Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 2, No. 3, 2024

Research on the traffic flow measurement method of intelligent connected vehicles based on series and cascade technology

Chang Li

Gac Toyota Motor Co.,Ltd, Guangzhou, 510000, China

Abstract: At present, the major cities in our country have entered the development of smart cities, the utilization rate and prevalence rate of intelligent connected cars in smart cities have reached a new peak,to manage the real-time traffic flow of intelligent connected vehicles effectively in smart cities. This paper proposes to calculate the flow of intelligent connected vehicles based on series cascade technology. Firstly, this paper collects multi-source heterogeneous data based on multiple collectors; Secondly, a generative adversarial network is employed for data fusion, synthetic data, closely mirroring the original, is generated for handling unknown variables during model training. Next, an autoencoder model generates input data for feature representation. Lastly, CatBoost, a regression model that leverages a cascaded machine learning algorithm, is utilized to forecast the traffic flow of intelligent connected vehicles in smart cities. The results show that the method based on CASCaded machine learning has remarkable effect in predicting smart city traffic flow and alleviating urban road congestion.

Keywords: smart city, traffic prediction, CatBoost algorithm, series cascade

1. Introduction

To promote the coordinated development of smart connected cars and smart cities, government departments such as the Ministry of Housing and Urban-Rural Development and the Ministry of Industry and Information Technology have launched several pilot city construction works. These pilot cities will explore new models and paths for the coordinated development of smart connected cars and smart cities. As an important part of the smart city transportation system, smart connected cars can improve traffic efficiency, reduce traffic accidents and ease traffic congestion through functions such as autonomous driving and remote control. However, in the process of building a smart city, Internet of Things applications and multi-source heterogeneous data collectors with functions such as data sharing and remote control are growing at an exponential rate, which also brings the demand for large volume data fusion and analysis. Due to the huge amount of data generated, it is extremely difficult to directly process the raw data. How to collect data from different distributed nodes and store the data in the blockchain node for further analysis, so as to help smart connected cars better realize intelligent driving, is extremely critical for the construction of smart cities. Traditional data fusion technologies cannot handle the complexity of heterogeneous data, so innovative technologies such as blockchain and machine intelligence need to be introduced. Introducing a machine intelligence-based approach to data fusion can provide new solutions for extracting important features from data collected from blockchain nodes, while maintaining data integrity to protect user privacy. To sum up, this paper aims to explore the effectiveness of a smart machine-based approach in data fusion and smart city traffic prediction.

In the academic aspect, the data fusion method of the collector has been widely concerned and deeply studied by many scholars. Raykar[1] introduced a particle filter algorithm that utilizes an ensemble of random particles to approximate the probability distribution inherent in nonlinear recursive filtering techniques for data fusion. This approach has proven efficacy in forecasting traffic flow within intelligent, networked urban environments. Kaur[2] proposed a collector data fusion method based on recurrent neural networks and long short term memory networks, and predicted the results of invisible data in intelligent traffic in smart cities. Zafar[3] conceptualized a hybrid deep learning model premised on GRU-LSTM, which has been employed to amalgamate citywide data from diverse sources. The method processes multi-source heterogeneous data including road data and vehicle data through customized data pipelines for prediction research. The research results show that on the basis of data fusion, The accuracy of this method in predicting road traffic is 95%. Yan Li [4] et al. have proposed a nuanced deep learning approach that incorporates attribute feature units, melding external factors into convolutional neural networks, and discerns traffic flow by taking multiple factors into account.Research findings suggest that the incorporation of meteorological variables serves to diminish forecasting inaccuracies, and the model exhibits commendable performance in the context of extended traffic flow prediction for smart cities. It can be seen that it is an essential task to predict the traffic flow of intelligent connected vehicles in the process of smart city construction.

In the field of business, technology based on machine intelligence provides a powerful solution for efficient and accurate data fusion in optical sensor networks, and promotes the development of AI-driven next-generation iot applications in Industry 4.0 and smart cities. Sensor nodes in optical networks produce data from multiple geographic locations and cover a wealth of information. These sensors gauge diverse environmental parameters and relay them to blockchain nodes via the optical network. Large amounts of data are generated in a continuous manner and can be processed to derive important information about future road traffic loads. The Internet of Things (IoV) seeks to use this data to propose energy-efficient solutions, alleviate traffic congestion in smart cities, optimise traffic flow and reduce carbon emissions. In addition to its application to road traffic forecasting, research on iot applications based on blockchain and machine intelligence algorithms can also be extended to other fields, such as predicting industrial energy consumption and traffic-induced pollution in smart cities.

To sum up, this paper proposes an intelligent connected vehicle traffic measurement method based on series and cascade technology to deal with the existing smart city intelligent connected vehicle traffic prediction accuracy and other problems. The combination of data fusion and CatBoost regression improves the accuracy and reliability of intelligent connected traffic prediction, and provides a theoretical basis to help smart cities optimize road networks, improve traffic flow, and create a more sustainable urban environment.

2. Research Methods

This research roadmap is divided into five steps. The first step involves collecting real-time data related to traffic conditions, including traffic flow sensors, occupancy sensors, weather sensors, etc.The second step, a deep learning model based on autoencoder and generative adversarial network technology is adopted for data processing. The third step entails data fusion, where the basic features of different data sets are combined to extract crucial information. The fourth step uses CatBoost cascade machine intelligence technology to predict the intelligent network traffic conditions, such as the intelligent network road traffic flow, the number of vehicles and the average vehicle speed. Figure 1 illustrates the research roadmap presented in this paper.

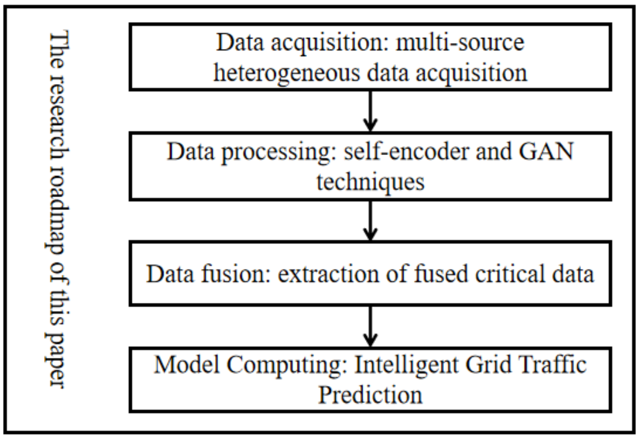

Fig.1 The research roadmap of this paper

2.1. Multi-source heterogeneous data acquisition

Multi-source heterogeneous data collection includes temperature and humidity data collection, as well as other types of data collection. Each data collector provides all the measurement data within the network over a specific period of time. Due to the characteristics of the network transmission by the collector, the process of data collection can be chaotic. Therefore, before the multi-source heterogeneous data summary, the original data should be pre-processed, the pre-processing steps include data cleaning, processing outliers, correcting error information, etc., to adjust all features to the same specification scale, filter out interference information, and ultimately form a unified standard and format matrix data. The mode of multi-source heterogeneous data acquisition is shown in Figure 2 below. These collectors can rapidly generate measurement data, which can then be analyzed using machine intelligence methods to calculate traffic condition flow.

Fig.2 Schematic diagram of multi-source heterogeneous data acquisition mode

In order to build and evaluate the collector data fusion model, the collected data needs to be divided into different subsets. In the process of data partitioning, try to avoid data inadvertently leaking into the training process of the model. This paper adopts the method of cross-validation to verify the data in the process of model training. After fusing the collector node data, it is necessary to check the data for balance and anomalies before classification. This helps improve the accuracy of the classifier and enhances the model's ability to learn from scarce data samples.

2.2. Data Processing

2.2.1. Autoencoder training

An autoencoder is a type of neural network that focuses on unsupervised learning tasks, capable of extracting key features from data and learning from specific environments. In this study, before providing the data to the encoder, data pre-processing is required. This process includes tasks such as data balancing, data normalization, and data de-anomaly, among others. The data after the pre-processing enters the coding layer, and then is processed by several hidden layers to gradually reduce the dimensionality. The hidden layer determines the compression degree of the data, and the neurons of the hidden layer regulate the data dimension of the hidden space.

2.2.2. GAN training

Training generative adversarial network (GAN) is a new neural network architecture technology. This technique is used to generate synthetic data similar to the modal distribution of the source collector, in order to predict traffic conditions on intelligent network roads. These synthetic data are then used to enhance the target collector modes and enrich the diversity and heterogeneity of the data set. The GAN framework has two core elements: a generator and a discriminator. The generator, as a neural network, is responsible for receiving random interference information and generating synthetic data samples that are highly similar to the source sensor data; The discriminator receives real data samples from the source modality and synthesized samples generated by the generator as inputs, and distinguishes authenticity through data comparison. When the generator receives interference information, it generates synthetic data that deviates from the source mode data.The generator is trained to enhance its performance by diminishing the discriminator's capacity to accurately differentiate between authentic and fabricated samples. Concurrently, the discriminator is trained to be better at distinguishing between actual and fake samples.

(1) Generator data

The generator takes random interference z as input and generates synthesized data samples, which are represented by G(z) in this paper. The generator attempts to learn the mapping from the input space (usually a multidimensional vector) to the data space (such as images, audio, etc.).

(2) Discriminator data

The discriminator takes synthesized data or real-time data as input variables and produces a single scalar value that proves the probability of the input data. As a binary classifier, the discriminator can effectively distinguish between real and fake samples.

The mathematical aspect representation: the discriminator is represented as $D: X \rightarrow [0,1]$, where X is the space of data samples and [0,1] represents the probability range.

2.3. Fusing data

Assuming that for each data collector modality i, the encoder can be represented as formula (1).

$$E_I: X_i \rightarrow Z_i \qquad (1)$$

Formula (1) analysis: X_i is the data sample space of data collector modality i, while Z_i is the latent spatial representation of modality i

The decoder can be expressed as formula (2).

$$D_I: Z_i \rightarrow X_i \qquad (2)$$

Formula (2) Parsing: D_I represents receiving latent spatial representations and reconstructing the original data samples of modality i

2.3.1. Expand the data

A generative adversarial network (GAN) is used to augment the data of the source collector modes, and an autoencoder is used to process the data of each collector.

When given a random interference vector z as input to the generator, the GAN is able to generate a composite sample with a similar distribution to the source collector mode data, denoted $G(z)$. The generator of a GAN is represented by formula (3).

$$G: Z \rightarrow X_s \qquad (3)$$

Formula (5) analysis: Z is the input interference information and the data sample space of the source mode.X_s

Discriminator D_s receives real data sample X_r from the source modality and synthetic sample X_f from the generator, and outputs a probability value indicating the likelihood that the sample is real.

2.3.2. Potential space

During autoencoder training, the latent space of each data collector modality can be obtained. The fusion function is used to express the latent space of n modalities $(z_1, z_2, \ldots, z_n)$ of the collector, which can be expressed as formula (4).

$$Z_1 \times Z_2, \cdots \times Z_n \rightarrow Z_f \qquad (4)$$

Formula (4) analysis: Z_f is the latent space after fusion. The fused latent space contains combined information from all Z_f collector modalities, which can capture the relationships and patterns between different modalities.

2.4. Model calculation

2.4.1. Introduction to CatBoost model

CatBoost model is a deeply improved version based on GBDT framework, which can be combined into a new set of category features by using the connection between features, greatly enriching the feature dimension of the model and improving the accuracy of the model. The basic model of CatBoost model adopts a symmetric tree, which can also prevent the model from overfitting.

2.4.2. Principles of CatBoost model

The most important principle of CatBoost model is the processing of feature values, which mainly includes two aspects: the first aspect is the processing of category features, and the second aspect is the recombination of features. After the recombination of feature values, the Gini coefficient of each segmentation scheme needs to be calculated. The segmentation with the smallest Gini coefficient is the best segmentation of the node, and the Gini coefficient calculation formula is shown in formula (5).

$$y_i = \varphi(x_i) = \sum_{k=1}^{k} f_k(x_i),\ fk \in F \qquad (5)$$

Where k is the number of basic tree models, F is the basic tree model. Based on this, assuming that the learning errors of the basic tree models are independent of each other, adding a new function F (t) to the model, the learning objective of the CatBoost model is to find $f_t(x)$ and minimize the objective function. The objective function calculation formula is shown in formula (6).

$$\hat{y}_i^{(t)} = y_i^{(t-1)} + f_t(x_i) \qquad (6)$$

The learning of the i-th sample in the t-round model needs to be combined with the results of the previous t-1 rounds to learn the current round model. Therefore, the objective function of the i-th sample during the t-round is shown in formula (7).

$$\begin{aligned} L^{(t)} &= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \Omega(f_t) \\ L^{(t)} &= \sum_{i=1}^{n} l\left(y_i, y^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \end{aligned} \qquad (7)$$

3. Research results

A series of standard performance indicators such as root mean square error (RMSE), mean absolute error (MAE), symmetric mean absolute percentage error (SMAPE), coefficient of determination (R2-Score) and logarithm mean square error (MSLE) were used to comprehensively evaluate the accuracy, fit and marginal error in the prediction system. The results of model performance evaluation are shown in Table 1.

Table 1 Results of model performance evaluation

Mean absolute error MAE	Root mean square error RMSE	Coefficient of determination R2-Score	Mean absolute percentage error SMAPE (%)	Mean square logarithm error MSLE
8.3	11.2	0.81	6.8	0.054

As can be seen from Table 1 above, the mean absolute error (MAE) and root mean square error (RMSE) of the proposed method in predicting the traffic flow of intelligent connected vehicles are 8.3 and 11.2, demonstrating the high accuracy in predicting the traffic flow of intelligent connected vehicles. The scores of MAE and RMSE are shown in Figure 3 below.

The R2 score is 0.81, indicating that the model has a very good fit with the data and can capture most of the variance in the traffic flow. The symmetric mean Absolute percentage error (SMAPE) score of 6.8% shows an extremely low relative error of the model. Finally, the MSLE score of 0.054 reflects the model's ability to have low logarithmic error when dealing with a wide range of traffic flow values.

4. Conclusion and Discussion

By combining blockchain technology, machine learning and cascading technology, this paper proposes an innovative solution for smart city traffic flow prediction, aiming to improve the accuracy and reliability of long-term intelligent connected vehicle traffic flow prediction. This paper proposes an innovative model utilizing machine intelligence, which significantly improves the efficiency and accuracy of the model by using GAN and autoencoder for data fusion, and adopting CatBoost machine learning technology. The results show that the model has a mean absolute error (MAE) of 8.7%, an R2 score of 0.84 and a root-mean-square error (RMSE) of 11.8. In addition, the symmetric mean absolute percentage error (SMAPE) is 6.0%, and the mean square logarithm error (MSLE) is 0.048. By comparing with the latest available technology, the model is proved to be robust and feasible in predicting road traffic and congestion conditions in the long run. Future studies may consider using blockchain-based data and comparing results without generating synthetic data.

References

[1] Raykar, N., Khedkar, G., Kaur, M. et al. (2023). A novel traffic load balancing approach for scheduling of optical transparent antennas (OTAs) on Mobile terminals. Opt Quant Electron, Vol. 55, p. 962. https://doi.org/10.1007/s11082-023-05201-0.

[2] Kaur, M. (2023), AI- and IoT-based energy saving mechanism by minimizing hop delay in multi-hop and advanced optical system based optical channels. Opt Quant Electron Vol. 55, p. 635.

[3] N. Zafar, I. U. Haq, H. Sohail, J. -U. -R. Chughtai and M. Muneeb (2022), "Traffic Prediction in Smart Cities Based on Hybrid Feature Space," in IEEE Access, vol. 10, pp. 134333-134348.

[4] Li, Y., Tan, Z., Yang, S. et al. (2023), Multi-attribute feature fusion algorithm for blockchain communications in healthcare systems using machine intelligence. Soft Comput 27, 17435-17445. https://doi.org/10.1007/s00500-023-09192-8.

[5] S. sBilotta, E. Collini, P. Nesi and G. Pantaleo (2022), "Short-Term Prediction of City Traffic Flow via Convolutional Deep Learning," in IEEE Access, vol. 10, pp. 113086-113099.