Advances in Applied Mathematics 应用数学进展, 2020, 9(9), 1364-1376
Published Online September 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.99162

Hans汉斯

基于XGBoost算法的短期交通流预测

刘　伟[1]，Subhash C. Bagui[2]，贾宏恩[1]

[1]太原理工大学数学学院，山西　晋中
[2]西佛罗里达大学数学与统计系，佛罗里达　彭萨克拉，美国
Email: 986924878@qq.com

收稿日期：2020年8月16日；录用日期：2020年9月1日；发布日期：2020年9月8日

摘　要

针对短期交通流预测问题，为完成实时精准预测，建立了一种基于Huber损失的极端梯度上升(Extreme Gradient Boosting, XGBoost)短时交通流预测模型。通过对交通流数据周期性、关联性的分析，提取时间特征，并进行时间特征重要性分析。利用该模型以及提取的特征进行交通流预测，实验结果表明：该模型优于基于均方误差损失的极端梯度上升模型以及基于平均绝对误差损失的极端梯度上升模型。同时，该模型较梯度提升回归模型、支持向量机回归模型具有更高的预测精度，各误差指标小，且模型训练时间短，符合短时交通流预测所要求的时效性。

关键词

交通流预测，极端梯度提升(XGBoost)，Huber损失函数，特征重要性分析

Short-Term Traffic Flow Prediction Based on XGBoost

Wei Liu[1], Subhash C. Bagui[2], Hong'en Jia[1]

[1]College of Mathematics, Taiyuan University of Technology, Jinzhong Shanxi
[2]Department of Mathematics and Statistics, The University of West Florida, Pensacola, FL, USA
Email: 986924878@qq.com

Received: Aug. 16th, 2020; accepted: Sep. 1st, 2020; published: Sep. 8th, 2020

Abstract

For short-term traffic flow prediction, in order to complete real-time accurate prediction, an ex-

文章引用: 刘伟, Subhash C. Bagui, 贾宏恩. 基于 XGBoost 算法的短期交通流预测[J]. 应用数学进展, 2020, 9(9): 1364-1376. DOI: 10.12677/aam.2020.99162

treme gradient boosting (XGBoost) short-term traffic flow prediction model based on Huber loss is established. By analyzing the periodicity and relevance of traffic flow data, time features are extracted and feature importance analysis is performed. Using this model and the extracted features for traffic flow prediction, the experimental results show that the model is superior to the extreme gradient boosting model based on mean square error loss and the extreme gradient boosting model based on average absolute error loss. At the same time, the model has higher prediction accuracy than gradient boosting regression model and support vector machine regression model, each error index is small, and the model training time is short, which meets the timeliness required by short-term traffic flow prediction.

Keywords

Traffic Flow Prediction, Extreme Gradient Boosting (XGBoost), Huber Loss Function, Feature Importance Analysis

1. 引言

随着经济水平的快速发展，道路交通的需求量也随之增加，然而城市交通拥堵导致车辆行驶速度降低、行程时间延长、排放尾气增加、出行成本上升，这直接成为了制约城市发展的重要问题之一。因此，短时交通流预测研究对城市智能交通系统的建设发展具有重要意义[1]。短时交通流预测是根据历史交通流数据对下一个时间间隔的交通流量进行预测的过程，既有缓解交通拥挤、提高运输效率的作用，也为城市交通规划提供了有效的依据。

从 20 世纪 60 年代开始，国内外学者就建立了多种模型用于短时交通流预测，主要可分为传统的数理统计模型、非线性理论模型以及人工智能模型。传统的交通流预测模型包括时间序列预测模型[2] [3]、卡尔曼滤波模型[4] [5]、自回归模型[6]、傅里叶变换模型[7]等。随着人工智能的发展，更多的机器学习模型和深度学习模型开始运用于交通流预测。神经网络自学习能力强而且自适应性强，在识别复杂非线性问题中应用广泛，但存在算法收敛速度慢、预测精度低等缺点。支持向量机模型可以避免神经网络结构选择和局部极小点问题，在非线性、高维空间泛化性高，但对缺失数据敏感且仅局限于小集群样本。梯度提升树属于 Boosting 集成算法的一种，是一种迭代决策树算法，其由多棵决策树组成，每棵树都是对前序模型的不足之处进行改进，结果由各决策树累加而得到，从而得到了一个强学习器，但存在容易过拟合的问题。极端梯度提升算法(Extreme Gradient Boosting, XGBoost)在梯度提升算法的基础上做了进一步的改进，它在目标函数中加入了正则项并且采取列抽样，不仅防止模型过拟合，而且运算速度得到了明显的提升。

因此，为提高短期交通流预测精度及训练速度，本文采用 XGBoost 算法进行短期交通流预测。近年来，研究者采用多种调节参数方法对 XGBoost 算法进行优化，却较少优化 XGBoost 算法的目标函数。在本文中，为了充分发挥 XGBoost 算法的框架作用，根据交通流数据的特征，选用 Huber 损失作为目标函数。实验结果表明：基于 Huber 损失的极端梯度上升模型具有更高的预测精度，各误差指标小，且模型训练时间短，符合短时交通流预测所要求的时效性。

2. XGBoost 原理

2.1. XGBoost 目标函数定义

极端梯度提升算法(Extreme Gradient Boosting, XGBoost)是对梯度提升算法(Gradient Boosting Decision Tree, GBDT)的一种改进，通过在目标函数中添加正则项，进一步提升了算法性能。其定义如式(1)所示：

$$\begin{cases} \hat{y}_i^{(0)} = 0 \\ \hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\ \vdots \\ \hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \end{cases} \tag{1}$$

式(1)中，$\hat{y}_i^{(t)}$为第 t 轮的预测值，其等于前 $t-1$ 轮预测值加上新的函数 $f_t(x_i)$。在定义目标函数时，决策树自身存在惩罚项[8]。惩罚项表达式如式(2)所示：

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \tag{2}$$

式(2)中：γ 为惩罚力度；λ 为参数；ω 为各叶子节点权重；T 为叶子节点个数。因此，包含惩罚项的目标函数如式(3)所示：

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t}\Omega(f_k) \\ &= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + \text{constant} \end{aligned} \tag{3}$$

用二阶泰勒展开式来近似原来的目标，泰勒展开式如式(4)所示：

$$f(x+\Delta x) \approx f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2 \tag{4}$$

根据式(4)的泰勒展开式以及式(3)的目标函数表达式，定义：

$$\Delta x = f_t(x_i) \tag{5}$$

$$f(x) = l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{6}$$

$$g_i = \frac{\partial l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^{(t-1)}} \tag{7}$$

$$h_i = \frac{\partial^2 l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^{(t-1)2}} \tag{8}$$

式中：g_i 对应于泰勒展开式中的一阶导数；h_i 对应于泰勒展开式中的二阶导数，故目标函数表达式可通过泰勒展开式转换为式(9)：

$$Obj^{(t)} \approx \sum_{i=1}^{n}\left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) + \text{constant} \tag{9}$$

将式(9)中的常数项移除后，得到目标函数的最终形式，如式(10)所示：

$$Obj^{(t)} \approx \sum_{i=1}^{n}\left[g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t) \tag{10}$$

从式(10)可以看出，该目标函数的值仅取决于 g_i 和 h_i。

2.2. XGBoost 的目标函数求解

我们每次添加的树要使我们的目标函数最优，那么我们的目标函数可以写成：

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^{n}\left[g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t) \\ &= \sum_{i=1}^{n}\left[g_i w_q(x_i)+\frac{1}{2}h_i \omega_q^2(x_i)\right]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)w_j+\frac{1}{2}\left(\sum_{i\in I_j} h_i+\lambda\right)\omega_j^2\right]+\gamma T \\ &= \sum_{j=1}^{T}\left[G_j w_j+\frac{1}{2}\left(H_j+\lambda\right)\omega_j^2\right]+\gamma T \end{aligned} \tag{11}$$

上式中，$G_j=\sum_{i\in I_j} g_i$ 表示叶子节点一阶导，$H_j=\sum_{i\in I_j} h_i$ 表示叶子节点二阶导。从上式可以看出 XGBoost 较于其他算法的优越性之一就表现在将样本遍历转化为叶子节点遍历，这样提高了运算速度。

最后的目标函数是关于 ω_j 的二次函数，这样它的极小值点和极小值分别为：

$$\omega_j^* = -\frac{G_j}{H_j+\lambda} \tag{12}$$

$$Obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T \tag{13}$$

最后一个式子正是衡量树结构好坏的标准，$q(x)$ 值越小代表结构越好。当我们指定损失函数后，就可以求出 G_j 和 H_j，继而可以得到目标函数的极小值。通常，我们选用的损失函数为均方误差(Mean Squared Error, MSE)或者平均绝对误差(Mean Absolute Error, MAE)。

有了这个标准后，我们理应尝试所有可能并选择最优，然而太费时。于是我们每次优化树的一层。我们假设一个叶子分裂为两个叶子，则它的得分增加为：

$$Gain=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]-\gamma \tag{14}$$

如果增益小于 γ，我们就不将此叶子分裂。

3. 基于 XGBoost 算法的交通流预测模型

3.1. 自定义 XGBoost 目标函数

机器学习中的所有算法都依赖于最小化或最大化某一个函数，我们称之为“目标函数”，最小化的这组函数被称为“损失函数”。在建立模型时，为了充分发挥 XGBoost 算法的框架作用，需要根据不同的数据集以及任务自定义目标函数，来达到更优的效果[9]。

均方误差(Mean Square Error, MSE)是目标变量与预测值之间距离的平方和，平均绝对误差(Mean Absolute Error, MAE)是目标变量和预测值之间距离的绝对值之和。以 MSE 为损失函数的模型会给离群点

赋予更高的权重，而以 MAE 为损失函数的模型对于小的损失值，其梯度也很大，而且其导数并不连续，求解效率较低。Huber 损失结合了 MSE 与 MAE 的优点，降低了对离群点的惩罚程度，能增强 MSE 对离群点的鲁棒性。

在本文中，根据交通流数据的特征，选用 Huber 损失作为目标函数。Huber 损失的定义如下：

$$L_\delta\left(y_i,\tilde{y}_i\right)=\begin{cases}\frac{1}{2}\left(y_i-\tilde{y}_i\right)^2, & \left|y_i-\tilde{y}_i\right|\le\delta\\ \delta\left|y_i-\tilde{y}_i\right|-\frac{1}{2}\delta^2, & \text{otherwise}\end{cases} \tag{15}$$

式中：y_i 表示真实值；$\tilde{y}_i$ 表示预测值；δ 是一个常数，表示邻域，可以通过调节参数得到 δ 的最优取值。

Huber 损失是不可导函数，所以在本文中用的是 Huber 损失的可导逼近形式(伪 Huber 损失函数作为目标函数)，其定义如下：

$$L_\delta\left(x\right)=\delta^2\left(\sqrt{1+\left(\frac{x}{\delta}\right)^2}-1\right) \tag{16}$$

式中：$x=\tilde{y}_i-y_i$，表示预测值与真实值的误差。

伪 Huber 损失函数的一阶导数为：

$$\frac{\partial}{\partial x}\left(\delta^2\left(\sqrt{1+\frac{x^2}{\delta^2}}-1\right)\right)=\frac{x}{\sqrt{1+\frac{x^2}{\delta^2}}} \tag{17}$$

伪 Huber 损失函数的二阶导数为：

$$\frac{\partial^2}{\partial x^2}\left(\delta^2\left(\sqrt{1+\frac{x^2}{\delta^2}}-1\right)\right)=\frac{1}{\left(1+\frac{x^2}{\delta^2}\right)^{\frac{3}{2}}} \tag{18}$$

当预测误差小于 δ 时，Huber 损失函数采用平方误差；当预测误差大于 δ 时，Huber 损失函数采用的是线性误差。

3.2. 交通流预测模型实现流程

基于交通流数据在时间上的变化规律，采用上述模型对交通流数据实现实时预测。实现流程可分为 4 个阶段：数据处理阶段，特征提取阶段，模型优化阶段，可行性分析阶段[10]。具体流程如图 1 所示。

数据处理阶段：对采集数据进行缺失值填充等预处理，处理后将数据集分为训练集与测试集。

特征提取阶段：对数据集的规律进行刻画描述，提取时间特征，并对各个特征进行重要性分析，选取重要性程度最高的 6 个特征。

模型优化阶段：对 XGBoost 模型的目标函数进行优化，并且采用 Hyperopt 方法对各个参数进行调节。

可行性分析阶段：利用训练好的模型在测试集上进行测试，将预测结果与真实值进行比较，获取模型预测性能。将本文模型与常用预测模型进行比较，分析模型可行性。

4. 实例分析

4.1. 数据来源

本文所用交通流数据来源于美国加利福尼亚州戴维斯附近的 I80 走廊，传感器设置如图 2 所示。数

据采集时间为 2016 年 1 月 1 日至 10 月 12 日，采集时间间隔为所 30 s。该数据集共包含 812,792 组数据。为构建基于 XGBoost 算法的短期交通流预测模型，本文采用 7、8 月份的数据来进行建模分析。具体的，用 24 天的数据(7 月 1 日~7 月 24 日)作为训练集，用 10 天(7 月 25 日~8 月 3 日)的数据作为测试集。

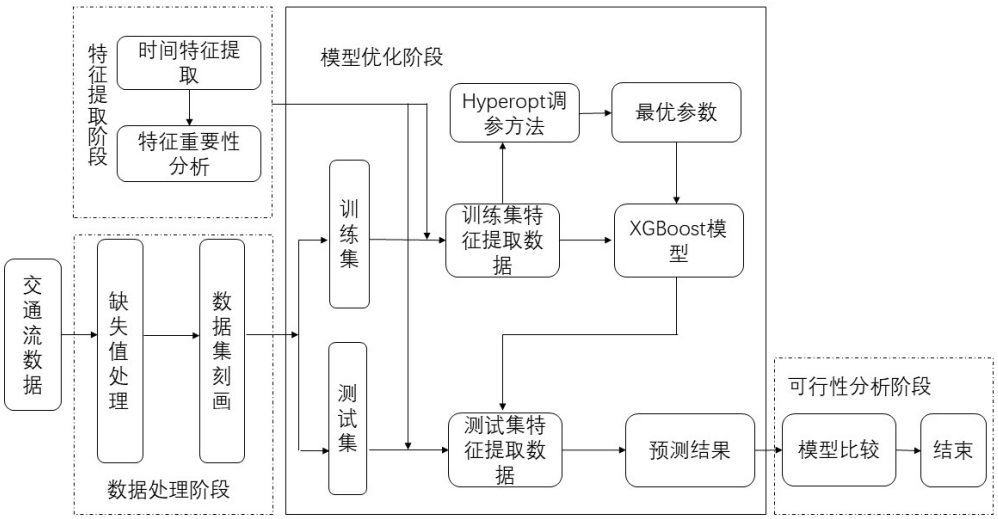

Figure 1. Realization process of traffic flow prediction model
图 1. 交通流预测模型实现流程

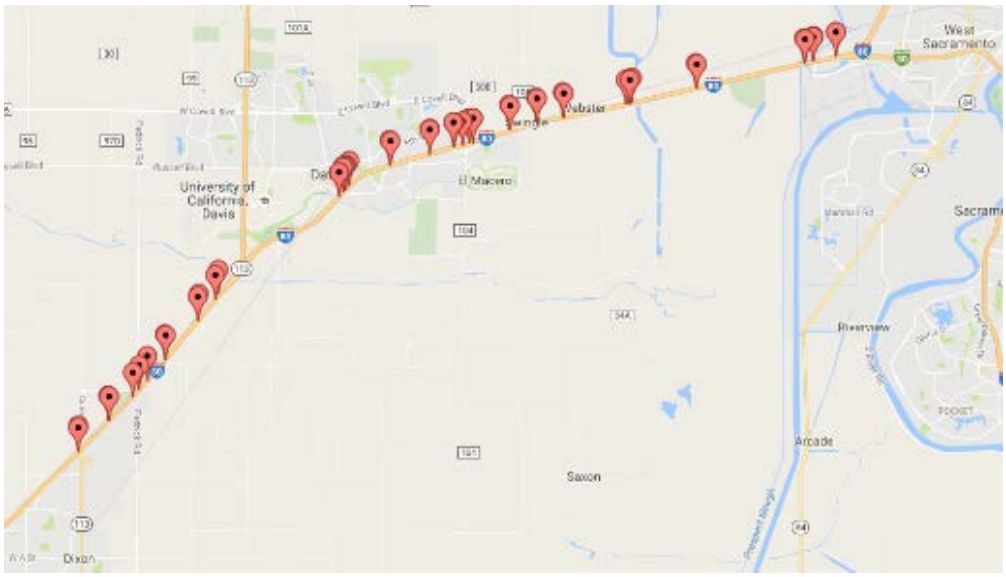

Figure 2. Sensor settings in I80 Corridor, California
图 2. 加利福尼亚州 I80 走廊传感器设置

4.2. 数据预处理

数据集中存在不规则小数据量间断性缺失情况。因此，需要对数据进行插补修复。考虑到交通数据的动态特征和即时特性，实验过程中使用历史均值法和最近邻均值法(取该缺失数据上下时间节点的平均值作为插补值)相结合的方法进行填充。历史均值法适合连续缺失数据，最近邻均值法适合单点缺失数据。

通过修复得到完整数据，之后继续将交通流处理成以 5 min 为间隔的数据，便于进一步分析、训练

和预测。

在得到完整的数据集后，将对交通流数据的分布进行刻画。随机选取某一周数据(2016 年 7 月 11 日~2016 年 7 月 17 日)和某一天数据(2016 年 7 月 11 日)进行考察，分别如图 3、图 4 所示。横坐标表示交通流量，纵坐标表示密度。可以发现，其分布很符合二维的混合高斯分布。高斯分布具有很好的数学性质，因此不需要进行数据变换。

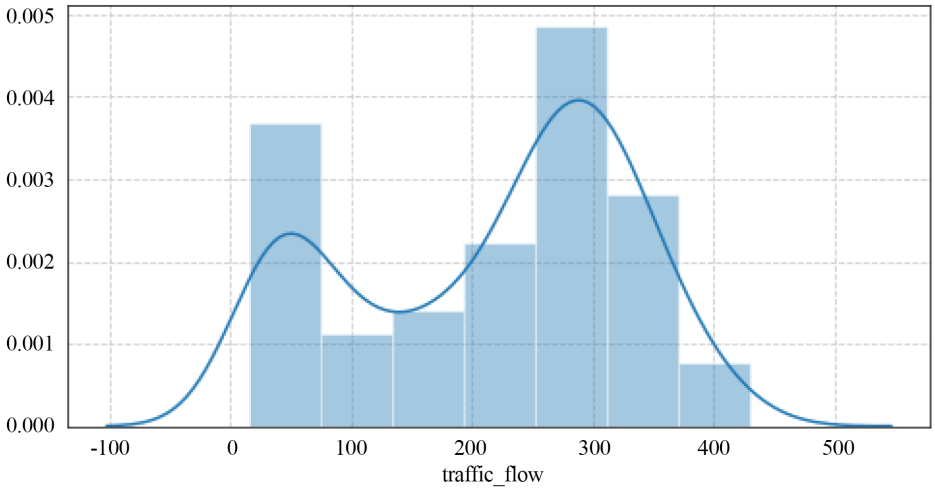

Figure 3. One-day traffic flow data distribution map
图 3. 一天交通流数据分布图

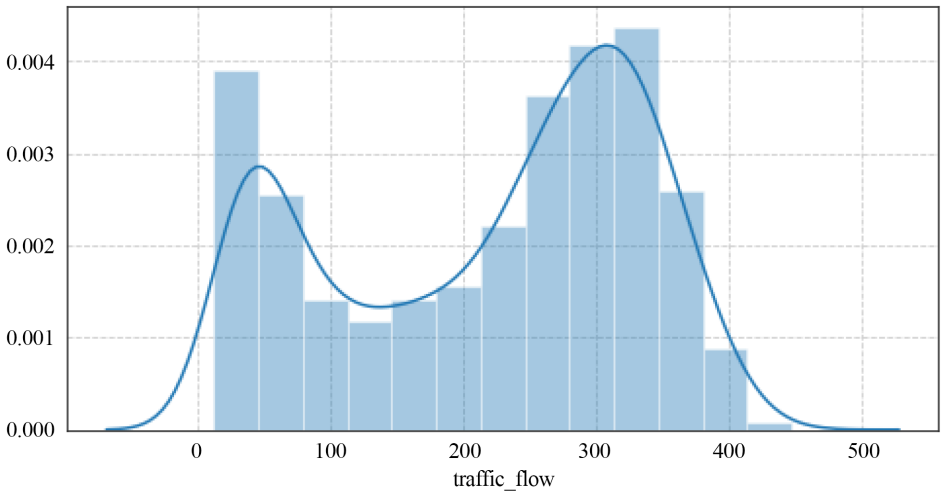

Figure 4. One-week traffic flow data distribution map
图 4. 一周交通流数据分布图

4.3. 特征提取

首先对交通数据规律进行描绘观察，随机选取连续两周(2016 年 7 月 1 日~2016 年 7 月 14 日)的交通数据进行刻画，通过观察处理后的数据(时间间隔为 5 min)容易发现，交通流数据具有极强的周期性及连续性，即时间相近的数据其状态也更相似，如图 5 所示(横坐标表示时间组数，每组时间为 5 min；纵坐标表示交通流量)。

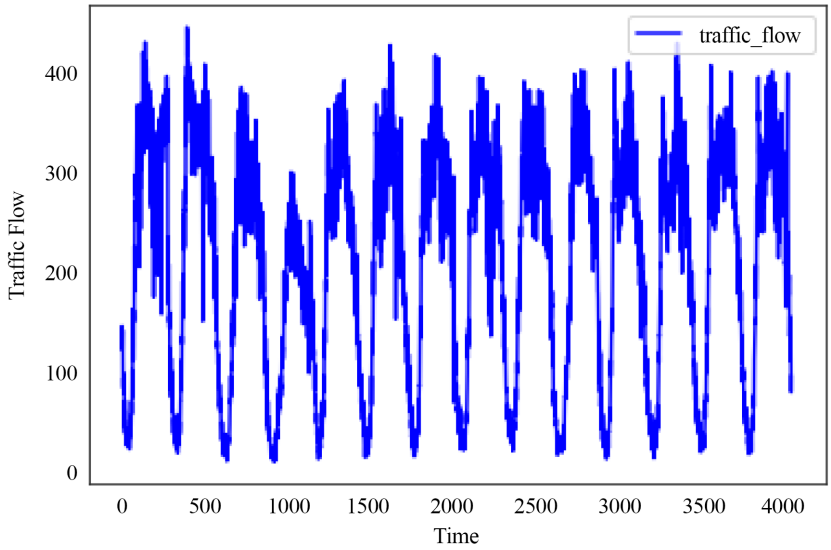

Figure 5. Two-week traffic flow data graph
图 5. 两周交通流数据图

实验中，提取交通流数据的月周期特征、周周期特征、最近前 3 天状态特征、最近前 2 天状态特征、最近前 1 天状态特征和最近 5 min 的状态特征，共 6 个特征作为数据属性，整理数据集用以训练和预测。

对模型各个特征进行重要性分析，如图 6 所示(f0 表示月周期特征，f1 表示周周期特征，f2 表示最近前三天状态特征，f3 表示最近前两天状态特征，f4 表示最近前一天状态特征，f5 表示最近 5 min 的状态特征)。

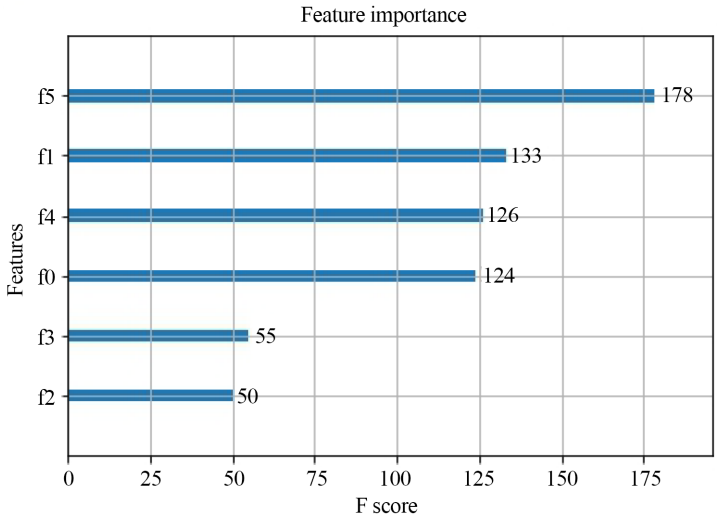

Figure 6. Feature importance analysis chart
图 6. 特征重要性分析图

由图 6 可知，最近 5 min 的状态特征最重要，因为交通流数据是时间序列数据，下一时刻的状态必定与上一时刻的状态紧密相关；周周期特征重要性排名第二，是否是工作日对交通流预测具有很重要的影响；重要性排名第三的特征是最近前 1 天状态特征，排名第四的是月周期特征，排名第五的是最近前 2 天状态特征，排名最后的是最近前 3 天状态特征。

4.4. XGBoost 参数调优

XGBoost 模型参数众多，使用 Hyperopt 方法对各个参数进行调节，如表 1 所示。

Table 1. Parameter value of traffic flow prediction model
表 1. 交通流预测模型参数取值

参数名	取值	参数名	取值
n_estimators	60	scale_pos_weight	0.996
learning_rate	0.10	subsample	0.71
max_depth	7.00	colsample_bytree	0.68
min_child_weight	5.00	gamma	0.65

1) n_estimators：弱学习器的最大迭代次数，或者说最大的弱学习器个数。n_estimators 太小，容易欠拟合，n_estimators 太大，又容易过拟合。

2) learning_rate：学习率，可以减少每一步的权重，提高模型的鲁棒性。

3) max_depth：数的最大深度。

4) min_child_weight：决定最小叶子节点样本权重和。

5) scale_pos_weight：样本十分不平衡时，将这个参数设置成正数，可以使算法更快收敛。

6) subsample：随机采样比例。

7) colsample_bytree：列采样率，也就是特征采样率。

8) gamma：分裂节点时，损失函数减小值只有大于等于 gamma，节点才分裂。

4.5. 模型预测结果及分析

为评估模型的效果，对未来 10 天(7 月 25 日~8 月 3 日)的交通流进行预测，如图 7 所示。

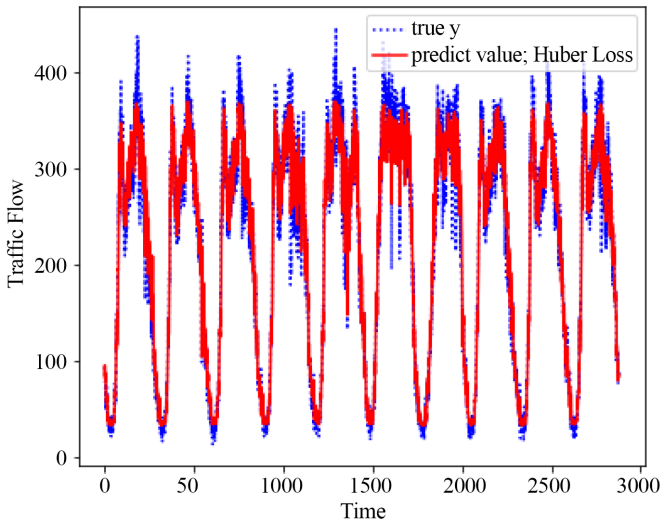

Figure 7. Ten-day prediction graph
图 7. 基于 Huber 损失的 XGBoost 模型十天的预测图

图 7 中蓝色曲线为 7 月 25 日~8 月 3 日实际的交通流数值，红色曲线表示由该模型进行预测的交通流预测值。从图中可以看出，红色曲线和蓝色曲线是基本重合的，这表示本文中提出的 XGBoost 模型能

够很好地拟合交通流数据，且预测精度非常高。为了更清晰地展示该模型的效果，可选择从 10 天预测数据中随机抽取某一天来进行展示，如图 8 所示。图 8 中蓝色曲线为随机抽取的 7 月 25 日的交通流实际值，红色曲线为模型预测值。

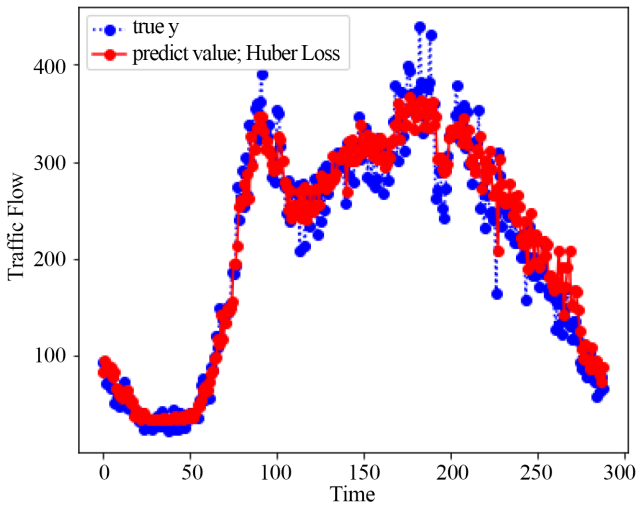

Figure 8. Single-day prediction graph
图 8. 基于 Huber 损失的 XGBoost 模型的单天预测图

5. 不同模型预测结果对比分析

5.1. 不同模型预测结果评价指标

使用 3 种常用的性能指标来对模型的预测效果进行判定，它们分别是均方根误差(Root Mean Square Error, RMSE)、平均绝对误差(Mean Absolute Error, MAE)和决定系数(R^2)，从不同侧面来反应模型的预测效果。各指标定义如下：

$$\mathrm{RMSE}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2} \tag{19}$$

$$\mathrm{MAE}=\frac{1}{m}\sum_{i=1}^{m}\left|y_i-\hat{y}_i\right| \tag{20}$$

$$R^2=1-\frac{\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{m}\left(y_i-\bar{y}_i\right)^2} \tag{21}$$

式中，y_i 为真实值，$\hat{y}_i$ 为预测值，$\bar{y}_i$ 为均值，m 为测试样本个数。

通过计算上述 3 个性能评价指数值，来评估所建模型的预测性能。均方根误差和平均绝对误差越小模型效果越好，R^2 越接近于 1，模型越好，反之，越差。

5.2. 不同模型预测结果对比

为评价基于不同自定义目标函数的 XGBoost 模型的预测性能，同时采用 Huber 损失、均方误差以及平均绝对误差作为 XGBoost 模型的目标函数对 7 月 26 日的交通流进行预测，并将预测结果与真实值进行比较，如图 9 所示。

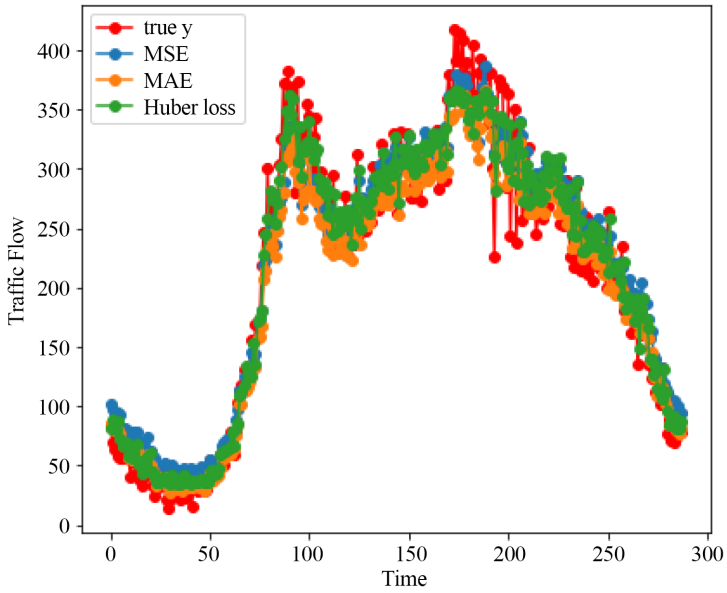

Figure 9. Comparison of different objective functions
图 9. 基于不同目标函数的 XGBoost 模型结果对比图

图 9 中红色曲线为 7 月 26 日实际的交通流数值，蓝色曲线表示基于均方误差的 XGBoost 模型的预测值，橙色曲线表示基于平均绝对误差的 XGBoost 模型的预测值，绿色曲线表示基于 Huber 损失的 XGBoost 模型的预测值。从图中可以看出，绿色曲线和红色曲线的重合度最高，这表示本文提出的基于 Huber 损失的 XGBoost 模型能够很好地拟合交通流数据，且预测精度非常高。

为直观获取各模型预测精度的差异，采用 5.1 节提出的性能评价指标对各模型预测结果进行评价，如表 2 所示。

Table 2. Comparison of different objective functions
表 2. 基于不同目标函数的 XGBoost 模型结果对比

目标函数	RMSE	MAE	R^2
Huber 损失	23.38	17.99	0.958
均方误差	24.31	18.20	0.954
平均绝对误差	26.55	20.24	0.946

由表 2 可知，将 Huber 损失作为目标函数的 XGBoost 模型所得到的均方根误差、绝对平均误差以及 R^2 均优于将均方误差损失作为目标函数的 XGBoost 模型与将平均绝对误差损失作为目标函数的 XGBoost 模型。将 Huber 损失作为目标函数的 XGBoost 模型误差最小，预测精度最高，预测效果是最佳的。

为评价基于 Huber 损失的 XGBoost 模型的预测性能，同时采用支持向量回归模型、梯度提升回归模型进行预测，并将预测结果与基于 Huber 损失的 XGBoost 模型的预测结果及真实值进行比较，如图 10 所示。

图 10 中蓝色曲线为 7 月 26 日实际的交通流数值，红色曲线表示基于 Huber 损失的 XGBoost 模型的预测值，橙色曲线表示梯度提升回归模型的预测值，绿色曲线表示支持向量回归模型的预测值。从图中可以看出，红色曲线和蓝色曲线的重合度最高，这表示本文提出的基于 Huber 损失的 XGBoost 模型能够更好地拟合交通流数据。

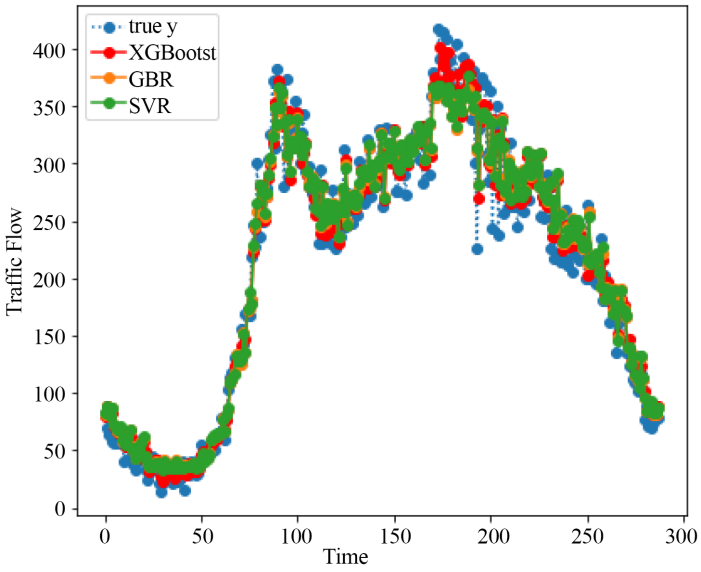

Figure 10. Comparison of different models
图 10. 不同模型结果对比图

为直观获取各模型预测精度的差异，采用 5.1 节提出的性能评价指标对各模型预测结果进行评价，如表 3 所示。

Table 3. Comparison of different model results
表 3. 不同模型结果对比

选用模型	RMSE	MAE	R^2
极端梯度提升模型(XGBoost)	23.38	17.99	0.958
支持向量机回归模型(SVR)	26.55	20.24	0.946
梯度提升回归模型(GBRT)	26.49	19.61	0.947

由表 3 可知，基于 Huber 损失的 XGBoost 模型所得到的均方根误差、绝对平均误差以及 R^2 均优于梯度提升回归模型和支持向量机回归模型，其误差最小，预测精度最高，预测效果是最佳的。因此，本文中提出的基于 Huber 的 XGBoost 模型能够从数据本身的特性出发，来对数据进行深层的探索，模型的建立也更加科学合理。

6. 结语

本文建立了一种基于 Huber 损失的极端梯度上升(Extreme Gradient Boosting, XGBoost)短时交通流预测模型，实验表明该模型优于基于均方误差损失的极端梯度上升模型以及基于平均绝对误差损失的极端梯度上升模型。同时，该模型较梯度提升回归模型、支持向量机回归模型具有更高的预测精度，各误差指标小，且模型训练时间短，符合短时交通流预测所要求的时效性，具有一定应用价值。需要指出的是，本文未考虑交通流的空间特性，下一阶段仍需研究多因素影响下的交通流预测，进一步加强模型鲁棒性，使其适用于复杂情形下的交通流预测。

参考文献

[1] 李敏, 黄迟. 集成学习下的短期交通流预测[J]. 济南大学学报(自然科学版), 2019, 33(5): 390-395.

[2] Ahmed, M.S. and Cook, A.R. (1979) Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Technique. *Transportation Research Board*, **722**, 1-9.

[3] Li, Y., Xiao, J., *et al.* (2016) Multiple Measures-Based Chaotic Time Series for Traffic Flow Prediction Based on Bayesian Theory. *Nonlinear Dynamics*, **85**, 179-194. https://doi.org/10.1007/s11071-016-2677-5

[4] Wei, H., Cheng, Z., Sotelo, M.A., *et al*. (2017) Short-Term Vessel Traffic Flow Forecasting by Using an Improved Kalman Model. *Cluster Computing*, No. 10, 1-10.

[5] Guo, J.H. and Williams, B.M. (2010) Real-Time Short-Term Traffic Speed Level Forecasting and Uncertainty Quantification using Layered Kalman Filters. *Transportation Research Record: Journal of the Transportation Research Board*, **2175**, 28-37. https://doi.org/10.3141/2175-04

[6] 李晓磊, 肖进丽, 刘明俊. 基于 SARIMA 模型的船舶交通流量预测研究[J]. 武汉理工大学学报(交通科学与工程版), 2017, 41(2): 329-332.

[7] 陆化普, 孙智源, 屈闻聪. 基于时空模型的交通流故障数据修正方法[J]. 交通运输工程学报, 2015, 15(6): 92-100.

[8] 钟颖, 邵毅明, 吴文文. 基于 XGBoost 的短时交通流预测模型[J]. 科学技术与工程, 2019, 19(30): 338-342.

[9] 苏美红, 张海. 基于不同损失函数的模型选择和正则化学习方法[J]. 纺织高校基础科学学报, 2014, 27(4): 464.

[10] 叶景, 李丽娟, 唐臻旭. 基于 CNN-XGBoost 的短时交通流预测[J]. 计算机工程与设计, 2020, 41(4): 1081-1086.