学号：23724590

《机器学习》
课程论文

基于 XGBoost 算法的短期交通流预测

姓　　名　郎志远

学　　号　23724590

论文评分

2024 年 1 月 29 日

目 录

基于 XGBoost 算法的短期交通流预测

[摘要]本文聚焦于解决城市交通拥堵问题，旨在通过精确的短期交通流预测来优化交通管理。为此，本文建立了一个基于 XGBoost 算法的预测模型，该模型专门针对交通流量的短期变化进行预测。通过深入分析交通流数据的时间特性，模型能够有效捕捉交通流量的周期性和趋势，从而为实时交通流量预测提供了一个高效且准确的解决方案。这一研究不仅有助于缓解城市交通拥堵，提高道路运输效率，而且为城市交通规划和智能交通系统的发展提供了科学依据。
[关键字]XGBoost，交通流预测，特征分析

Short-term traffic flow prediction based on XGBoost algorithm

Abstract: This paper focuses on solving the problem of urban traffic congestion and aims to optimize traffic management through accurate short-term traffic flow prediction. To solve this problem, this paper establishes a prediction model based on XGBoost algorithm, which is specifically designed to predict short-term changes in traffic flow. By deeply analyzing the temporal characteristics of traffic flow data, the model can effectively capture the periodicity and trend of traffic flow, thus providing an efficient and accurate solution for real-time traffic flow prediction. This research not only helps to alleviate urban traffic congestion and improve the efficiency of road transportation, but also provides a scientific basis for the development of urban transportation planning and intelligent transportation systems.

Key words: XGBoost; Traffic flow prediction; Feature analysis

1、引言

随着经济的快速发展，城市交通需求不断增长，但交通拥堵问题也随之加剧，导致车辆行驶速度下降、行程时间延长、尾气排放增加以及出行成本上升，这些问题已成为制约城市发展的关键因素。[1]因此，短期交通流预测对于构建智能交通系统和推动城市可持续发展具有至关重要的作用。短期交通流预测旨在利用历史数据预测未来某一时段的交通流量，这不仅有助于缓解交通拥堵，提升运输效率，而且为城市交通规划提供了科学依据。自 20 世纪 60 年代以来，学者们已经开发了多种模型来应对短期交通流预测的挑战，包括传统的数理统计模型、非线性理论模型以及新兴的人工智能模型。这些模型涵盖了时间序列分析[2]、卡尔曼滤波[3]、自回归模型和傅里叶变换[4]等。随着人工智能技术的进

步，神经网络和支持向量机等机器学习模型也被广泛应用于交通流预测，尽管它们在处理复杂非线性问题方面表现出色，但也存在收敛速度慢和对缺失数据敏感等局限性。为了克服这些挑战，本研究采用了 XGBoost 算法这一种基于梯度提升的决策树算法，通过迭代优化每棵树来提高整体模型的性能。XGBoost 通过在目标函数中引入正则项和列抽样技术，有效防止了过拟合，并显著提升了运算效率。

2、XGBoost 原理

将包含 m 个特征、容量为 n 的数据集记为$D = \{(x_i, y_i | x_i \in R^m, y_i \in R, |D| = n)\}$，所有 CART 树的集合记为$F = \{f(x) = w_{q(x)}, q: R^m \to T, w \in R^T\}$。其中，q 代表样本映射到相应的叶子节点的决策规则，T 代表一棵树的叶子节点数量，w 代表叶子节点的得分。f 代表 CART 树，包括树的结构 q 和叶子节点的得分 w。基于 XGBoost 算法的y_i的预测值可以表示为公式(1)。

$$\check{y} = \varphi(x_i) = \sum_{k=1}^{k} f_k(x_i) \tag{1}$$

其中$f_k \in F$，K 为 CART 树的数量。XGBoost 算法在每一次模型训练时保留前面 t-1 轮的预测不变，加入新函数f_t到模型中，$\check{y}_i^t = \check{y}_i^{t-1} + f_t(x_i)$为第 i 个样本在第 t 次模型训练时的预测结果。假设基学习器的误差相互独立，XGBoost 算法的学习目标是找到f_t，最小化目标函数，其计算如公式(2)和公式(3)所示。

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \check{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) + constant \tag{2}$$

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3}$$

其中，$l(\cdot,\cdot)$是训练误差，描述预测值与真实值之间差异的损失。$\Omega(\cdot)$ 是模型复杂度的正则项惩罚函数，γ 是复杂度参数，λ 是一个固定系数。

XGBoost 算法采用贪心算法[5]从根节点开始，递归地选择树结构的最优特征，据此特征对训练数据进 行分割。假设I_L和I_R分别是分割点左边和右边的样本集，$I = I_L \cup I_R$。计算每个分割方案的信息增益，信息增益最大的分割为该节点的最优分割，其计算如公式(4)所示。

$$L_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I_L} h_i + \lambda}\right] - \gamma \tag{4}$$

其中$I_j = \{i | q(x_i) = j\}$为节点 j 上的样本集合，$g_i = \partial_{\check{y}_i^{t-1}} l(y_i, \check{y}_i^{t-1})$，$h_i = \partial^2_{\check{y}_i^{t-1}} l(y_i, \check{y}_i^{t-1})$分别为训练误差的一阶和二阶梯度统计量。$\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda}$、$\frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_L} h_i + \lambda}$、$\frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I_L} h_i + \lambda}$分别为左子树分数、右子树分数、不分割时的分数，$\gamma$ 为加入新叶子节

点引入的复杂度代价,当L_{split}<0 时,放弃分割。集成学习的关键是如何生成好而不同的个体学习器,个体学习器准确性越高、多样性越大,则集成效果越好。相对于 LR、NB 等稳定分类模型,树模型是一种对样本扰动比较敏感的不稳定分类模型。决策树因为其简单直观,具有很强的可解释性,常被作为集成学习的个体学习器。相比于 ID3 和 C4.5,CART[6] 采用二元递归划分方法构建二叉树,分裂特征可以重复使用,既可以用于分类也可以用于回归。基于集成学习的优良性,本文在模型底层选用以 CART 树为基学习器的 XGBoost 算法建立评分预测模型。

3、基于 XGBoost 算法的交通流预测模型

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。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。

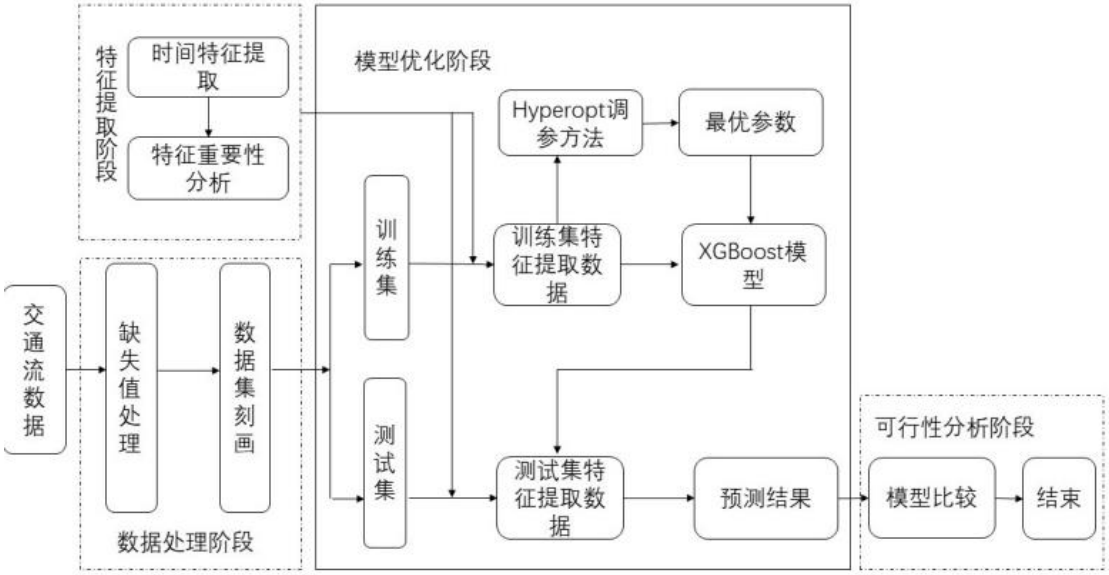

图 1 XGBoost 交通流预测模型

4、实例分析

本文使用的数据集来自 Kaggle(Traffic Flow)，这个数据集中的数据根据不同的路口编号为文件命名，文件中每五分钟记录一组数据，一个完整的文件的时间跨度约在 60 天左右。

4.1 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 5min 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和某一天数据进行考察，分别如图 2、图 3 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

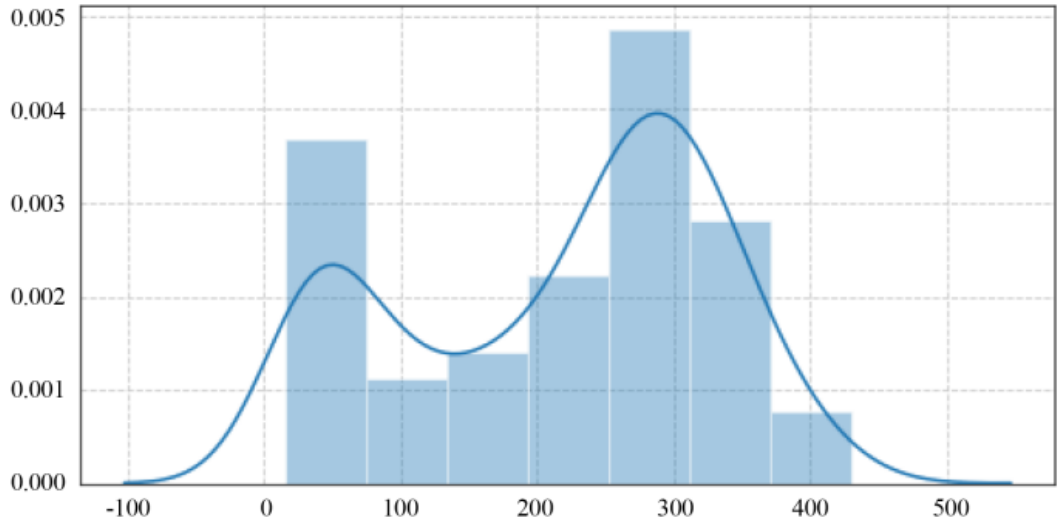

图 2 一天交通流数据分布

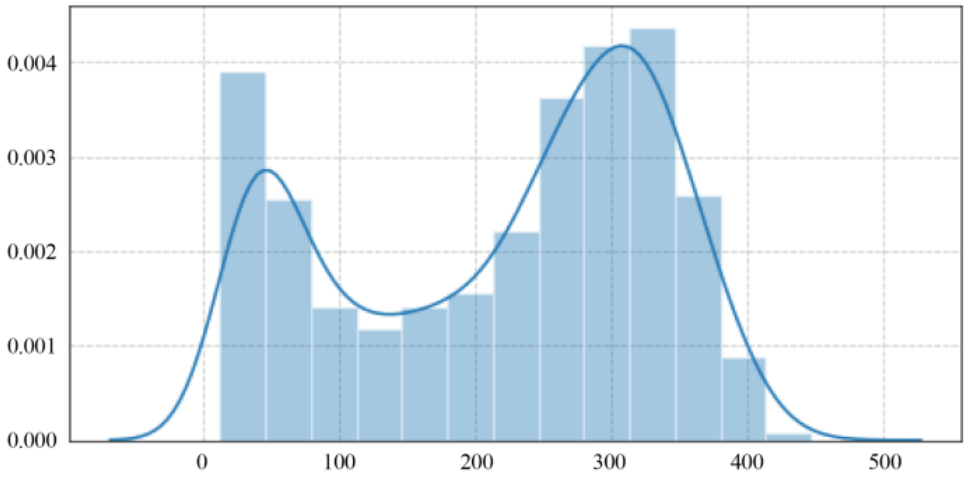

图 3 一周交通流数据分布

4.2 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 4 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)。

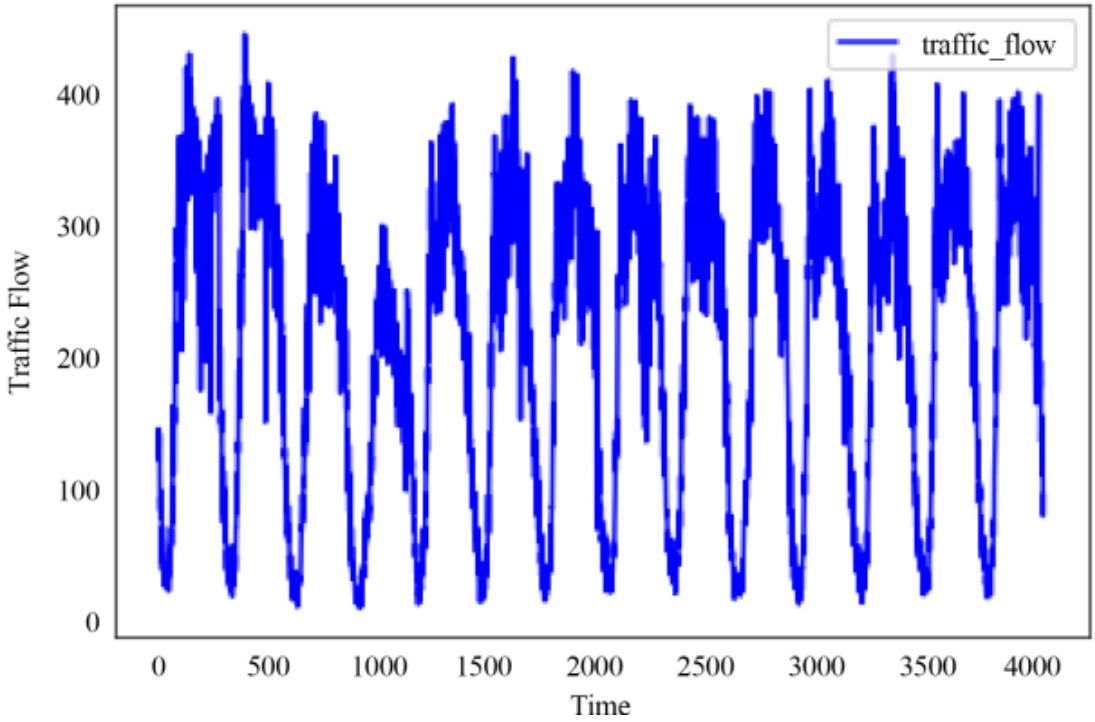

图 4 两周交通流数据图

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 5min 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 5 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 5min 的状态特征)。

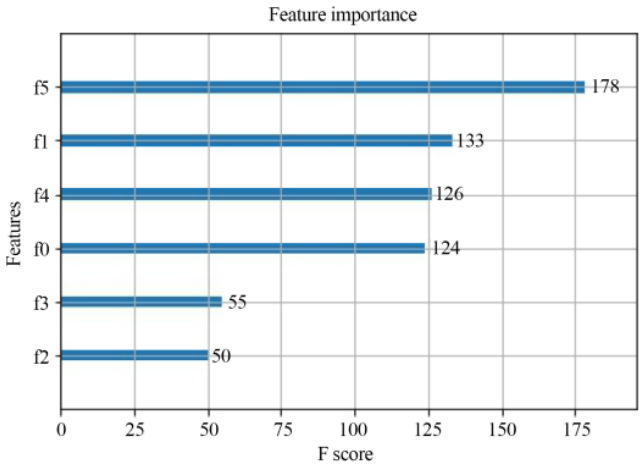

图 5 特征重要性分析图

由图 5 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。

4.3 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。

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**	gamm	**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。

4.4 模型预测及结果分析

为评估模型的效果，对未来 10 天的交通流进行预测，如图 6 所示。

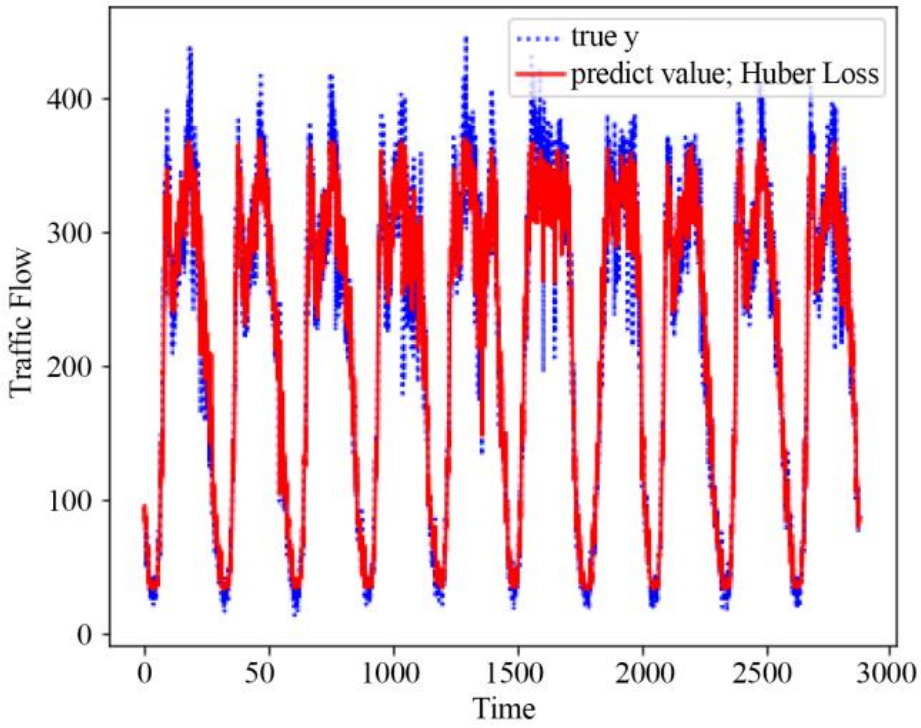

图 6　XGBoost 模型十天的预测图

图 6 中蓝色曲线为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 7 所示。图 7 中蓝色曲线为随机抽取的 第三天的交通流实际值，红色曲线为模型预测值。

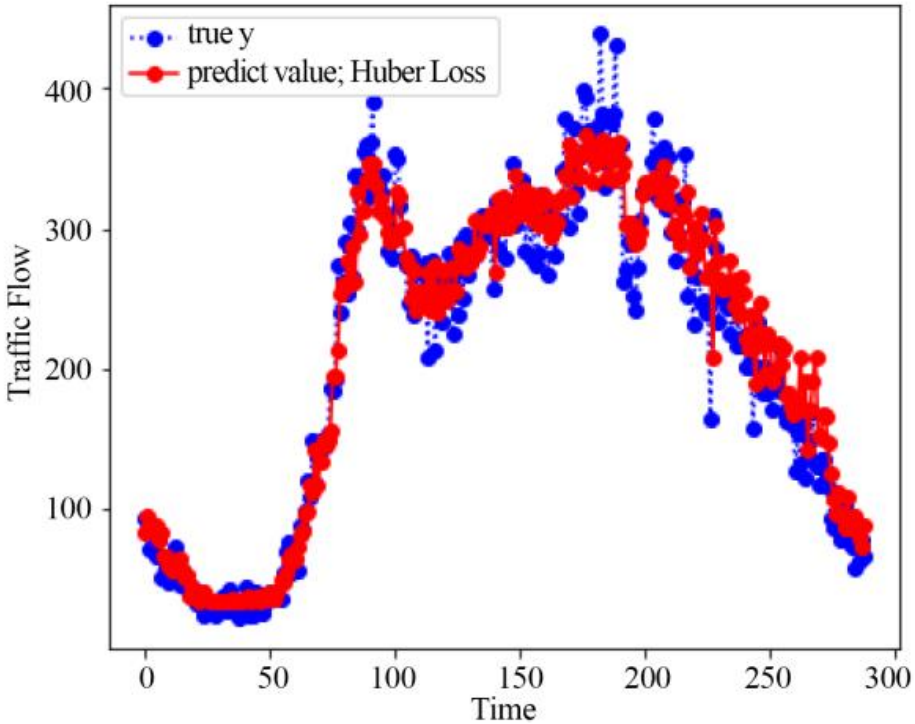

图 7　XGBoost 模型的单天预测图

为评价基于 XGBoost 的模型的预测性能，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 XGBoost 的模型的预测结果及真实值进行比较，如图 8 所示。

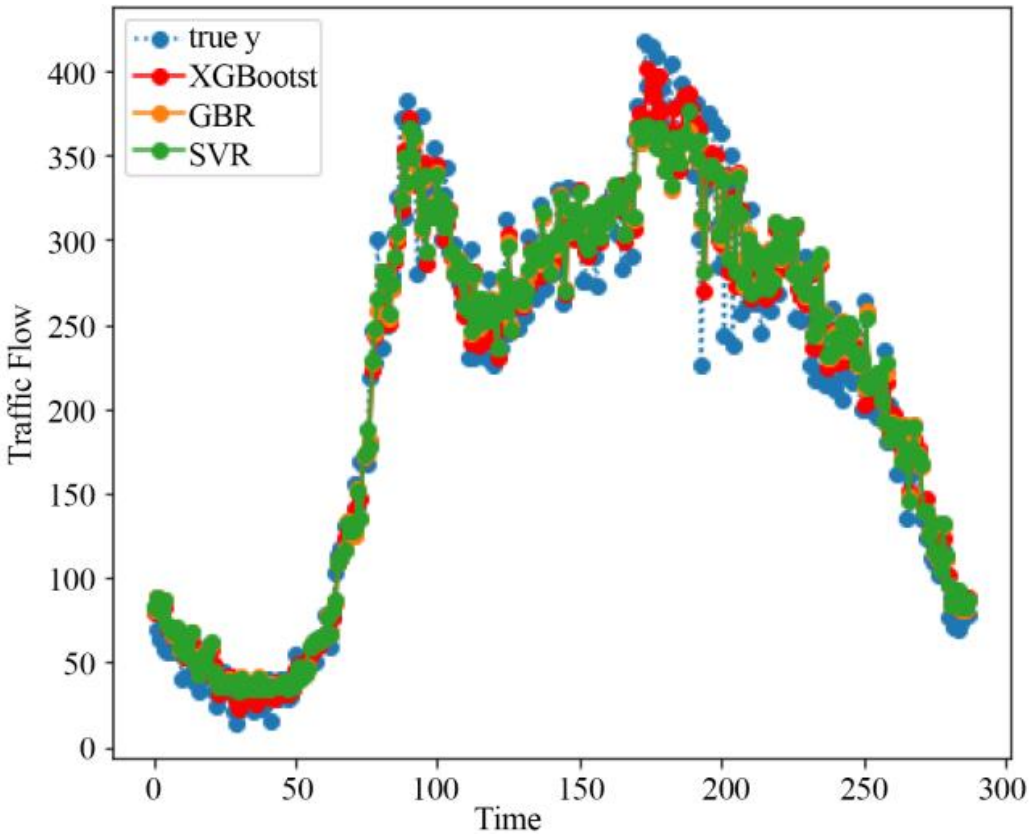

图 8 不同模型结果对比图

5、总结与展望

本研究提出了一种新的短期交通流量预测模型，该模型采用一种改进的梯度提升技术提高预测性能。实验结果表明，相本模型展较优秀优的预测效果。此外，与梯度提升回归和支持向量机回归模型相比，本模型在预测精度上也有所提升，同时训练时间更短，满足了短期交通流量预测对时效性的要求，显示出其实际应用潜力。但本研究尚未涉及交通流量的空间分布特性，未来的工作将需要探索在多种因素影响下如何提高模型的预测准确性和鲁棒性，以适应更复杂的交通流量预测场景[7]。

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] 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.

[4] 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.
[5] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32.
[6] 应维云. 随机森林方法及其在客户流失预测中的应用研究[J]. 管理评论，2012，24(2): 140-145. (Ying Weiyun. The Research on Random Forests and the Application in Customer Churn Prediction[J]. Management Review, 2012, 24(2): 140-145.)
[7] 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.