Multi-model Hybrid Traffic Flow Forecast Algorithm Based on Multivariate Data

Jie Zhou, Yuling Sun(✉), and Liang He

East China Normal University, Shanghai 200062, China
jzhou@ica.stc.sh.cn, {ylsun,lhe}@cs.ecnu.edu.cn

Abstract. Traffic flow forecast is a fine-grained task in urban intelligent transportation systems. Accurate traffic flow forecast can effectively support the development of intelligent transportation systems, reduce congestion, and improve the quality of residents' travel. The forecast of traffic flow is affected by many random factors such as weather, holidays and seasons. It has a certain degree of randomness and uncertainty, which makes the traditional single model prediction result extremely unstable, and the consideration of random factors is incomplete. As a result, the final forecast results are quite different from the actual situation. To address this problem, this paper proposes a multi-model hybrid traffic flow forecast algorithm based on multivariate data, which considers various random factors from multiple aspects, and captures different features through multiple models to improve the accuracy. The experiments on the dataset of KDD CUP 2017 demonstrate the effectiveness of our approach.

Keywords: Traffic flow forecast · Multivariate data · Multi-model-based

1 Introduction

The continuous construction of digital cities and the rapid development of various high-tech have spawned the birth of various smart city construction concepts [1, 2]. Among them, smart transportation is the inevitable trend of the development of traditional transportation system under the background of informatization [3–6]. Smart transportation refers to the full use of modern electronic information technology such as Internet of Things and mobile Internet in the transportation field, collecting various types of traffic information through analyzing, mining and applying data processing technologies such as data modeling and data mining [3, 5, 7, 8], in order to realize the systematic and real-time nature of intelligent transportation, and enhance the interactivity of information exchange and the extensiveness of services.

The idea of intelligent transportation provides a revolutionary solution to many problems in traditional transportation systems. Taking the congestion problem of expressways as an example, the congestion problem of high-speed toll stations has always been a pain point in traditional transportation networks. In recent years, with the exponential growth of private cars, congestion of toll stations has become the norm of highways, greatly reducing the high-speed travel experience of residents. The construction and development of the smart transportation system provides an innovative solution to this problem. The development of various Internet of Things and mobile

Y. Sun et al. (Eds.): ChineseCSCW 2018, CCIS 917, pp. 188–200, 2019.
https://doi.org/10.1007/978-981-13-3044-5_14

Internet devices, such as GPS-based Google Map, Waze, Baidu Map, and Gaode Map, provides a comprehensive and real-time way for crowdsourced traffic data collection; it is possible to effectively analyze and predict future high-speed traffic conditions and Estimated Time of Arrival (ETA) by constructing a variety of targeted data models based on these real-time data. On the one hand, it can provide users with real-time travel guides, which can effectively avoid the peak traffic and improve the travel experience. On the other hand, it also provides data basis for the real-time effective traffic supervision program design of the traffic control department, so that the traffic control department can real-time, effective staffing and traffic grooming based on the forecast results to avoid large-scale congestion.

Traffic flow and ETA prediction are the basic functions of the intelligent transportation system, and also the research focus and difficulty in the field of urban computing and social computing [7, 8]. At present, a large number of domestic and foreign researchers have studied this problem and proposed a series of algorithms for traffic flow prediction [9], such as historical trend method [10], nonparametric regression method [11], neural network based prediction methods [12], etc. These algorithm models can obtain better prediction results without emergency traffic events, special weather and special time, but the model results are extremely unstable in the case of many interference factors. However, the actual high-speed traffic speed will be affected by various random factors such as weather conditions, holidays, and commuting peaks, which further makes the final prediction results produced by such algorithms often differ greatly from the actual situation. At the same time, the algorithms such as support vector machine [13], KNN [14], ARIMA prediction algorithm [15], XGBoost [16] are also widely used in traffic flow prediction problems. However, actual high-speed traffic data is a typical multivariate data, and the existing methods tend to focus on single-model prediction. There are various advantages and disadvantages that vary from method to method, and it is difficult to comprehensively mine the characteristics of various aspects of traffic data.

In response to the current research situation, this paper proposes a multi-model-based traffic flow forecast algorithm based on multivariate data. The algorithm models various types of data based on different model angles from multivariate data, and then predicts real-time traffic flow through a combination of multiple model advantages. The algorithm proposed in this paper effectively combines the advantages of various models, and avoids the shortcomings of different models, and the effect is greatly improved.

2 Problem Description

2.1 Problem Definition

Traffic flow refers to the number of vehicles passing through a certain monitoring point or section within a unit time [9]. Affected by a variety of random factors such as weather, holidays, time, etc., traffic volume often has a large degree of suddenness and randomness. According to the prediction time span, the forecast of traffic flow can be divided into long-term forecast, medium-term forecast and short-term forecast: long-

term forecast is generally based on the year, medium-term forecast is generally based on month, day and hour, and the forecasting unit short-term forecast is generally 5–15 min; according to the different forecasting objects, the traffic flow forecast can be further divided into the traffic flow forecast of the road section and the forecast of the monitoring point [9]. This paper focuses on the short-term traffic forecast problem for toll stations.

The research data comes from 1 month's traffic data of a high-speed road section, involving the inflow and outflow historical data of the three stations of the toll station 1, the toll station 2 and the toll station 3, and the road network topology of the target area (e.g. Figs. 1, 2 and 3), vehicle trajectory, toll station historical traffic volume, weather data, etc. We conduct model training based on historical data from last month to predict traffic conditions at specific peak hours in the next month. As shown in Fig. 4, the specific research question can be described as how to design the algorithm and model based on the historical data of the previous month. Based on the model, the traffic data of the green time slot is given in the next month (06:00–08:00, 16:00–18:00), the traffic volume in the time slot of the red mark (08:00–10:00, 17:00–19:00) is predicted in each 20 min. Which means to predict the inflow and outflow of the toll station 1, the toll booth 2, and the toll booth 3. Among them, the toll gate 2 is a one-way toll gate that can only be entered, and the other ports are two-way toll gates that can enter and exit.

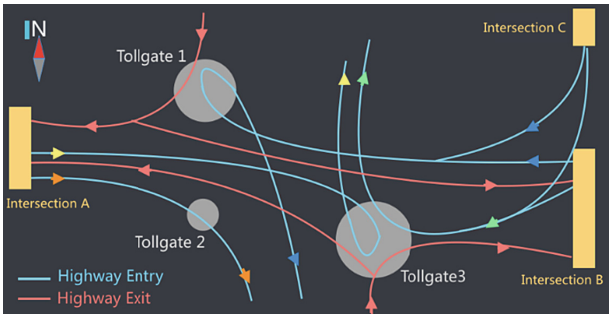

Fig. 1. Road network topology of target area

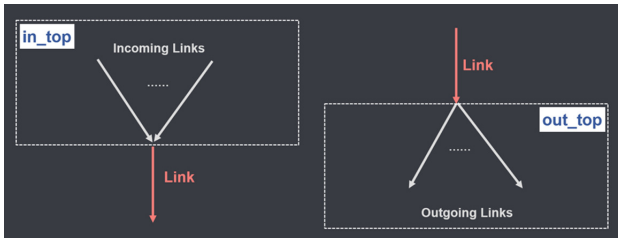

Fig. 2. In_top and Out_top for a road link

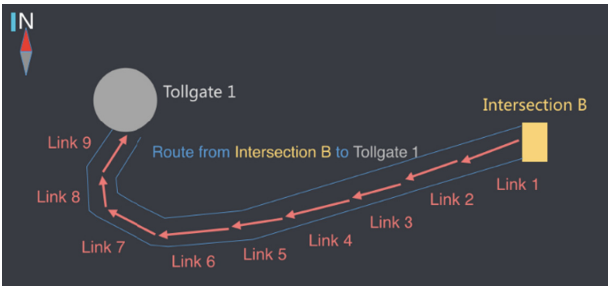

Fig. 3. Link sequence for the route from intersection B to tollgate 1

Fig. 4. Time windows for traffic prediction (Color figure online)

2.2 Data Description

In traffic flow forecasting, road network traffic information is typical multivariate information, mainly reflected in different sources, different environments, different spatial locations, and so on. In the source, it includes vehicle flow information, vehicle attribute information, weather information, and so on; in spatial location, the information comes from different toll stations, different areas, and so on; in data collection, there are also different sources, different sensors such as ETC, video and GPS. In order to improve the correctness and reliability of traffic information, it is necessary to analyze and process data such as traffic and weather from multiple sources. The multi-transportation data involved in this paper mainly includes the following parts:

- Road data, including the length, width, number of ramps, ramp widths, links to toll stations, toll stations, and so on;
- Vehicle data, including the time the vehicle entered the route, the vehicle's travel path, vehicle capacity and vehicle type, and so on;
- Vehicle trajectory data, including vehicles, toll stations, intersections, entry path times, travel times, and so on;
- Weather data, including air pressure, wind speed, wind direction, temperature, humidity and precipitation in the selected area.

The algorithm will use this data for feature extraction and model training, so that the multivariate data can be reasonably coordinated and the useful information can be fully integrated.

3 Multi-model Hybrid Traffic Flow Forecast Algorithm Based on Multivariate Data

The multi-model hybrid traffic flow forecast algorithm proposed in this paper includes three steps: data analysis and preprocessing, feature extraction of multivariate data, and multi-model fusion. This chapter will introduce the proposed algorithm with examples.

3.1 Data Analysis and Data Preprocessing

Data analysis and preprocessing are mainly to obtain the initial awareness of the data, and provide the basis for the subsequent feature extraction and mining. In this paper, we mainly analyze and pre-process the traffic flow statistics of each site at a fixed time period, the time-continuous traffic flow of a single site, and the traffic flow of a single site on random extraction dates. The analysis results are shown in Figs. 5, 6 and 7, specifically:

– Statistical analysis of traffic flow at each station for a fixed period of time: We analyze the time distribution of traffic flow from 6 to 10 o'clock every day at a toll station. The analysis results are shown in Fig. 5. The results mainly confirm the abnormal factors such as holidays have strong interference effects on traffic flow. As can be seen from the figure, since the traffic volume has changed suddenly during the National Day from October 1 to October 7, we define this part of the data as an outlier. In the final training data, the algorithm deletes the data during this period to reduce noise.

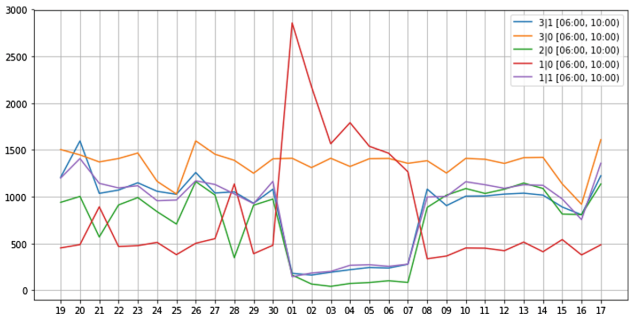

Fig. 5. Traffic statistics for each station from 6:00–10:00 daily

– Time-continuous traffic analysis at a single site: We performed a time-continuous traffic analysis on toll station 3-0. The results are shown in Fig. 6. The results

confirm the periodic characteristics of the traffic data, such as daily periodicity, weekly periodicity, and so on. It can be seen from the figure that the flow curve has a relatively obvious periodicity, and the traffic flow distribution of the station is also relatively close, which provides a basis for the feature extraction of the model.

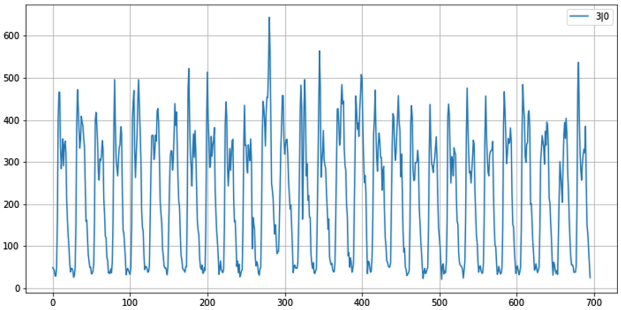

Fig. 6. Traffic flow statistics for every 20 min imported from toll station No. 3

– Traffic analysis of individual sites on a random pick date: We randomly analyzed traffic flow data of four days at toll station 3-0 (September 20, October 11, September 27, and October 21). The analysis results are shown in Fig. 7. The results mainly confirm the interference effect of the daily time period characteristics on the vehicle traffic flow. It can be seen from the figure that the traffic flow distribution from 6:00 to 10:00 in the four days of random selection is basically the same: the traffic volume at 6 o'clock starts to increase slowly, and the highest peak of traffic flow is reached from 8:00 to 9:00, what we usually call the peak hours of work. This is also very consistent with people's travel rules.

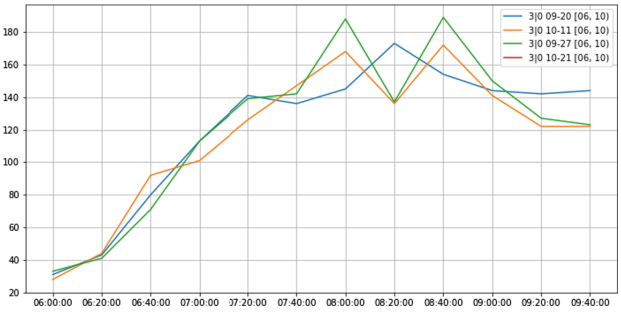

Fig. 7. Traffic statistics for the In_top of the No. 3 toll station from 6:00 to 10:00 of September 20, October 11, September 27, and October 21 four days

3.2 Feature Extraction of Multivariate Data

The forecast of traffic flow is not only related to people's work schedule, but also related to weather, road structure, emergencies, and so on. For this reason, multiple data needs to be considered in extracting features. Through the data analysis of the previous step, we confirmed the interference effects of random factors such as holidays, weather, and time periods on traffic flow. For example, in sunny weather, more people would like to go out and play, and holidays often cause high traffic, 8:00–10:00 am and 5:00–7:00 pm is the peak period of commuting, and so no, and the periodicity of traffic flow is also found. Therefore, we propose a multi-model-based forecast method based on multivariate data. The multivariate data here refers to traffic flow data that combines various factors such as weather, holidays, time periods, dates, etc. The second step of the algorithm is feature extraction from multivariate data. Based on the above analysis, we mainly extracted the following features:

- Toll_id;
- Direction;
- Week;
- Weather characteristics;
- Traffic flow every 20 min for the first two hours;
- Statistics of two-hour traffic (including maximum, minimum, median, average, variance, and so on);
- Statistics of traffic flow per hour (including maximum, minimum, median, average, variance, and so on);
- Special date, including whether it is a working day, whether it is the first day of work, whether it is the first day of vacation, whether it is the first day of work, whether it is the last day of work, whether it is a holiday, and so on;

For discrete variables, we convert them to one-hot vector processing.

3.3 Multi-model-Based Traffic Flow Forecast Algorithm

The traffic flow prediction algorithm proposed in this paper is regarded as a regression problem and uses historical data to establish a model to predict the future traffic flow. To enable us to capture the characteristics of more dimensions, we used three different machine learning models: LightGBM, XGBoost, GBRT, and selected three different strategies for model training. In addition to the overall training prediction, the algorithm also uses the tollgate-direction-differentiated prediction scheme and the combination of the predictions of the morning and the afternoon to carry out the model training. Finally, the algorithm performs weighted ensemble on the three prediction schemes of the three models to obtain the final forecast results, see Fig. 8 for the specific process.

An overview of the three models is as follows:

- LightGBM: Microsoft's gradient boosting framework uses a learning algorithm based on decision tree, which has the advantages of fast training efficiency, low memory usage, high accuracy, support for parallel learning, and processing of large-scale data.

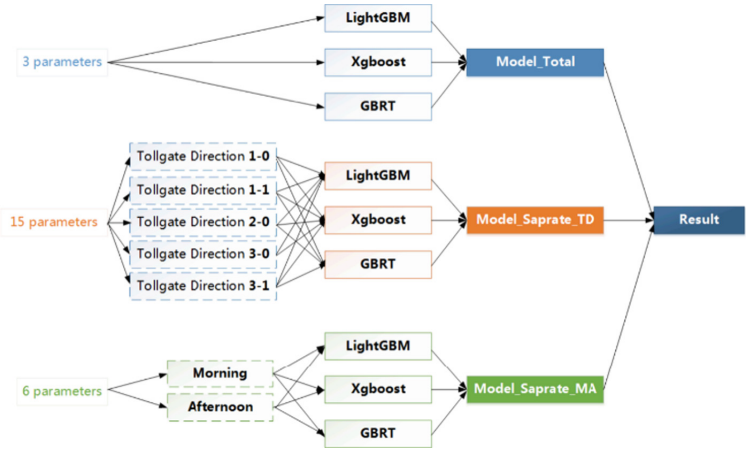

Fig. 8. The framework of ensemble

- XGboost: A tree learning algorithm for processing sparse data. Using approximate tree learning theory to use fractional descriptions, a reasonable weighting weight can be given to each training strength. Parallel and distributed design makes the algorithm very fast learning and modeling speeds, end-to-end systems can process large amounts of data with minimal cluster resources.
- GBRT: A promotion of boosting, which can deal with regression problems, can directly deal with the characteristics of mixed types, and has better robustness to outliers in output space.

We will explain in detail the multi-model fusion traffic flow prediction algorithm flow proposed in this paper below.

For a data set with n samples and m features $\mathrm{D} = \{(x_i, y_i)\}(|\mathrm{D}| = \mathrm{n}, x_i \in R^m, y_i \in \mathrm{R})$, a model containing K trees can be expressed as:

$$\tilde{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \epsilon F \tag{1}$$

where $\mathrm{F} = \{\mathrm{f(x)} = \mathrm{w}_{q(x)}\}(\mathrm{q} : R^m \rightarrow \mathrm{T}, \mathrm{w} \in R^T)$ represents the function space composed of trees, q represents the mapping of x to leaf nodes, and w represents the weight of leaf nodes.

Loss function is defined as:

$$\mathrm{L}(\phi) = \sum_i^l (\tilde{y}_i, y_i) + \sum_k \Omega(f_k) \tag{2}$$

Where $\Omega(f) = \lambda T + \frac{1}{2}\lambda w^2$ is the regular term which aims to reduce the complexity of the model. T is the number of leaf nodes, and w is the weight of the leaf nodes.

The model uses the Taylor formula of the loss function to obtain the loss approximation representation at step t to find the optimal submodel. The formula is:

$$L^{(t)} \simeq \sum_{i=1}^{n} \left[l\left(y_i, \tilde{y}^{t-1}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{3}$$

where $g_i = \partial_{\tilde{y}^{t-1}} l(y_i, \tilde{y}^{t-1})$, $h_i = \partial^2_{\tilde{y}^{t-1}} l(y_i, \tilde{y}^{t-1})$ represents the first derivative and the second derivative of the loss, respectively, after removing the constant:

$$\tilde{L}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{4}$$

The algorithm defines $\mathrm{I}_j = \{i | q(x_i) = j\}$ as the sample set of leaf node j. The above formula can be used to sum each leaf node:

$$\begin{aligned} \tilde{L}^{(t)} &= \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \sum_{i \in I_j} h_i + \lambda w_j^2 \right] + \gamma T \end{aligned} \tag{5}$$

For a given tree structure q(x), when the derivative of the above formula is 0, the minimum value can be obtained to obtain the value of the leaf node:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{6}$$

Furthermore, when the tree is q, the optimal value of the loss function is:

$$\tilde{L}^{(t)}(\mathrm{q}) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{7}$$

Since the complexity of traversing all possible tree structures to get the optimal loss is too high, the algorithm adopts a layer-by-layer construction method. That is, a root node is continuously split in the construction. The process of splitting is subject to the change of loss:

$$L_{split} = \frac{1}{2} \left[\frac{\left(\sum_{i \in I_L} g_i \right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i \right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i \right)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \tag{8}$$

After building the tree model, we can use this model for regression prediction.

The multi-model-based traffic flow forecast algorithm based on multivariate data uses three different models to build the model. Three different strategies are also selected for model training, as follows:

- Strategy 1: All the data of the toll stations are trained in a model, which can increase the amount of data, and learn the common characteristics of the traffic flow distribution of different toll stations, such as 6:00–8:00 in the morning. Traffic increases, but the characteristics of each site are weakened.
- Strategy 2: Different models are used for training in the morning and afternoon. The morning traffic is trained with morning data, while the afternoon traffic is trained with afternoon data, so that the distribution characteristics of the previous and afternoon predictions can be obtained separately, but This reduces the amount of data.
- Strategy 3: For each site, we use a separate model to train. In this way, we can learn the individual characteristics of each site well, and will not affect the distribution characteristics of the site because of the overall data. For example, the traffic volume of this site is much larger than other sites.

In addition, because of the small number of training samples, we use time window sliding to greatly increase the data sample.

4 Experiments and Analysis

4.1 Metric Function

In this paper, we choose Mean Absolute Percentage Error (MAPE) to evaluate the result. The MAPE for traffic flow forecast is defined as

$$MAPE = \frac{1}{C}\sum_{c=1}^{C}\left(\frac{1}{T}\sum_{t=1}^{T}\left|\frac{f_{ct}-p_{ct}}{f_{ct}}\right|\right) \tag{9}$$

where C is the number of tollgate-direction pairs (as aforementioned: 1-entry, 1-exit, 2-entry, 3-entry and 3-exit), T be the number of time windows in the testing period, and f_{ct} and p_{ct} be the actual and predicted traffic volume for a specific tollgate-direction pair c during time window t.

The classical regression loss function adopted by most machine learning package:

$$\mathrm{MSE} = \frac{1}{2}(f_{ct}-p_{ct})^2 \tag{10}$$

$$\mathrm{MAE} = |f_{ct}-p_{ct}| \tag{11}$$

We take a log-transform of target y, we can obtain a sample mathematical approximation:

$$|logp_{ct} - logf_{ct}| = \left|log\frac{p_{ct}}{f_{ct}}\right| = \left|log\left(1+\frac{p_{ct}-f_{ct}}{f_{ct}}\right)\right| \approx \left|\frac{p_{ct}-f_{ct}}{f_{ct}}\right| \tag{12}$$

Through the log-transform, the measure MSE and MAE become relatively measure close to the spirit of MAPE.

4.2 Experiment Analysis

In order to verify the effectiveness of our proposed algorithm, we use the data of KDD CUP 2017 to conduct experiments. Table 1 shows our experimental results on KDD CUP 2017 data. As can be seen from the table, our multi-model fusion strategy has a significant effect on improving the prediction results. First, from the results of models 1, 2, and 3, it can be seen that each model has different learning ability for the same data, and can capture different features, and performance also is different. When the results of the three models are merged, the results obtained by the model 4 are significantly better than those of the single model. The model 8 combined by models 5, 6, and 7 and the model 12 combined by models 9, 10, and 11 also can verify such conclusions. At the same time, we can also find from the data that the same model can learn different characteristics when using different strategies. Compared with models 1, 5 and 9, we find that the results of the three models are different. Finally, when we fuse the results of models 4, 8, and 12 to get model 13, the results of the model are improved again, and nearly 10% improvement is achieved compared to the single model. Therefore, it can be seen from the experimental results that the multi-model fusion strategy used in this paper can effectively improve the accuracy and robustness of prediction.

Table 1. The results of the experiment

#	Model	MAPE
1	lightgbm	0.1213
2	xgboost	0.1216
3	gbrt	0.1210
4	ensemble_3model (ensemble of model1, 2 and 3)	0.1187
5	lightgbm_separate_tollgate_direction	0.1154
6	xgboost_separate_tollgate_direction	0.1170
7	gbrt_separate_tollgate_direction	0.1149
8	ensemble_eparate_tollgate_direction (ensemble of model 5, 6 and 7)	0.1133
9	lightgbm_separate_morning_afternoon	0.1180
10	xgboost_separate_moring_afternoon	0.1181
11	gbrt_separate_moring_afternoon	0.1214
12	ensemble_separate_morning_afternoon	0.1160
13	ensemble_3ensemble (ensemble of mode 4, 8 and 12)	0.1125

In addition, in order to view the learning effect of the proposed model, we show the forecast results of the toll station 1-1 on October 26th in Fig. 9. As can be seen from the figure, the models 4, 8, and 12 are overall. The trends are consistent, but there are large differences, and the combined results can be a good synthesis of the characteristics of these models, making the results more stable and reliable. The experimental results show that the fusion model integrates the advantages of other single model predictions, with small errors and high prediction accuracy, and has great advantages in traffic flow

prediction tasks. In addition, we used this method to participate in the KDD CUP traffic flow prediction competition, and obtained the second price of the competition, which further verified the feasibility and effectiveness of the proposed method.

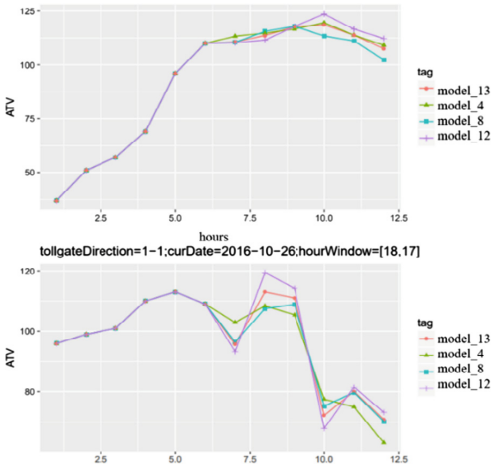

Fig. 9. One example of the forecast

5 Conclusion

In this paper, we propose a multi-model hybrid traffic flow forecast algorithm based on multivariate data. This approach comprehensively mines the characteristics of various aspects from multivariate data, models various types of data based on different model, and then predicts real-time traffic flow through a variety of model ensemble methods. The experimental results show that the approach can effectively combines the advantages of various models, and avoids the shortcoming of different models and the results obtains a greatly improvement.

References

1. Zhang, J., Zheng, Y., Qi, D.: Deep spatio-temporal residual networks for citywide crowd flows prediction. In: AAAI (2017)
2. Zhang, S., et al.: Effective and efficient: large-scale dynamic city express. IEEE Trans. Knowl. Data Eng. **28**(12), 3203–3217 (2016)
3. Zheng, Y.: Methodologies for cross-domain data fusion: an overview. IEEE Trans. Big Data **1**(1), 16–34 (2015)
4. Zheng, Y.: Trajectory data mining: an overview. ACM Trans. Intell. Syst. Technol. (TIST) **6** (3), 29 (2015)
5. Ma, S., Zheng, Y., Wolfson, O.: T-share: a large-scale dynamic taxi ridesharing service. In: ICDE (2013)

6. Ma, S., Zheng, Y., Wolfson, O.: Real-time city-scale taxi ridesharing. IEEE Trans. Knowl. Data Eng. **27**(7), 1782–1795 (2015)
7. Qiao, S., et al.: Short-term traffic flow forecast based on parallel long short-term memory neural network. In: 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS). IEEE (2017)
8. Yuan, P.C., Lin, X.X.: How long will the traffic flow time series keep efficacious to forecast the future? Physica A: Stat. Mech. Appl. **467**, 419–431 (2017)
9. Zhao, Z., Yang, Z.: Traffic flow forecast based on residual modification GM(1, 1) model. Comput. Sci. (2017)
10. Stephanedes, Y.J., Michalopoulos, P.G., Plum, R.A.: Improved estimation of traffic flow for real-time control (discussion and closure). Transportation Research Record (1981)
11. Wang, X., Zhicai, J., Miao, L., et al.: The application of nonparametric regressive algorithm for short-term traffic flow forecast. In: International Workshop on Education Technology and Computer Science, pp. 767–770. IEEE (2009)
12. Chan, K.Y., Dillon, T.S., Singh, J., et al.: Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and Levenberg–Marquardt algorithm. IEEE Trans. Intell. Transp. Syst. **13**(2), 644–654 (2012)
13. Rong, Y.U., Wang, G., Zheng, J., et al.: Urban road traffic condition pattern recognition based on support vector machine. J. Transp. Syst. Eng. Inf. Technol. **13**(1), 130–136 (2013)
14. Yu, B., Wu, S., Wang, M., et al.: K-nearest neighbor model of short-term traffic flow forecast. J. Traffic Transp. Eng. **12**(2), 109–115 (2012)