__Original Research Article__

Forecasting of road traffic flow based on harris hawk optimization and XGBoost

ABSTRACT

Aims: By predicting short-term traffic flow to assist intelligent transportation system decision-making, it can more effectively solve the problem of congestion and improve road capacity.
Study design: In this paper, a multi-step prediction HHO-XGBoost model is proposed by combining XGBoost and Harris Hawk optimization, which is applied to traffic flow prediction.
Place and Duration of Study: Southwest University of Science and Technology, between July 2021 and March 2022.
Methodology: The hyperparameters of the XGBoost model are optimized using Harris Hawk Optimization. The XGBoost parameter optimization is treated as an optimization problem and solved using an intelligent optimization algorithm to obtain the optimal parameters of the model, and the proposed model is applied to traffic flow prediction together with seven other representative models.
Results: In the prediction case, the proposed model has the lowest MAPE, and the proposed model has the best prediction performance and the most stable generalization compared to the comparison models. In addition, Harris Hawk Optimization has more powerful global search ability than Salp swarm algorithm
Conclusion: The HHO-XGBoost model has a strong strong potential in traffic flow prediction field. The parameter adjustment of the machine learning model based on the intelligent optimization algorithm is a way to improve machine learning. The HHO has stronger global search ability than SSA, and can find the global optimal solution with fewer iterations.

Keywords: {Taffic flow, Extreme gradient boosting, Harris hawk optimization, Multi-step forecast }

UNDER

1. INTRODUCTION

1.1 Backgroud

With the progress of urbanization in the world and the popularization of automobiles, traffic problems are increasing day by day, such as traffic congestion, frequent traffic accidents, and traffic environment deterioration. Frequent traffic accidents and traffic congestion will prolong people's travel time and cause more traffic accidents to a certain extent. With the development of society and economy, the desire of human beings for a better life continues to breed, resulting in an increasing trend of owning one private car per capita rather than one private car per household[1].With the gradual increase of private vehicles, the situation of urban traffic congestion is becoming more and more serious, and traffic accidents are increasing year by year. Traffic congestion not only causes serious pollution to the environment, but also has a great negative impact on human health[2].

The intelligent transportation system relies on the monitoring and guidance of the running state of the driving vehicles to optimize the distribution of traffic flow in the road network[3], which can effectively alleviate road traffic congestion and traffic accidents. It is currently the most effective way to solve traffic problems recognized by the world. As one of the key technologies in the field of intelligent transportation research, traffic flow prediction has become a research hotspot at home and abroad.

Traffic flow prediction is the basis for realizing traffic guidance and reasonable traffic control, and it is a prerequisite for the realization of intelligent transportation. However, the randomness and dynamic characteristics of traffic flow determine the difficulty of prediction, which has always been a difficult and hot spot in traffic flow research.

1.2 Related work

Up to now, a large number of researchers have carried out a lot of research work on the short-term forecast of traffic flow, using the knowledge of various disciplines, and proposed a variety of forecasting theories and methods. According to the nature of the prediction method itself, the current traffic flow prediction models can be roughly classified into the following categories: statistical models, machine learning models, deep learning models, and gray system models. Cools(2009) et al.[4] used ARIMAX and SARIMAX models to study changes in daily traffic flow. Yu Guoqiang(2003)[5] predict short-term traffic flow based on Markov chain model. Li Mengzhang and Zhu Zhanxing adopted spatial-temporal fusion graph neural networks for traffic flow forecasting[6]. Zhao Wentian et al.(2019)[7] used deep temporal convolutional networks for short-term traffic flow prediction and achieved very dazzling results. Jiawei Cao et al.(2020)[8] adopted XGBoost to predict short-Term traffic flow on highways. Xiao Xingping[9,10,11] has introduced the grey system model into the field of traffic flow prediction many times, and achieved satisfactory results and results.

2. METHODOLOGY

2.1 Extreme Gradient Boosting

XGBoost is a decision tree-based algorithm proposed by Chen Tianqi et al. in 2016[12]. It efficiently implements the GBDT algorithm and makes many improvements in algorithm and engineering. It is widely used in Kaggle competitions and many other machine learning competitions and achieved got good results.

The core idea of XGBoost is based on Boosting, and the general process is shown in the Eq.1.

$$
\begin{aligned}
&\hat{y}_i^{(0)} = 0 \\
&\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_{1(x_i)} \\
&\hat{y}_i^{(1)} = f_{1(x_i)} + f_{2(x_i)} = \hat{y}_i^{(1)} + f_{2(x_i)} \\
&\vdots \\
&\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i{}^{(t-1)} + f_{t(x_i)}
\end{aligned}
\qquad (1)
$$

Where $\hat{y}_i^{(t)}$ represents the forecasting result of the t-th round model, and $\hat{y}_i^{(t-1)}$ is the forecasting result of the $t-1$ round model. What's more, the objective function of XGBoost is mainly composed of two parts: loss function, regularization term, such as Eq.2.

$$obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_i) \qquad (2)$$

UNDER PEER REVIEW

$$= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f(x_i)\right) + \Omega(f_t) + C$$

Where $\Omega(f) = \gamma T + \frac{1}{2}\lambda\left||w|\right|^2$, C is a constant. We use the second-order expansion of Taylor's formula to approximate the objective function, remove the constant term at the same time, and expand ω. Besides, map all samples to the tree structure through the function $q(x)$ to obtain Eq.3:

$$\begin{aligned} obj^{(t)} &= \sum_{i=1}^{n}[g_i f_t(x_i) + h_i f_t^2(x_i)] + \Omega f(t) \\ &= \sum_{i=1}^{n}\left[g_i \omega_{q(x)} + h_i \omega_{q(x)}^2\right] + \gamma T + \frac{\lambda}{2}\sum_{j=1}^{T}\omega_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i \in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right)\omega_j^2\right] + \gamma T \end{aligned} \qquad (3)$$

Where $I\_j = \{i|q(x_i) = j\}$ as the set of subscripts of the samples on each leaf node j. Define $G_j = \sum_{i \in I_j} g_i$, $H_j = \sum_{i \in I_j} h_i$:

$$obj^{(t)} = \sum_{j=1}^{T}\left[G_j \Omega_j + \frac{1}{2}(H_j + \lambda)\omega_j^2\right] + \gamma T \qquad (4)$$

When the structure of the tree $q(x)$ is determined, the optimal weight of the leaf can also be easily calculated. At this time, the objective function value can be calculated by Eq.5.

$$obj^{(t)} = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j}{H_j + \lambda} + \gamma T \qquad (5)$$

The result of Eq.5 is used to evaluate the quality of the tree structure. The smaller the result, the better the tree structure.

2.2 Harris Hawk Optimization

Inspired by the predation behavior of Harris Eagle, in 2019 Heidari et al. proposed the Harris Hawk Optimization[13]. The algorithm is mainly composed of three parts: search, conversion and development of search and development, and has strong global search ability.

2.2.1 Exploration phase

In HHO, Harris Hawk is the candidate solution, and the best candidate solution in each step is considered to be the expected prey or close to the optimal solution. Harris Hawks will perch somewhere randomly and find prey through two strategies:

$$E(i+1) = \begin{cases} E_{rand}(i) - c_1|E_{rand}(i) - 2c_2 E(i)|, q \geq 0.5 \\ \left[E_{food}(i) - E_m(i)\right] - c_3[lb + c_4(ub - lb), q < 0.5 \end{cases} \qquad (6)$$

Where $E(i)$ and $E(i+1)$ are the position of the individual at the current and next iteration, respectively, i is the number of iterations and $E_{rand}(i)$ is an individual randomly selected from the population, $E_{food}(i)$ is the prey position. Besides, $c1$, $c2$, $c3$ and q are random numbers inside$[0,1]$. $E_m(i)$ is the average position of the current population of hawks. which can be calculated by Eq.7.

$$E_m(i) = \sum_{k=1}^{M} E_k(t)/M \qquad (7)$$

$E_k(t)$ represents the position of each individual in the Kth iteration, and M represents the population size.

UNDER PEER REVIEW

2.2.2 Transition from exploration to exploitation

The HHO algorithm switches between searching and different exploitation behaviors according to the escape energy of the prey, which is defined as:

$$W = 2W_0(1 - \frac{i}{I}) \qquad (8)$$

Among them, W_0 is the initial energy of the prey, which is a random number between $[-1, 1]$, which is automatically updated at each iteration, and I is the maximum number of iterations. Enter the search phase when $|E| \geq 1$, and enter the development phase when $|E| < 1$.

2.2.3 Exploitation phase

Define r as a random number between $[0,1]$. When $0.5 \leq |W| < 1$ and $r \geq 0.5$, a soft siege strategy is adopted to update the position:

$$E(i+1) = \Delta E(i) - W|LE_{food}(t) - E(i)| \qquad (9)$$

When $|W| < 0.5$ and $r \geq 0.5$, adopt a hard siege strategy for position update:

$$E(i+1) = E_{food}(i) - W|\Delta E(i)| \qquad (10)$$

When $0.5 \leq |W| < 1$ and $r < 0.5$, the soft encircling strategy of asymptotic fast dive is adopted to update the position:

$$E(i+1) = \begin{cases} Y, f(Y) < f(E(I)) \\ Z, f(Z) < f(E(I)) \end{cases} \qquad (11)$$

where $Y = E_{food}(i) - W|LE_{food}(i) - E(i)|$, $Z = Y + S * Tf(2)$. f is a fitness function, S is a two-dimensional vector, and its elements are random numbers in $[0,1]$, TF is the mathematical expression of Levi's flight.

When $|W| < 0.5$ and $r < 0.5$, adopt the hard encircling strategy of asymptotic fast dive to update the position:

$$E(i+1) = \begin{cases} Y, f(Y) < f(E(I)) \\ Z, f(Z) < f(E(I)) \end{cases} \qquad (12)$$

where $Y = E_{food}(i) - W|LE_{food}(i) - E_m(i)|$.

In the above four formulas, $\Delta E(i) = E_{food}(i) - E(i)$ represents the difference between the prey position and the individual's current position, and L is $[0, 2]$ between random numbers.

2.3 Establishment of a novel hybrid model

The optimization of hyperparameters is a difficult and hot issue in the field of machine learning. The quality of a model depends largely on the choice of hyperparameters.Most of the time, researchers adjust the hyperparameters based on their own experience, but this method is highly subjective and may not be able to obtain satisfactory results. Therefore, in this study, we introduce an intelligent optimization algorithm to optimize the hyperparameters of the model.

The construction process of the novel prediction model based on HHO and XGBoost is mainly divided into the following five steps: (1) Determining model parameters and initialization algorithm. (2) Define the objective function of HHO, (3) Continuously update the position based on the objective function value. (4) If is metting the maximum iteration, continue to run the algorithm. Otherwise, return the step(3). (5) Obtain the optimal model and apply the model to predict traffic flow. The detailed process of the model construction is shown in Figure.1.

UNDER PEER REVIEW

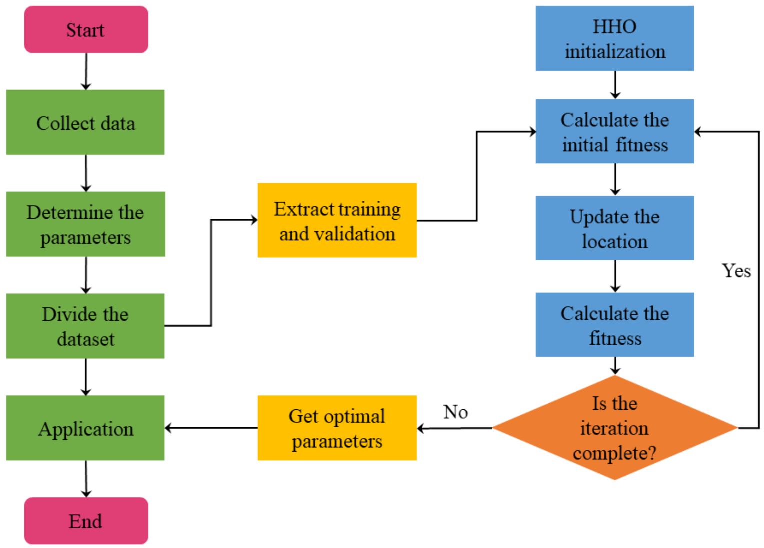


Figure 1 The construction process of HHO-XGBoost model

3. RESULTS AND DISCUSSION

3.1 Data description

The quality of a model depends largely on its generalization. Therefore, in this study, our data is derived from the OpenITS[14].The data comes from the Whitemud Drive Highway in Canada, where road traffic information is collected every twenty second. In our research, the data is processed into two granularities of ten minutes and twenty minutes.

The relevant information and some statistical characteristics of the two datasets are shown in Table 1.

Table 1 Some statistical features of the dataset

The time granularity of the dataset	Mean	Standard Deviation	Median	Skewness coefficient	Kurtosis coefficient
10 minutes	835.861	552.897	924.000	0.049	-1.117
20 minutes	778.058	518.233	813.000	0.093	-1.088

3.2 Forecasting result

In order to quantitatively analyze the prediction performance of the model, we introduce the Mean Absolute PercentageError (MAPE) to evaluate the prediction results of the model. The smaller the MAPE, the better the prediction performance. In addition, in order to verify the competitiveness of the proposed model, a variety of models and algorithms were introduced in this study for comparison, including Salp swarm algorithm (SSA), Random Forest (RF), LightGBM, Multilayer Perceptron (MLP) and Support Vector Regression (SVR).

We apply the proposed model to two datasets respectively. One step of prediction on the dataset with time granularity of 10 minutes represents the prediction of road traffic flow in the next ten minutes. On the dataset with time granularity of 20 minutes The prediction step represents the prediction of the road traffic flow in the next 20 minutes. In this study, the road traffic flow for the next fifty minutes and the next one hundred minutes was predicted respectively. The detailed prediction results are shown in Table 2.

UNDER PEER REVIEW

Table 2 MAPE(\%) of prediction results for both datasets

Dataset	Model / Step	HHO-XGB	SSA-XGB	XGB	HHO-RF	HHO-LGB	SVR	MLP
10 minutes	1-step	7.611	16.693	20.963	72.438	80.284	17.101	11.606
	2-step	7.942	20.709	23.505	73.857	81.767	18.811	11.45
	3-step	7.474	23.716	25.761	75.416	83.398	21.493	11.263
	4-step	7.631	26.161	28.949	76.429	84.456	24.213	12.426
	5-step	7.564	28.722	33.163	77.728	85.814	26.977	12.912
20 minutes	1-step	10.985	13.884	13.578	28.384	17.832	43.043	11.016
	2-step	12.779	14.558	14.906	29.208	21.517	42.863	13.844
	3-step	13.88	14.362	16.095	33.475	25.368	42.679	15.641
	4-step	15.29	14.916	17.01	36.094	29.942	42.639	16.939
	5-step	16.857	15.596	18.554	40.056	34.742	42.636	18.759

As can be seen from the table, HHO-XGBoost has a very prominent advantage among all competitors. It is worth mentioning that in the two datasets with different statistical characteristics, the results of the HHO-XGBoost model are better than other competitors, which also shows from the side that our proposed novel model not only has excellent prediction performance. It also has very good generalization and can be applied to a variety of different situations.

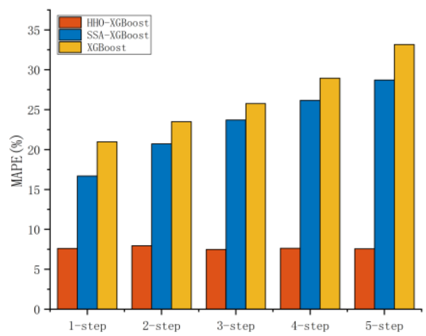
Among all the comparison models, we not only selected random forest and lightGBM which belong to the same decision tree model as XGBoost, but also selected SVR and MLP which are completely different from XGBoost principle, which is enough to illustrate the powerful predictive ability model of XGBoost and can be applied to traffic field of flow prediction.

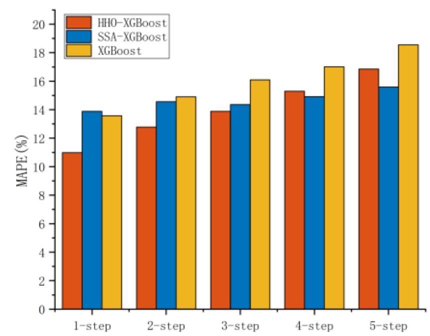
3.3 Discussion

In this section we have a separate discussion on intelligent optimization algorithms. Figure 2 shows the MAPE of the three models HHO-XGBoost, SSA-XGBoost, and XGBoost. It can be seen from the figure that the optimization algorithm has a very good effect on improving the accuracy of the model. This is enough to show that the intelligent optimization algorithm is reasonable and effective for the parameter adjustment of the machine learning model. Figure 3 shows the error drop graph of the HHO and SSA algorithms with the increase of the number of iterations. It can be clearly seen from the figure that compared with SSA, HHO has a faster decrease speed. When the number of iterations is only 60 The second time, the HHO algorithm has reached the optimal value, which indicates that the global search ability of HHO is stronger than that of SSA.

To sum up, it is a very effective method to use the intelligent optimization algorithm to optimize the hyperparameters of machine learning. Compared with the traditional GridSearchcv and empirical parameter tuning, it has the advantages of faster speed, smaller memory consumption and better prediction accuracy. high characteristic. However, it is worth mentioning that in some data sets, because the intelligent optimization algorithm is too focused on achieving the optimum on the validation set, it is prone to overfitting when making predictions.

UNDER PEER REVIEW





(a) 10 minutes (b) 10 minutes

Figure 2 Algorithm tuning and comparison of default parameters

4. CONCLUSION

In order to solve the existing problems in the field of traffic flow prediction and effectively improve the prediction accuracy and generalization performance of the model, in the research, we use HHO to optimize the hyperparameters of XGBoost, and propose a new HHO-XGBoost for Predict the traffic flow on the road. In order to demonstrate the good enough predictive ability of the hybrid model, we compare the prediction results of HHO-XGBoost with the prediction results of seven other representative models, and finally we also discuss the intelligent optimization algorithm used in the study.

To sum up, we can draw the following three conclusions: (1) The HHO-XGBoost algorithm shows strong prediction ability and generalization performance in different prediction steps and different datasets, which shows that the HHO-XGBoost model has a strong application in traffic flow. The prediction field has strong potential (2) The prediction performance of the two hybrid models HHO-XGBoost and SSA-XGBoost is better than that of XGBoost. It can be concluded that the parameter adjustment of the machine learning model based on the intelligent optimization algorithm is a way to improve machine learning. Effective means of model accuracy. (3) HHO has stronger global search ability than SSA, and can find the global optimal solution with fewer iterations.

REFERENCES

1. Angayarkanni, S. A., R. Sivakumar, and Y. V. Ramana Rao. "Hybrid Grey Wolf: Bald Eagle search optimized support vector regression for traffic flow forecasting." *Journal of Ambient Intelligence and Humanized Computing* 12.1 (2021): 1293-1304.
2. Zhang, Fan, et al. "Influence of traffic activity on heavy metal concentrations of roadside farmland soil in mountainous areas." *International journal of environmental research and public health* 9.5 (2012): 1715-1731.
3. Li, Xiang. "Intelligent transportation systems in big data." *Journal of Ambient Intelligence and Humanized Computing* 10.1 (2019): 305-306.
4. Cools, Mario, Elke Moons, and Geert Wets. "Investigating the variability in daily traffic counts through use of ARIMAX and SARIMAX models: assessing the effect of holidays on two site locations." *Transportation research record* 2136.1 (2009): 57-66.

UNDER PEER REVIEW

5. Yu, Guoqiang, et al. "Short-term traffic flow forecasting based on Markov chain model." *IEEE IV2003 Intelligent Vehicles Symposium. Proceedings (Cat. No. 03TH8683)*. IEEE, 2003.
6. Li, Mengzhang, and Zhanxing Zhu. "Spatial-temporal fusion graph neural networks for traffic flow forecasting." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. No. 5. 2021.
7. Zhao, Wentian, et al. "Deep temporal convolutional networks for short-term traffic flow forecasting." *IEEE Access* 7 (2019): 114496-114507.
8. Xiao, Xinping, and Huiming Duan. "A new grey model for traffic flow mechanics." *Engineering Applications of Artificial Intelligence* 88 (2020): 103350.
9. Xiao, Xinping, et al. "An improved seasonal rolling grey forecasting model using a cycle truncation accumulated generating operation for traffic flow." *Applied Mathematical Modelling* 51 (2017): 386-404.
10. Duan, Huiming, Xinping Xiao, and Qinzi Xiao. "An inertia grey discrete model and its application in short-term traffic flow prediction and state determination." *Neural Computing and Applications* 32.12 (2020): 8617-8633.
11. Cao, Jiawei, et al. "Short-Term Highway Traffic Flow Forecasting Based on XGBoost." *2020 15th International Conference on Computer Science & Education (ICCSE)*. IEEE, 2020.
12. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.
13. Heidari, Ali Asghar, et al. "Harris hawks optimization: Algorithm and applications." *Future generation computer systems* 97 (2019): 849-872.
14. Peng,L. &OpenITS Org. OpenData V7.0-Canada Whitemud Drive highway data https://www.openits.cn/openData1/700.jhtml (2021). Accessed: 2021-12-12

UNDER PEER REVIEW