**基于拓扑结构的轨道交通人流量预测**

Prediction of Human Flow in Rail Transit Based on Topology

（通过训练得到一个预测模型，用来预测任何一个站点任何一个时段的客流量）

王天羿

重庆市南岸区崇文路2号（重庆邮电大学）重庆 400065

E-mail: 932363572@qq.com

联系电话: 18883285019

**摘要：**网络化运营时代的到来要求在满足大客运量的运输安全和运营效率的前提下，逐渐提升轨道交通客运技术，实现以人为本的客运服务理念。然而，由于客运需求与客运能力供给的分布不均衡性，常常会在城市轨道交通网络上造成难以预见的网络客流拥塞状况，给城市居民出行带来极大的困扰。基于此，本文提出一种结合轨道交通拓扑结构和提升树算法的流量预测方法。通过分析轨道交通网络的拓扑结构，提出一种轨道交通网络断面客流量预测方法，再结合提升树算法（Boosting Tree）预测各站点短时客流量。

**关键字：**城市轨道交通，拓扑，提升树，断面客流量

# 1.引言

城市规模的快速发展、路网总量的不断提升、机动车保有量快速增长，交通问题已经成为城市发展和管理最重要的问题之一。随着道路通行能力的有限供给相对于交通需求总量快速增长的落后差距拉大，交通需求与交通供给之间矛盾日益凸显，交通拥堵问题已经成为城市人非常关心的问题之一。而且，随着生活素质不断提升，人们对城市交通的安全、快捷、便利的需求越来越高，快捷的交通路网、方便的交通信息服务已经成为大家日益关心和谈论的话题。不仅如此，交通拥堵如果解决不好，还会阻碍一个城市或地区的经济发展、影响一个城市或地区的形象。交通问题不仅在发展中国家，就是在发达国家也是令人困扰的严重问题。由于严重的生命和财产损失，公众和政府更加关注交通问题[1,2]。

城市轨道交通管理不可能只依赖于传统的管理方法和技术，而必须发展智能交通系统），其中当前发展重点应该是智能交通管理系统（Intelligent Traffic Management System, ITMS），主要包括以下几个系统：自适应交通控制、交通流采集、交通诱导、道路交通视频监控、电子警察（含新型电子警察系统）、交通事件检测、交通综合指挥、面向公众的交通信息服务平台等。

而城市轨道交通客流量预测在各种城市轨道交通控制系统中有相当重要的作用。主要体现在：①在城市集成交通控制系统中的信号控制子系统和动态导引子系统的前提是准确的交通参数的预测。②递阶分层控制中的战略控制层和决策层也要各种交通参数。③智能交通系统中的营运车辆调度管理系统，先进的驾驶员信息系统，先进的交通管理系统等子系统以及自动高速公路系统的网络层和连接层对流量等集总参数进行控制也需要对数据进行预测。

由于传统预测方法的预测精度不够理想且效率很低，影响交通控制。可靠和高效的短期城市轨道交通客流量预测是实现有效的城市轨道交通管理所必需的。短期交通量预测是一项复杂的工作，近几十年来一直是许多研究的课题。所以，城市轨道交通客流量预测已经引起了人们的关注。

本课题在前人的研究基础上，采用多分类算法结合轨道交通拓扑结构实现城市轨道交通客流量预测，算法整体框架如下图所示，本文的贡献如下：

（1）提出一种城市轨道交通断面客流量统计算法；

（2）提出一种城市轨道交通拥堵状态划分算法；

（3）本文提出一种结合轨道交通拓扑结构和提升树算法的流量预测方法。

这个图要换



图 整体框架

# 2.相关工作

现有的轨道交通量预测方法按有无参数可分为参数法和非参数法两类[3-5]。一个典型的参数化方法是季节性自回归移动平均数（Seasonal Autoregressive Integrated Moving Average ,SARIMA）[6-8]。移动平均数认为近期数据与最新数据相似，并使用最新数据加权平均作为预测。基于平稳时间序列相依观测的自回归模型。一个固定时间序列的值在一个常数方差范围内，考虑到现有数据和未来数据具有恒定平均值。在一个固定的系列，在历史与未来之间相互依存，可以用来进行预测。如果原始时间序列是非平稳的，则需要在模型中建立一个集成的部分，通过行为差异使其平稳。此外，也有研究通过将交通流量数据划分为多个模糊区间再做预测[9]。交通数据通常在中午和午夜后达到高峰。此外，周末与工作日相比有不同的交通情况。这些模式为蓝本的季节性部分的平均。

一个重要的非参数算法是k近邻（k-Nearest Neighbor ,KNN）[12,13]。分类和回归任务都可以用KNN处理。对于时间序列回归，KNN的核心思想是寻找具有最新数据相似模式的历史数据。使用KNN解决非线性问题，灵活性高、易于理解且易于实现[14]。一些研究表明，KNN算法优于传统的方法（例如卡尔曼滤波器和SARIMA）[15,16]。然而，也有研究报告表明，KNN与传统的方法有类似的性能[17,18]。

此外，长时交通流量的预测通常指预测未来一个月或一年的交通流量值。由于交通情况变化迅速，长时交通流量的预测在现实应用中很难发挥作用[1]。因此，短时的交通流量预测将会更有用处。例如，我们在驾车出行时想知道未来十分钟前方各个路口的流量情况来决定行驶路线，预测周期太长就会失去实际意义。

深度神经网络模型[19,20]也已经应用到了短时交通流量预测上。预测短时交通流量很大程度上依赖于历史和实时的交通数据，这些数据可以从各种传感器，例如线圈，雷达，GPS（Global Positioning System）以及智能交通卡等多媒体数据中获取。深度学习技术的发展正如火如茶，深度学习模型在交通问题上的应用吸引了大量科研工作人员的注意[2]。深度学习模型己经大量应用在了分类、自然语言处理、降维、目标检测和运动建模等任务上。对交通流量预测也有过早期的尝试，比如使用栈式自动编码器(Stacked AutoEncoder)[20]和深度置信网络（Deep Belief Network，DBN）[21]。这两种方式都是将原有的输入进行预学习处理再将结果作为预测输入进行交通流量预测的。

这些模型和方法可以被归结为以下五个大类：①基于线性统计理论的预测模型、②基于非线性预测理论的模型、③基于组合预测的模型、④基于知识发现的智能预测模型和⑤基于交通仿真的模型[22]，本文着重介绍其中三个。

## 2.1 基于传统统计理论的模型

这类方法是用数理统计的方法处理交通历史数据，对交通流、交通速度、旅行时间等数据进行预测。一般来说统计模型使用历史数据进行预测，方法是假设未来预测的数据与过去的数据有相同的特性。运用于交通预测的传统统计理论方法主要有历史平均模型(History Average Model)[23]、回归分析预测方法[24]、时间序列模型(Time Serial Model)[25]以及卡尔曼滤波模型(Kalman Filtering Model)[26]等。研究较早的历史平均模型方法简单，但精度较差，虽然可以在一定程度内解决不同时间、不同时段里的交通流变化问题，但静态的预测有其先天性的不足，因为它不能解决非常规和突发的交通状况，如交通事故等。

回归分析预测模型是一种通过分析事物之间的因果关系和影响程度进行预测的方法，常用于对多条路段进行分析，其中运用逐步回归方法建立多元回归预测模型受到了极大重视。回归分析预测方法是在可以获得多路段交通数据的基础上，建立起各路段参数之间的线性回归方程，当数据有限时，此方法无法实现[3]。时间序列模型是描述时间序列统计特性的一种常用方法，它是参数化模型处理动态随机数据的一种实用方法。主要有线性平稳模型和非线性平稳模型。线性平稳模型主要有：自回归模型、滑动平均模型、自回归滑动平均混合模型；非线性平稳模型主要有：ARIMA模型和IMA（Integrated Moving Average）模型。自回归求和ARIMA是一种应用得最为广泛的时间序列模型，是自回归模型和滑动平均模型的混合形式。该模型不像其它时间序列方法一样需要固定的初始化模拟，它将某一时刻的交通流量看成是更为一般的非平稳随机序列。在大量不间断数据的基础上，此模型拥有较高的预测精度，但需要复杂的参数估计，而且计算出的参数不能移植。在实际情况中，经常由于各种各样的原因容易造成数据遗漏，导致模型精度降低，而且依赖大量的历史数据，成本很高。

卡尔曼滤波理论由Kalman于1960年提出，该理论采用由状态方程和观测方程组成的线性随机系统的状态空间模型来描述滤波器，并利用状态方程的递推性，按线性无偏最小均方误差估计准则，采用一套递推算法对该滤波器的状态变量作最佳估计，从而求得滤掉噪声的有用信号的最佳估计。由于卡尔曼滤波采用较灵活的递推状态空间模型，因此卡尔曼滤波方法既适应于处理平稳数据，又可用于非平稳数据处理，且对状态变量作不同的假设，可使其描述及处理不同类型的问题，同时减少了计算机存储量和计算时间；模型具有线性、无偏、最小均方差性。卡尔曼增益矩阵可在计算中自动改变，调节信息的修正作用以保持滤波估计的最佳性，具有在线预测的功能。但该方法是线性模型，所以在预测非线性、不确定性的交通流时，模型性能变差。在每次计算时都要调整权值，因此，计算量过大，预测输出值有时要延迟几个时间段。

总体来说，基于传统统计理论的预测方法理论简单、容易理解，但是由于大部分模型都是基于线性的基础，当预测间隔变小，例如小于10分钟，由于交通流量变化的随机性和非线性加强，使得模型的性能变差。预测时仅仅利用了本路段的历史资料，没有考虑相邻路段的影响，这是影响其预测精度的原因之一。

## 2.2 基于神经网络的预测模型

目前，城市轨道交通客流量预测主要都是运用神经网络模型，人工神经网络诞生于20世纪40年代。1964年，Hu应用自适应线性网络进行天气预报，开创了人工神经网络预测的先河；1993年，Vythoulkas PC首次提出用系统识别和人工神经网络进行城市道路网络交通状态的预测[27]。随着神经网络的发展，基于神经网络的短期交通流预测的研究也越来越多。Haitao Li等人使用动态模糊神经网络来做交通流预测[28]；More Rohan等人研究人工神经网络在道路交通量预测与拥塞控制的工作[29]；Li Zhenxing等人提出PLS剪枝算法的前馈神经网络网络流量预测方法[30]。

基于神经网络模型的预测原理为：用一部分数据训练模型，即确定网络结构（包括隐含层数、各层节点数、各层连接权值、各层神经元的传递函数），网路结构确定以后，用剩余部分数据进行预测。总结起来，大体可以分成3类：单一的一类神经网络模型；多种神经网络相结合的优化模型；有神经网络结合其他方法进行预测的综合模型等[31]。神经网络具有识别复杂非线性系统的特性，因此比较适合用于短期交通流预测。目前已从一个简单的单一神经网络模型进行预测，发展到用不同的神经网络模型进行组合，提高了模型的精度和预测效果，再到用神经网络模型和其他领域的先进理论结合进行预测，弥补了神经网络的不足，提高了预测的精度。

但是，神经网络用于短期交通流预测的局限性和不足，由于神经网络的“黑箱”式学习模式，训练过程需要大量的原始数据，且效率较低。同时，没有结合城市轨道交通的自身拓扑结构来研究。神经网络模型的训练过程只能通过调整神经元的权值进行数据处理，即只有神经元外部的处理能力，这种不足导致这类网络存在着局部极小、收敛速度慢、推广能力差以及难以实现在线调整等问题。

## 2.3 基于非线性理论的预测模型

非线性预测主要以混沌理论、耗散结构论、协同论、自组织理论等非线性系统理论为理论基础，利用有关混沌吸引子概念、分形概念、相空间重构方法，数字生态模拟法(Sata Ecology)等建立预测模型。其中发展较成熟的预测方法是混沌理论和小波分析。混沌学是一门新兴学科，混沌理论研究的是非线性动力学系统的混沌。混沌(Chaos)是指一种貌似无规则的运动，在确定性非线性系统中，不需附加任何随机因素亦可出现类似随机的行为(内在随机性)。要用混沌理论对交通流进行分析，首先要判别交通流的混沌特性，交通流系统是有人的群体参与的、开放的复杂巨系统，因此交通中存在着混沌。罗婷等[32]将混沌算法的思想引入到模拟退火粒子群算法中，设计了模拟退火混沌粒子群算法(Simulated Annealing Particle Swarm Optimization,SACPSO)，该算法融合了混沌和模拟退火粒子群算法各自的优点，采用单点单步的方式，来构建SACPSO-BP（Back Propagation）神经网络交通流预测模型。从理论上讲用混沌理论对非线性、不确定性很强的交通流进行预测是非常适合的，所以这类模型将会有很好的发展应用前景。小波分析是20世纪80年代中期发展起来的一门新兴的分析方法，并已在逼近论、微分方程、分形识别、计算机视角及非线性科学等方面都有成功地应用。因小波具有良好的时—频局部化性质，人们将其应用于混沌、非平稳时问序列的分析与预测。近年来发展了一些基于小波分析的预测方法，如将小波框架神经网络用于交通流量预测，取得了较好的效果。但由于该网络固有的参数较多且不易确定，使小波网络在交通流预测中的优势未能充分地发挥。正交小波网络不仅具有小波框架神经网络的优点，而且由于尺度函数的正交性，其参数更易确定，可以提供一个更加有效的函数表示或逼近。

然而，该方法同样没有考虑轨道交通的拓扑结构，此外，鲁棒性较差的问题也妨碍了小波网络的工程应用；应用于预测领域的还有小波理念与kalman滤波的结合模型、小波理论与时间序列的结合模型。小波理论与混沌理论的结合模型、非参数小波算法等。基于小波分析的模型有很大的应用前景，但在交通流预测领域应用不多，将有待继续研究。

# 3 方法

## 3.1 城市轨道交通客流特征分析

3.1.1 时间特征分析

客流量预测是以历史客流数据为依据的，轨道交通路网随着城市交通压力的增大而越来越复杂，客流量也随着时间的不断变化而变化。为了反映历史客流的这种变化情况，得到较为准确的预测结果，在预测之前，首先应该掌握客流的时序特征及变化趋势，如：日规律、周规律、节假日规律等。对于轨道交通客流短期预测来说，历史客流量的数据统特征分析，是有针对性地提出预测方案、改进预测方法及得到较为准确预测结果的前提基础，这也正是本章所要研究的内容。

本文使用的数据是上海市2015年4月份的城市轨道交通刷卡数据。

（1）时间间隔划分

目前轨道交通流量预测主要是指短时客流量的预测，其时间跨度一般不超过20min，时间跨度越小，难度越大，预测精度越难精确。原始数据的时间是以秒为单位，这里时间间隔也选取为20分钟，将一天24小时分为72个时间段，每个时间段相隔20分钟。原始数据格式如下表所示。

表 地铁原始数据格式

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 用户ID | 日期 | 时间 | 地名 | 乘车方式 | 费用 | 是否优惠 |
| 2601377141 | 2015-04-01 | 08:23:50 | 1号线莘庄 | 地铁 | 0.00 | 非优惠 |
| 2601377141 | 2015-04-01 | 19:18:30 | 1号线莘庄 | 地铁 | 4.00 | 非优惠 |

上表中的数据通过时间的转换，得到下表。

表 地铁原始数据格式

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 用户ID | 日期 | 时间 | 地名 | 乘车方式 | 费用 | 是否优惠 |
| 2601377141 | 2015-04-01 | 101 | 1号线莘庄 | 地铁 | 0.00 | 非优惠 |
| 2601377141 | 2015-04-01 | 232 | 1号线莘庄 | 地铁 | 4.00 | 非优惠 |

统计上海轨道交通一天的总客流量分布，结果如下图所示，可以看出呈现出双峰型特点。第一个高峰期出现在上午8点到9点之间，第二个高峰期出现子下午6点到7点之间。符合现实生活中地铁早高峰和晚高峰出现的时间。



图 地铁日进站客流量

（2）星期划分

一个月刷卡数据统计如下图所示，可以看出数据呈现明显的周期性规律。横坐标为2015年4月1号到30号对应的星期。忽略特殊日期（清明假期：4月第一个星期6，7，1）的影响，可以看出每周在周五的时候客流量最大，周日的时候客流量最小。



图 上海地铁一个月刷卡数据

3.1.2 空间特征分析

（1）地铁站点特征



图 同一天不同地铁站客流量

由上图可见，不同地铁站存在明显的客流量大小存在明显的差异。所以，有必要对不同地铁站每天的客流量分布进行分析比较。



图 不同地铁站点客流量随时间变化情况

19 41 20 26

（2）OD矩阵特征

看论文与关联分析或者算法求OD矩阵

## 3.2 提升树算法

基于现有研究存在的不足，本文使用提升树算法进行城市轨道交通客流量的多分类预测。提升方法实际采用加法模型（即基函数的线性组合）与前向分步算法，以决策树为基函数的提升方法称为提升树。提升树可以表示为决策树的加法模型：

（1）

其中，表示决策树；为决策树的参数；为树的个数。

提升树最基本的组成部分叫做分类与回归树（classification and regression tree，CART），CART会把输入根据输入的属性分配到各个叶子节点，而每个叶子节点上面都会对应一个实数分数。

提升树和决策树有所不同，可以将提升树理解为决策树的一个扩展。从简单的类标变换到数值之后，我们可以做很多工作，例如概率预测，排序等。

已知一个训练数据集，，为输入空间，，为输出空间。对于一个集合树模型，输出的表达式如下：

（2）

是给定第次CART树时的预测结果。

对于给定的，最终的预测结果是每棵树的预测之和。

将单个CART表示为，定义实例对应树中相应的叶子：。根据公式2，对于给定的实例，第颗CART的预测结果可以写成如下格式：

（3）

表示CART中叶子结点的权重，表示由树结构定义的映射函数。

树集合模型的学习目标可以被理解为求以下损失函数的最小值：

（4）

误差函数表示目标结果与公式1中预测结果的差异性。表示实例的总数。表示算法中CART的总数。这里为正则化项，用来惩罚复杂模型，避免模型过度拟合。可以表示为如下公式：

（5）

其中，和是正则花参数。

训练模型的目的是为了求总体损失函数最小值。然而传统的优化算法不能在欧式空间使目标函数最小化，因为公式4中的损失函数依赖参数，同样依赖每棵树的结构。为了高效地解决这个问题，梯度下降提升树算法被提出，关于该算法这里不再赘述。

## 3.2 基于时空属性的轨道交通断面客流量统计算法

为了便于陈述，表列出了相关的符号及其含义。

表 符号表

|  |  |
| --- | --- |
| 符号 | 含义 |
|  | 上行起始站 |
|  | 上行终点站 |
|  | 时刻 |
|  | 站点客流量 |
|  | 进站客流 |
|  | 出站客流 |

目前对断面客流的计算方法还较少，已有的研究中用到的断面客流数据也多没有说来源，也不能实时掌握网络化运营的轨道交通线网中各站台候车客流、各班次列车载客量以及各断面客流的分布情况。这给城市轨道交通客流量预测以及运营调度带来了不便。基于此，本文提出一种结合时间片段和关联站点的城市轨道交通断面客流量递归统计算法。

有如下一条城市轨道交通线路：



图 轨道交通拓扑图

将地铁站名转换为ID，并定义ID从小到大的的方向为地铁的上行方向。

这里首先考虑上行方向，将站点在时刻的客流量表示为如下形式：

下行方向的站点客流量表示为如下形式：

通过上面的公式可以看出，除了起始站，其他站点的断面客流量都与上一个站点的断面客流相关。

## 3.3 城市轨道交通拥堵状态分类算法

思路：使用聚类的方法结合地铁车辆的容量及发车时间间隔来计算。

# 4 实验及结果

## 4.1.数据预处理

本文使用的数据是上海市2015年4月份的交通数据，其中包括地铁刷卡、公交车刷卡及出租车轨迹等数据。这里选取其中一个月的地铁刷卡数据如下图所示。可以看出数据呈现明显的周期性规律。



图 上海地铁一个月刷卡数据

原始数据格式如下：

表 地铁原始数据格式

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 用户ID | 日期 | 时间 | 地名 | 乘车方式 | 费用 | 是否优惠 |
| 2601377141 | 2015-04-01 | 08:23:50 | 1号线莘庄 | 地铁 | 0.00 | 非优惠 |
| 2601377141 | 2015-04-01 | 19:18:30 | 1号线莘庄 | 地铁 | 4.00 | 非优惠 |

数据预处理流程：

（1）每5分钟分为一个时间段

（2）将日期转换为星期

（3）计算相同时间段的人流量

由于整个上海地铁十分复杂，共有19条地铁线路，共计287个站点。本文只以上海市轨道交通16号线为研究对象，16号线共有13个站，其中罗山路和龙阳路为换乘站点。定义滴水湖站到龙阳路站为地铁上行方向。得到如下的16号线地铁站ID表。

表 16号线站点及对应ID

|  |  |  |
| --- | --- | --- |
| 站名 | ID |  |
| 滴水湖 | 1 | 198 |
| 临港大道 | 2 | 41 |
| 书院 | 3 | 43 |
| 惠南东 | 4 | 136 |
| 惠南 | 5 | 135 |
| 上海野生动物园 | 6 | 20 |
| 新场 | 7 | 141 |
| 航头东 | 8 | 223 |
| 鹤沙航城 | 9 | 277 |
| 周浦东 | 10 | 81 |
| 罗山路 | 11 | 215 |
| 华夏中路 | 12 | 67 |
| 龙阳路 | 13 | 287 |

# 5 总结