右。随机选择一千名用户，应用 Prophet 与高斯过程回归算法，结果显示将近 80%的用户 RMSE 在 1 以下，对比基于小波变换的 ARIMA 算法，60%的用户 RMSE 在 1 以下。

第六章 总结与展望

6.1 总结

本文基于运营商的真实用户业务数据，分析用户业务流量数据时域特征，并挖掘用户应用软件使用行为与用户业务流量数据之间的相关性。最后提出基于 Prophet 与高斯过程回归的用户网络流量预测算法。

第二章对所用数据进行了深入详尽的分析。分别从全网用户与个人用户角度出发，对用户业务流量数据进行了分析与挖掘。首先应用 STL 分解方法对全网用户业务流量时间序列进行分解，观察其整体的趋势。应用时间序列自相关函数与部分自相关函数，刻画分析全网用户业务流量的强周期性。然后应用累计分布函数对个人用户流量分布进行分析，发现了用户流量分布的长尾效应，同时也发现用户间的流量使用时域分布模式存在着较大差异。最后继续分析用户移动性，发现用户移动性与用户流量使用行为之间的相关性。

第三章基于用户流量使用数据对用户进行聚类。首先应用因子分析法，对单个用户流量时间序列的特征进行了提取及压缩，并对用户进行了聚类分析，根据其使用流量时域特征，挖掘用户之间的相似性，将用户分成六种类型。不同类型的用户倾向于在每周的不同时段使用流量，且使用流量的数量大小也各异。对用户类型的挖掘分析，有助于运营商更好地了解用户业务流量需求，从而制定更加合理的计费策略或定制化业务流量套餐。

第四章主要对用户使用手机应用软件的行为进行了重点分析，应用潜在语义分析进行特征提取与降维处理。然后用 k 均值聚类基于用户使用手机应用软件行为对用户进行聚类分析，将用户分为六种不同类型。结合分析不同类型用户使用手机应用软件行为与其使用流量的行为，结果发现，用户使用流量的行为与其使用手机应用软件行为紧密联系。可以在预测用户流量是引入其手机应用软件行为的影响，有助于提高预测准确度。

第五章提出一种基于小波变换的 Prophet 与高斯过程用户网络流量预测算法。小波变换将用户流量时间序列分解为高频子序列与低频子序列。其中高频子序列反映了时间序列的突变性与无规律的波动性特征，而低频子序列则反映了时间序列的周期性与长程依赖性。针对高频子序列与低频子序列的特点，应用 Prophet 模型预测低频子序列，用高斯过程回归模型预测高频子序列。最后再进行离散小波逆变换，重构得到最终的网络流量预测结果。对比提出的算法与传统的时间序列预测算法，提出算法的预测效果得到了很大的提升，预测误差降低了一半左右。

6.2 展望

鉴于时间和个人能力有限，本文还有以下几点内容值得继续深入研究。

本文在第三章节对用户基于流量使用行为进行聚类分析，将用户分为不同类型。可

以进一步结合聚类结果，利用用户之间的相似性，提高用户流量预测准确度。

文章挖掘了用户移动性以及用户应用软件使用与用户流量使用之间的相关性，可以考虑更加深入地用数学模型来定量分析。另一方面，还可以考虑应用聚类结果到用户应用软件推荐系统中。

进一步地，基于个人用户使用流量预测结果对移动通信网络的资源分配策略优化与改进提供指导，例如研究基于用户流量预测结果的定制化业务套餐以及计费策略等。

致谢

首先在此由衷地感谢潘志文老师与刘楠老师的指导。在攻读硕士及撰写论文期间，我获得两位老师诸多的悉心指导与帮助。在每周的汇报工作会议上，潘志文老师和刘楠老师总是及时地为我指引研究方向，答疑解惑。除了专业知识的教授外，两位老师还教会我如何寻找课题，如何有效查找文献以及学会独立思考并提出问题。感谢他们的言传身教让我学习并热爱我所研究的领域，养成了一个硕士研究生应具备的科学研究素质。

同时，感谢我要感谢东南大学提供的学习环境与机会。在攻读硕士期间，我有幸获得了学院提供的出国学习的机会，赴法国学习先进的科学技术，增长见识。学校的图书馆藏书与数据库文献为我撰写论文提供了重要的学习资料。

最后，我要感谢学院各位老师在各项行政手续中的帮助。感谢我的师兄师姐师弟师妹们，与你们一起努力进步是我人生中重要的经历。感谢我的家人朋友，一直默默支持我的学业。

再次感谢所有人对我的帮助与鼓励。

参考文献

[1] Halepovic E, Williamson C. Characterizing and modeling user mobility in a cellular data network [C]. In: Acm International Workshop on Performance Evaluation of Wireless Ad Hoc (PEWA). Montreal, 2005. 71-78

[2] Paul U, Subramanian A P, Buddhikot M M, et al. Understanding traffic dynamics in cellular data networks [C]. In: 2011 IEEE INFOCOM, Shanghai, 2011. 882-890

[3] Douglass R W, Meyer D A, Ram M, et al. High resolution population estimates from telecommunications data [J]. EPJ Data Science, 2015, 4(1): 1-13

[4] Xu F L, Xia T, Cao H C, et al. Detecting popular temporal modes in population-scale unlabelled trajectory data [C]. In: ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Singapore, 2018. 1-25

[5] He G H, Chen J, Hou C, et al. Characterizing individual user behaviors in wlans [C]. In: the 10th ACM Symposium on Modeling, analysis, and simulation of wireless and mobile systems (MSWiM), Chania, 2007. 132-137

[6] Zhang Y. User mobility from the view of cellular data networks [C]. In: IEEE INFOCOM 2014 - IEEE Conference on Computer Communications IEEE, Toronto, 2014. 1348-1356

[7] Sun W, Miao D, Qin X, et al. Characterizing user mobility from the view of 4G cellular network [C]. In: 2016 17th IEEE International Conference on Mobile Data Management (MDM), Porto, 2016. 34-39

[8] Zhao Z, Zhang P, Huang H, et al. User mobility modeling based on mobile traffic data collected in real cellular networks [C]. In: 2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS), Surfers Paradise, 2017. 1-6

[9] Feng J, Li Y, Zhang C, et al. DeepMove: Predicting human mobility with attentional recurrent networks. [C]. In: 2018 International World Wide Web Conferences Steering Committee, Lyon, 2018. 1459-1468

[10] Xu Q, Erman J, Gerber A, et al. Identifying diverse usage behaviors of smartphone apps [C]. In: 2011 ACM SIGCOMM Internet Measurement Conference (IMC), Berlin, 2011. 329-344

[11] Liao Z X, Lei P R, Shen T J, et al. AppNow: Predicting usages of mobile applications on smart phones [C]. In: Conference on Technologies and Applications of Artificial Intelligence (TAAI), Tainan, 2012. 300-303

[12] Liao Z X, Lei P R, Shen T J, et al. Mining temporal profiles of mobile applications for usage prediction [C]. In: 2012 IEEE 12th International Conference on Data Mining Workshops, Brussels, 2012. 890-893

[13] Liao Z X, Li S C, Peng W C, et al. On the feature discovery for app usage prediction in

smartphones [C]. In: 2013 IEEE 13th International Conference on Data Mining, Dallas, 2013.1127-1132

[14] Li R, Zhao Z, Zheng J, et al. The learning and prediction of application-level traffic data in cellular networks [J]. IEEE Transactions on Wireless Communications, 2017, 16(6): 3899-3912

[15] Lu Z, Feng Y H, Zhou W J, et al. Inferring correlation between user mobility and app usage in massive coarse-grained data traces [C]. In: ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Singapore, 2018. 99-120

[16] Silva F A, Domingues A, Braga S. Discovering mobile application usage patterns from a large-scale dataset [J]. ACM Transactions. on Knowledge Discovery from Data, 2018, 12(5): 1-36

[17] Yang L, Yuan M X, Wang W, et al, Apps on the move: A fine-grained analysis of usage behavior of mobile apps [C]. In: IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications, San Francisco, 2016. 1-9

[18] Yu D H, Li Y, Xu F L, et al. Smartphone app usage prediction using points of interest [C]. In: ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Maui, 2018. 1-21

[19] Zeng M, Lin T H, Chen M, et al. Temporal-spatial mobile application usage understanding and popularity prediction for edge caching [J]. IEEE Wireless Communications, 2018, 25(3): 36-42

[20] Ricardo B Y, Di J, Fabrizio S. et al. Predicting the next app that you are going to use [C]. In: The Eighth ACM International Conference on Web Search and Data Mining (WSDM), Shanghai, 2015. 285-294

[21] Zhao X, Qiao Y, Si Z, et al. Prediction of user app usage behavior from geo-spatial data [C]. In: The Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data, San Francisco, 2016. 1-6

[22] 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

[23] Xu F, Li Y, Wang H, et al, Understanding mobile traffic patterns of large scale cellular towers in urban environment [J]. IEEE/ACM Transactions on Networking, 2017, 25(2): 1147-1161

[24] Shi H Z, Li Y. Discovering periodic patterns for large scale mobile traffic data: method and applications [J]. IEEE Transactions on Mobile Computing, 2018, 17(10): 2266-2278

[25] Naboulsi D, Fiore M, Ribot S, et al. Large-scale mobile traffic analysis: A survey [J]. IEEE Communications Surveys and Tutorials, 2016, 18(1): 124-161

[26] Le L V, Sinh D, Tung L P, et al. A practical model for traffic forecasting based on big data, machine-learning, and network KPIs [C]. In: 2018 15th IEEE Annual Consumer Communications and Networking Conference (CCNC), Las Vegas, 2018. 1-4

[27] Le L V, Sinh D, Lin B S, et al. Applying big data, machine learning, and SDN/NFV to 5G traffic clustering, forecasting, and management [C]. In: 2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft), Montreal 2018. 168-176

[28] Zhang Z, Liu F, Zeng Z, et al. A traffic prediction algorithm based on Bayesian spatio-temporal model in cellular network [C]. In: 2017 International Symposium on Wireless Communication Systems (ISWCS), Bologna, 2017. 43-48

[29] 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

[30] Trinh H D, Giupponi L, Paolo D, Mobile traffic prediction from raw data using LSTM networks [C]. In: IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Bologna, 2018. 1827-1832

[31] Wang J, Tang J, Xu Z, et al. Spatiotemporal modeling and prediction in cellular networks: A big data enabled deep learning approach [C]. In: IEEE INFOCOM 2017 - IEEE Conference on Computer Communications, Atlanta, 2017. 1-9

[32] Qiu C, Zhang Y, Feng Z, et al. Spatio-temporal wireless traffic prediction with recurrent neural network [J]. IEEE Wireless Communications Letters, 2018, 7(4): 554-557

[33] Huang C W, Chiang C T, Li Q. A study of deep learning networks on mobile traffic forecasting [C]. In: 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Montreal, 2017. 1-6

[34] Fang L, Cheng X, Wang H, et al. Mobile demand forecasting via deep graph-sequence spatiotemporal modeling in cellular networks [J]. IEEE Internet of Things Journal, 2018, 5(4):3091-3101

[35] Nie L, Jiang D, Yu S, et al. Network traffic prediction based on deep belief network in wireless mesh backbone networks [C]. In: 2017 IEEE Wireless Communications and Networking Conference (WCNC), San Francisco, 2017. 1-5

[36] Zhang S, Zhao S L, Yuan M X, et al. Traffic prediction based power saving in cellular networks: a machine learning method [C]. In: The 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL), Los Angeles, 2017. 256-277

[37] Dawoud S, Uzun A, Gondor S, et al. Optimizing the power consumption of mobile networks based on traffic prediction [C]. In: 2014 IEEE 38th Annual Computer Software and Applications Conference (COMPSAC), Vasteras, 2014. 279-288

[38] Yang J, Qiao Y, Zhang X, et al. Characterizing user behavior in mobile internet[J]. IEEE Transactions on Emerging Topics in Computing, 2015, 3(1): 95-106

[39] Jin Y, Duffield N, Gerber A, et al. Characterizing data usage patterns in a large cellular network [C]. In: 2012 ACM Workshop on Cellular Networks: Operations, Challenges, and Future Design, Helsinki, 2012. 7-12

[40] Cleveland R B, Cleveland W, McRae J E, et al. STL: A seasonal-trend decomposition procedure based on Loess [J]. Journal of Official Statistics, 1990, 6(1): 3-33

[41] Daubechies I. The wavelet transform, time-frequency localization and signal analysis [J]. IEEE Transactions on Information Theory, 1990, 36(5): 961-1005

[42] Mulaik S A. Foundations of factor analysis [M]. CRC Press, 2009

[43] Cerny C A, Kaiser H F. A study of a measure of sampling adequacy for factor-analytic correlation matrices [J]. Multivariate Behavioral Research, 1977, 12(1), 43-47

[44] Jolliffe I T. Principal Component Analysis [M]. Springer, 2010

[45] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning [M]. Springer, 2009. 106-119

[46] Kaiser H. An index of factor simplicity [J]. Psychometrika, 1974, 39: 31–36

[47] Hartigan J A, Wong M A. A k-means clustering algorithm [J]. Applied Statistic, 1979, 28(1):100-108

[48] Davies D L, Bouldin D W. A cluster separation measure [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, 1(2): 224-227

[49] Furno A, Fiore M, Stanica R. Joint spatial and temporal classification of mobile traffic demands [C]. In: IEEE INFOCOM 2017 - IEEE Conference on Computer Communications, Atlanta, 2017, 1-9

[50] Wu J, Zeng M, Chen X, et al. Characterizing and predicting individual traffic usage of mobile application in cellular network [C]. In: the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp18), New York, 2018. 852-861

[51] Lai Y, Wu Y, Yu C, et al. Mobile data usage prediction system and method [C]. In: 2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA), Taipei, 2017. 484-486

[52] Welke P, Andone I, Blaszkiewicz K, et al. Differentiating smartphone users by app usage [C]. In: 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Heidelberg, 2016. 519-5251

[53] Zhao S, Ramos J, Tao J, et al. Discovering different kinds of smartphone users through their application usage behaviors [C]. In: 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Heidelberg, 2016. 498-509

[54] B. Leng, J. Liu, H. Pan, S. Zhou and Z. N. Tsinghua, Topic model based behavior modeling and clustering analysis for wireless network users [C]. In: 2015 21st Asia-Pacific Conference on Communications (APCC), Kyoto, 2015. 410-415

[55] Christopher D. M, Raghavan P, Schütze H, et al. Introduction to information retrieval [M], Cambridge University Press, 2008

[56] Yang J, Li W, Qiao Y, et al. Characterizing and modeling of large-scale traffic in mobile

network [C]. In: 2015 IEEE Wireless Communications and Networking Conference (WCNC), New Orleans, 2015. 801-806

[57] Shafiq M Z, Ji L S, Liu A X, et al. Characterizing and modeling internet traffic dynamics of cellular devices [C]. In: ACM Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS), San Jose, 2011. 305-316

[58] Shu Y T, Yu M F, Liu J K, et al. Wireless traffic modeling and prediction using seasonal ARIMA models [C]. In: IEEE International Conference on Communications (ICC), Anchorage, 2003. 1675-1679

[59] Zhou B, He D, Sun Z. Traffic predictability based on ARIMA/GARCH model [C]. In: 2006 2nd Conference on Next Generation Internet Design and Engineering (NGI), Valencia, 2006. 207-220

[60] Guo J, Peng Y., Peng X Y, et al, Traffic forecasting for mobile networks with multiplicative seasonal ARIMA models [C]. In: 2009 9th International Conference on Electronic Measurement and Instruments, Beijing, 2009, 377-380

[61] Li J, Shen L, Tong Y. Prediction of Network Flow Based on Wavelet Analysis and ARIMA Model [C]. In: 2009 International Conference on Wireless Networks and Information Systems, Shanghai, 2009. 217-220

[62] Aldhyani T, Joshi M R. Integration of time series models with soft clustering to enhance network traffic forecasting [C]. In: 2016 Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata, 2016. 212-214

[63] Nikravesh A Y, Ajila S A, Lung C H, et al. Mobile network traffic prediction using MLP, MLPWD, and SVM [C]. In: IEEE International Congress on Big Data, San Francisco, 2016:402-409

[64] Taylor S J, Benjamin L. Forecasting at Scale [J]. The American Statistician, 2017, 27-45

[65] Rasmussen C E, Williams C. Gaussian processes for machine Learning, MIT Press 2006

攻读硕士学位期间的主要成果

- **攻读硕士学位期间发表的论文**

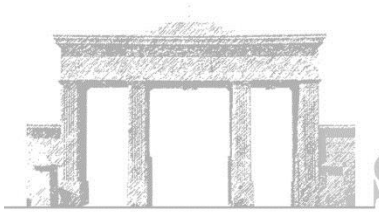
[1] 李玉. 中继无线网络分布式空时码与相关合并技术比较性研究. 第 33 届南京地区研究生通信年会. 2018

[2] Li Yu, Ma Ziang, Pan Zhiwen, Liu Nan, You Xiaohu. Prophet Model and Gaussian Process Regression Based User Traffic Prediction in Wireless Networks, submitted to SCIENCE CHINA-Information Sciences

- **攻读硕士学位期间申请的专利**

[1] 潘志文，李玉，刘楠，尤肖虎。基于小波变换的 Prophet 与高斯过程网络流量预测方法，201910427803.8，2019.5.22

心求至善



SOUTHEAST UNIVERSITY