# 基于拓扑结构的轨道交通人流量预测

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：**我国进入网络化运营的时代背景要求在满足大客运量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，常常会在交通网络上造成难以预见的网络客流拥塞状况，并且由于流量的动态演化性而产生各种急剧突发和影响重大的客运安全问题，为安全运营带来极大的挑战。随着我国城市轨道交通网络复杂化程度提髙，在面临严峻的客运安全问题时，当采取相应管控措施时，目前经验判断为主，缺少对于拥塞条件下的客流分布状态的准确把握和分析，缺少对于网络流量演化的精确认知和评估，缺失网络流量拥塞的有效疏导和控制。而这些问题也是目前运营管理人员、国内外研究学者以至整个社会都高度重视的问题，因此需要从理论上进行合理解释，从技术方法上进行科学建模，从实践应用上进行有效验证。解决以上问题有利于提高城市轨道交通网络化运营下的大客流评估和分析技术，有利于提高网络化运营条件下的拥塞状态分析和管理能力，有利于提高城市轨道交通网络拥挤管理控制的决策水平。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。

**关键字：**城市轨道交通，拓扑，提升树

# Prediction of Human Flow in Rail Transit

# Based on Topology

**Abstract:** The urban rail transit(URT) operation and management department has been required to gradually improve the transport technology and service quality, and implement the service concept of people-oriented in order to satisfy the safety and effectiveness of the huge volume of passenger ridership under the background that China has entered the era of network operation of urban rail transit. However, the distribution imbalance between the passenger demand and capacity supply often cause unpredictable congestion state int the network. And because of the dynamic evolution of the passenger flow, a variety of sudden and outburst safety transport problem with huge impact will be generated, which brings great challenge to the safe operation. As the complexity degree of the URT network increasing, the operation departments will take the management and control measures based on the existing experience at present when faced with the severe passenger security issue, which lacks of the accurate grasp and analysis of the passenger distribution under congestion state, lacks of the accurate understanding and evaluation of the passenger flow evolution, lacks of the effective grooming and control strategy to network grooming. These problems are also highly valued with the operation department, domestic and foreign researchers and the whole society. Therefore we need to bring out reasonable explanation from theory, scientific modeling from the technical method and effective verification from practical application. To solve the above problems will help to improve the large passenger flow evaluation and analysis technology, the congestion state analysis and management technology, the crowd and congestion grooming and control technology in decision making under network operation. Based on the above problems , This paper presents a traffic forecasting method that combines the topological structure of rail transit and the boosting tree algorithm.

**Keywords:** Urban Rail Transit, Topology, Boosting Tree

## 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]。有效的交通管理和自动事故检测是现代智能交通系统（Intelligent Transport System ,ITS）的关键要求。

城市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。

而交通流量预测在各种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，影响交通控制。可靠和准确的短期交通预测是实现有效的交通管理和事故检测所必需的。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，交通流预测已经引起了广泛的关注。

本文组织结构如下：第一部分介绍交通流量预测的背景及国内外现状；第二部分介绍涉及到的相关知识，着重介绍网络拓扑、图以及监督学习算法；第三部分介绍本文提出的结合网络拓扑与监督学习的轨道交通流量预测方法，以及如何建立网络拓扑结构和如何构建训练集；第四部分为实验，通过将本文提出的方法用代码实现并于目前常用的方法作比较；第五部分总结本文的内容，分析实验结果，并指出未来的工作。

## 2. 国内外研究现状

现有的短期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。

在非参数范畴中，提升树得到了广泛的应用。梯度提升是改进提升树的一种方法。它通过专注于预测的实体来改进树模型。极端的梯度提升是一种快速、正规化的算法，其计算速度更快，模型更好。Xgb（eXtreme Gradient Boosting）显示出出色的性能和工作效率，这要优于随机森林[10,11]。

另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，也有研究报告说，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]。

### 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、旅行时间等用于预测。一般来说统计模型使用历史数据进行预测，它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，是一种在现代控制理论中被广泛采用的先进的埋单序列方法，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，当预测间隔小于5min时，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

基于知识发现的交通流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。但应该看到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；训练完成的网络只适合于当前研究路段；同时，神经网络的学习算法采用经验风险最小化原理，不能使期望风险最小化，在理论上存在缺陷。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。目前神经网络在交通流预测领域的研究大多属于验证性的研究，即用人工神经网络方法完成预测并与其他方法进行对比，没有指出神经网络的适用范围和应用条件。每一种预测模型或方法都有其适用性，应进一步研究各种神经网络模型的适用条件和范围。

### 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，鲁棒性较差的问题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 数据处理

本文使用的数据是上海市15年的交通数据，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。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20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

## 4 方法

（1）提升树模型

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（Boosting Tree）。提升树可以表示为决策树的加法模型：



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。从简单的类标到分数之后，我们可以做很多工作，例如概率预测，排序等。

（2）提升树算法

（3）梯度提升

（一种基于贝叶斯超参数优化的提升决策树信用评分方法）

2.1 集合树和学习目标

已知给定数据集如下，

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：



是给定第k次CART树时的预测结果。下图1描述了集合树模型。

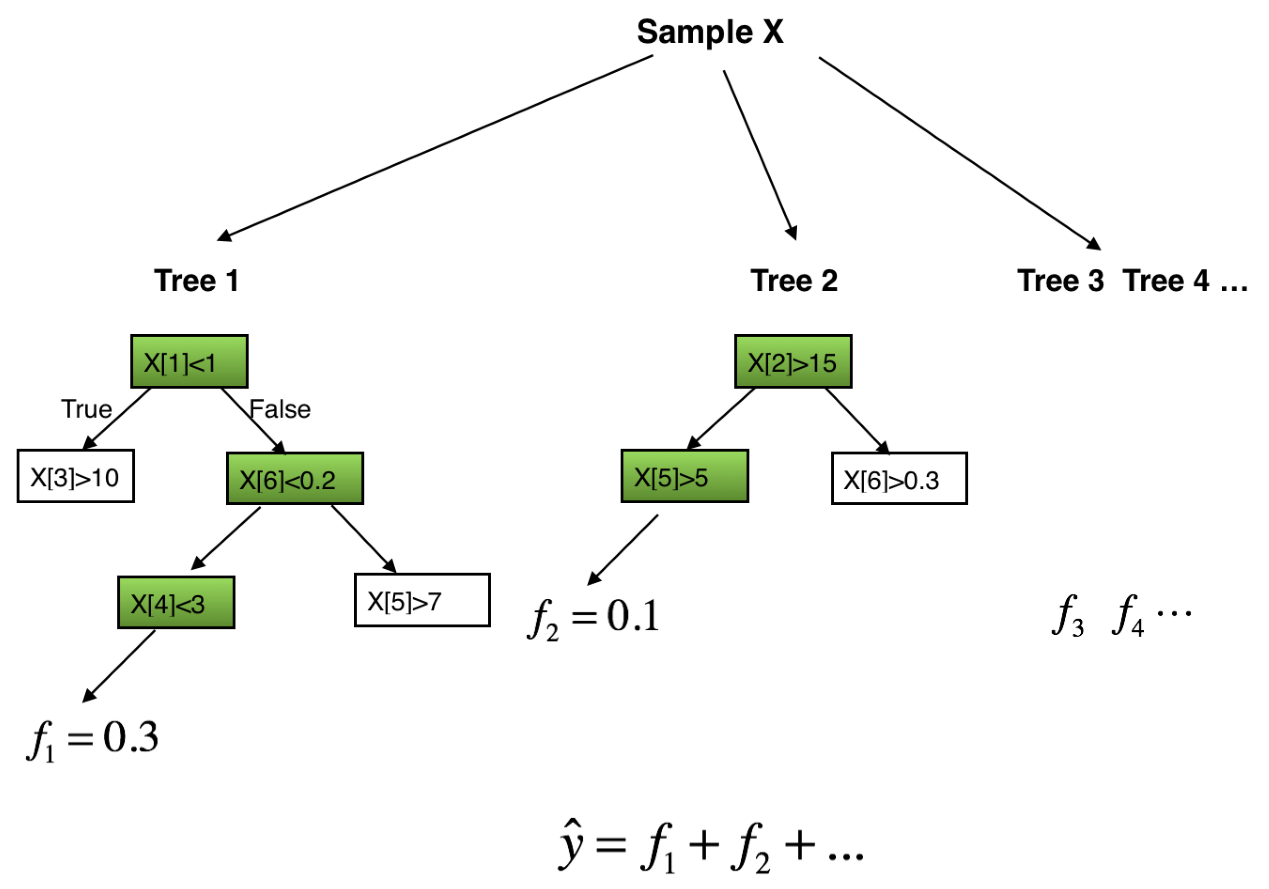


图1 集合树模型。

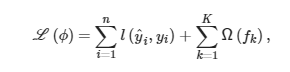
对于给定的xi，最终的预测结果Yˆ是每棵树的预测之和。

将单个CART表示为T，定义实例x对应树中相应的叶子：。根据公式1，对于给定的实例xi，第k颗CART的预测结果可以写成如下格式：



W标识CART中叶子结点的权重，q(xi)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：



误差函数l(yi,yi)表示目标结果yi与公式1中预测结果yi的差异性。N表示实例的总数。K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：

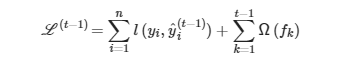




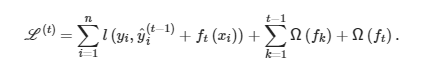
2.2 梯度下降提升树

训练模型的目的是为了最小化总体损失函数。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3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。

在训练模型中，第t-1次迭代的损失函数定义如下：

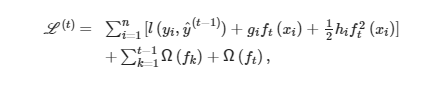


表示第t-1次迭代的预测结果。在下一次迭代也就是第t次迭代中加入ft，目标函数变成如下形式：

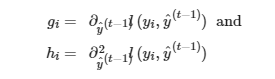


换句话说，根据公式3，贪婪地添加树来改善模型。所以，在第t次迭代，加入第t颗CART，目的是为了是最小。

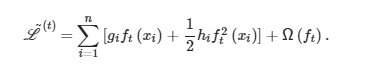
首先，用泰勒表达式展开目标函数：



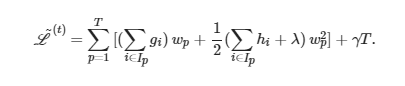
其中：



这个目标函数有一个非常明显的特点，它只依赖每个数据点在误差函数上的一阶导数和二阶导数。在第t次迭代，前一颗树t-1被改善，它的目标函数可以看成一个常量，因此，我们在上式中移除它，得到下面这个 简单的目标函数：



结合公式2和公式4，此表达式等价于下面这个表达式：

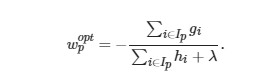


表示叶子p的数据实例集。所有的数据实例都被映射到叶子p上。

因此，对于一个固定的树结构q(x),叶子p的最优权重由下面的最小化方程定义：



由方程10定义，这个函数的结果如下：

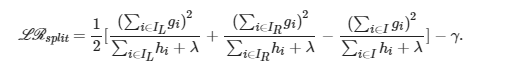


有了这个最优权重，得到如下的最优目标函数方程：



因此，对于每一次迭代，我们的目标变为寻找最优的树q是目标函数最小。

但是，通过枚举所有可能的树是不切实际的。相反，可以贪婪算法解决这个问题。从一颗简单树出发，我们可以迭代地分割树节点并向树中添加分支。对于每次迭代，在分割之前定义数据实例集为I，在分割之后定义数据集的左节点为IL，右节点定义为IR。因为，使用方程13，分解后的目标函数为：



因此，为了找到最优树，该算法通过选择最大化的分裂迭代地增加分支。

## 5 实验结果

## 6 总结