ORIGINAL ARTICLE

Prediction with coastal environments and marine diesel engine data based on ship intelligent platform

Yuxue Liu[1] · Shuli Jia[1] · Yuan Yu[2] · Liyong Ma[2]

Abstract

An intelligent platform prototype is established for a coastal environment monitoring ship. LSTM and GBDT methods are developed for pH value and fuel consumption prediction in the intelligent platform. The results of applying the general prediction algorithms to actual environments' data and marine diesel engine data are reported. GBDT has the best predictive results with the smallest error. SVM and SVR have similar prediction effects, while FNN has the largest error. As the prediction time increases, the error of LSTM becomes large. The ship intelligence platform can provide unified data support and general intelligent algorithms for data-driven applications, and it has the potential to be widely used in coastal environmental monitoring applications.

Keywords Ship intelligent platform · Coastal environment · Marine diesel engine · Prediction

Introduction

Sensing, communication, computing, and information technology are developing rapidly, and ships can provide more and more digital information. The amount of this information data is huge, and the data format and update frequency vary significantly. The centralized data center approach cannot provide the necessary support for the increasing number of applications with very different requirements. A ship intelligence platform to support big data came into being. Ship intelligence platforms generally use distributed storage and hierarchical data architecture to provide different data support for different applications. This intelligent platform can not only meet the underlying ship equipment control requirements, but also support advanced decision applications. In recent years, researches based on ship intelligence platforms have received more and more attention, such as the big data architecture of ship intelligence platforms (Jia et al. 2018), intelligent obstacle avoidance based on data fusion (Wang et al. 2019; Ma et al. 2019a, b), intelligent navigation applications, and so on. Ship intelligence platforms are very important, but there are very few intelligent platforms and their applications that have been established.

Water quality prediction and fuel consumption prediction are two important prediction applications. Water quality monitoring is an important part of coastal environmental monitoring (Ahmed et al. 2020).Water quality prediction has always been one of the focus issues being studied (Sun et al. 2020).Chlorophyll concentration in coastal waters has been studied, water pollution is predicted with the least square support vector machine method, PCA-SVR method is employed for seawater quality prediction by Sun et al. (2020). In this paper, the pH value, which is very important in water quality, is predicted. The fuel consumption prediction of marine diesel engines has always been the focus of energy efficiency research (Manigandan et al. 2020; Pauline et al. 2021; Saravanan et al. 2020). Le et al. (2020) and Jeon et al. (2018) used neural network methods to predict the fuel consumption of the ship's main engine. Gkerekos et al. (2019) not only compared the effects of different forecasting methods, but also analyzed the factors that affect fuel consumption. In this paper, the diesel consumption of the marine engine is predicted.

Although the ship intelligence platform is very important, there are few reports about the establishment of intelligent applications on the platform. There is little research on the application of coastal environmental based on ship

✉ Shuli Jia
shulijia@yeah.net

[1] Automation Engineering Department, Shanghai Marine Diesel Engine Research Institute, Shanghai, China

[2] School of Information Science and Engineering, Harbin Institute of Technology, Weihai, China

intelligent platform. Recently, an intelligent platform for a coastal environment monitoring ship was successfully constructed by Shanghai Marine Diesel Engine Research Institute. The intelligent platform stores many sensor data in a distributed manner. These sensors collect a large amount of data on marine power equipment including a marine diesel engine. The platform also stores the detection data of the monitoring ship on the coastal environment. To better provide support for data analysis and utilization, the ship intelligence platform provides some modular analysis algorithms. The purpose of this paper is to evaluate the effectiveness of these general analysis algorithms of the intelligent platform applied to real data. In the intelligent platform prototype system, the water measurement data of the environmental monitoring ship is stored. The prediction algorithms LSTM (long short-term memory) and GBDT (gradient boosting decision tree) are selected for the prediction of diesel engine state and water quality of environmental monitoring data.

The main contributions of this paper include an intelligent platform prototype which is established for a coastal environment monitoring ship, and the general intelligent prediction algorithms are developed for the platform. The results of applying the general prediction algorithms to actual data are reported in this paper. This paper is organized as follows. After introduction section, LSTM and GBDT algorithms are described in methods section, and data collection is also described. In results and discussion section, LSTM and GBDT algorithms are compared with other common methods, and results are reported. The conclusion is given at the end of the paper.

Methods

In the prototype of the ship intelligence platform, two important prediction algorithms, LSTM and GBDT, were implemented. Two prediction algorithms will be introduced, and the actual data applying these algorithms will be described in detail.

LSTM prediction algorithm

The main feature of RNN (recurrent neural network) design is that the current state is related to the previous state. The emergence of RNN solves the two problems of processing variable-length input and analyzing the sequence information of the sequence. Since each state of the RNN contains the output of the previous state, this architecture leads to a long-term dependency problem. LSTM is specially designed to solve this problem. To predict the time series of some future time values, in addition

to the known time series, such as the current value of series x_t, the LSTM prediction algorithm also sends two pieces of information backward in the time dimension, namely the hidden layer state h_{t-1} and the cell state c_{t-1}. The state of h_{t-1} can be regarded as the prediction of the value of the LSTM algorithm at the future time based on the current information and a small amount of memory information, or it is called one of the LSTM outputs at time $t-1$. Obviously, another output of LSTM is the cell state c_t, which is a pure memory state used to tailor the memory information for the hidden layer to prevent the hidden layer from processing too long information.

Unlike RNN, which has only one tanh layer in its repeated structure module, the LSTM model uses a state and gate mechanism to preserve long-term memory. The gate in LSTM is a fully connected network layer. Its input can be multiple vectors and its output is a real number. This real number indicates the degree of the door opening. In deep learning, this gate represents the degree of influence of variables on variables. LSTM introduces three gates to control the inflow and outflow of long-term state c. The three gates are the forgotten gate, input gate, and output gate. LSTM has many successful prediction application (Shahid et al. 2021).

The forgotten gate can be expressed as

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right). \tag{1}$$

Let f_t be the output of the forgotten gate, σ be the activation function, W be the network weight matrix, h_{t-1} be the output of the hidden layer, x_t be the network input, and b be the network bias. The forget of the input gate can be expressed as

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right). \tag{2}$$

The current input cell status is

$$\overline{c}_t = \tanh\left(W_c \cdot [h_{t-1}, x_t] + b_c\right). \tag{3}$$

The current cell status is

$$c_t = f_t \circ c_{t-1} + i_t \circ \overline{c}_t, \tag{4}$$

where the symbol ∘ means multiply by element. The output gate is used to control how much of the united state c_t is input to h_t. The output of the output gate is

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right). \tag{5}$$

Then the final output of the LSTM is

$$h_t = o_t \circ \tanh(c_t). \tag{6}$$

LSTM uses a chain structure to connect repeated neural network modules to achieve the purpose of memorizing values of indefinite length of time.

GBDT prediction algorithm

GBDT is a machine learning model with good performance for completing regression, classification, and ranking tasks, and it is widely used in the industry. GBDT belongs to the boost algorithm cluster, and its idea is to predict the data after combining multiple weak classifiers. The core of the gradient boosting algorithm is as follows: first, the residual of the decision tree model is evaluated, then negative gradient obtained from loss function is employed for boost residual, and finally, a fitted regression tree is gradually obtained by iteration. Therefore, gradient boosting is an iterative process that continuously fits the residuals and reduces the residuals.

The GBDT model is

$$F_M(x) = \sum_{m=1}^{M} T(x;\Theta_m), \tag{7}$$

where $T(x;\Theta_m)$ is the decision tree, x is the input vector, Θ_m is decision tree parameters, and M is the number of decision trees. The squared loss is chosen as the loss function of GBDT, which is

$$L(y, T_m(x)) = \frac{1}{2}\left(y - T_m(x)\right)^2, \tag{8}$$

where y is the output. Then the residual is equal to the negative gradient, and the direction of the residual is the direction of global optimization. GBDT can determine the parameters $\hat{\Theta}_m$ of the next decision tree by minimizing the loss function L as

$$\hat{\Theta}_m = \arg\min \sum_{m=1}^{M} L(y_i, T_{m-1}(x_i) + T(x, \Theta_m)). \tag{9}$$

It can be seen from the above formula that GBDT has the characteristics of a good fit to the real distribution, especially feature selection and model integration are its advantages.

Coastal environment and marine diesel engine data for prediction

Water quality parameters are measured and recorded hourly by the coastal environment monitoring ship. The pH value is one of the most important parameters reflecting water quality (Ahmed et al. 2020; Tiyasha et al. 2020; Sun et al. 2020), and it is selected as the environmental prediction data. pH value is measured with pH meter PHSJ-6L (Inesa Instrument, China). A total of 10 consecutive days of measurement data are used for prediction experiments. Seven days

of data are used as training dataset, and others are used as test dataset.

To predict diesel consumption, the coastal environment, and marine engine data are used. The coastal environment data include wind speed, wind direction, air temperature, and sea wave height, and these are measured with the instrumentAM-4836C (Reall tech Instrument, China). The engine and navigation system provide diesel consumption data and ship speed. All the data are collected when the ship is stable for more than 20 min. All these data from intelligent platform and sensors are employed for data analysis. 30 pieces data of 4-h voyage on the same ship route are collected. The total of 24 pieces are regarded as the training dataset, while the other pieces are used as test datasets. When forecasting, the first hour of data are used as the starting sequence, and the remaining 3 h data are used as a multivariate time series for prediction.

Results and discussion

Other algorithms used in the literature for water quality prediction and diesel engine prediction are used to compare with algorithms in the intelligent platform. These algorithms include support vector machine (SVM) (Nie et al. 2019), regression vector machine (RVM) (Sun et al. 2020), and feedforward neural network (FNN) (Jeon et al. 2018). FNN is a typical neural network method, which is widely used in nonlinear models. SVM is the result of the application of statistical learning theory. It uses the kernel function to solve the convex quadratic programming problem to maximize the classification interval. RVM is a more robust statistical learning regression method.
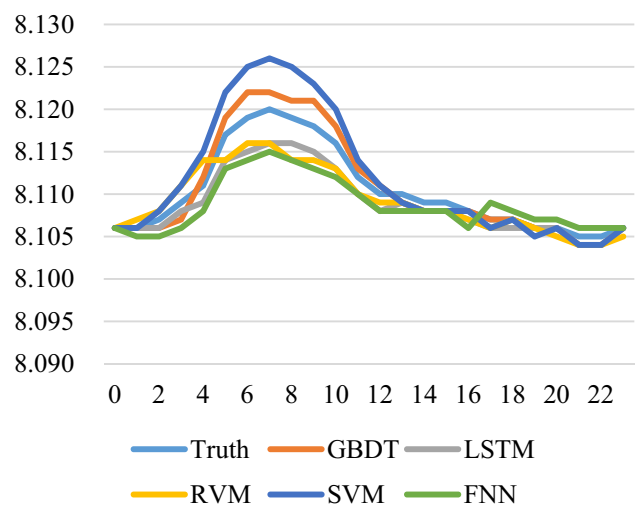

Fig. 1 pH prediction result of different methods for every hour

Table 1 MAE prediction comparison of pH with different prediction periods

Method	1 h	2 h	3 h	4 h	Average
FNN	0.00225	0.00304	0.00362	0.00402	0.00303
SVM	0.00204	0.00286	0.00294	0.00320	0.00276
RVM	0.00171	000,264	0.00284	0.00302	0.00255
LSTM	0.00150	0.00286	0.00416	0.00552	0.00351
GBDT	0.00104	0.00112	0.00108	0.00114	0.00110

The performance indicator means absolute error (MAE) is employed to evaluate these algorithms. It is defined as

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|. \tag{10}$$

The true value of the pH value of a whole day and the prediction results of different methods are illustrated in Fig. 1, where the predicted pH value is for every hour. The prediction performance of different algorithms, the prediction capabilities of different algorithms in 2 h, 3 h, and 4 h were also predicted. The MAE indicators of the prediction results of these algorithms at different time lengths are listed in Table 1.

The prediction of fuel consumption is a multivariate prediction problem. First, the fuel consumption is predicted every 20 min, which is consistent with the measurement sampling interval of the main diesel engine. To illustrate the prediction result figure, the principal component analysis (PCA) method (Sun et al. 2020) is used to determine the two most important parameters affecting fuel consumption: they are wind speed and ship speed. Fuel consumption prediction results of different methods for every 20 min are illustrated in Fig. 2. In the figure, wind speed and ship speed are measured values. Second, fuel consumption is predicted separately for every 40 min and 60 min. The MAE for all these predictions with different methods is calculated and listed in Table 2.

For the prediction of pH, the MAE of the GBDT method is the smallest, indicating that the prediction result of this method is the best. FNN has the worst prediction effect, and MAE is 2.16 times to 3.53 times that of the GBDT method. The results of RVM is similar to SVM, and their MAEs are better than FNN. Although FNN can express nonlinear models, for prediction problems, due to the insufficient depth of its neural network, its ability to solve prediction problems is limited. SVM and RVM use statistical learning models, and the best decision position can be selected through a small amount of data, so the effect is good. GBDT uses complex network architecture and has the advantage of integrating and synthesizing the results of multiple models, so the best results are obtained. The LSTM method is quite special, and its effect is better for forecasts with a small interval of 1 h. The LSTM method is worse than GBDT and better than
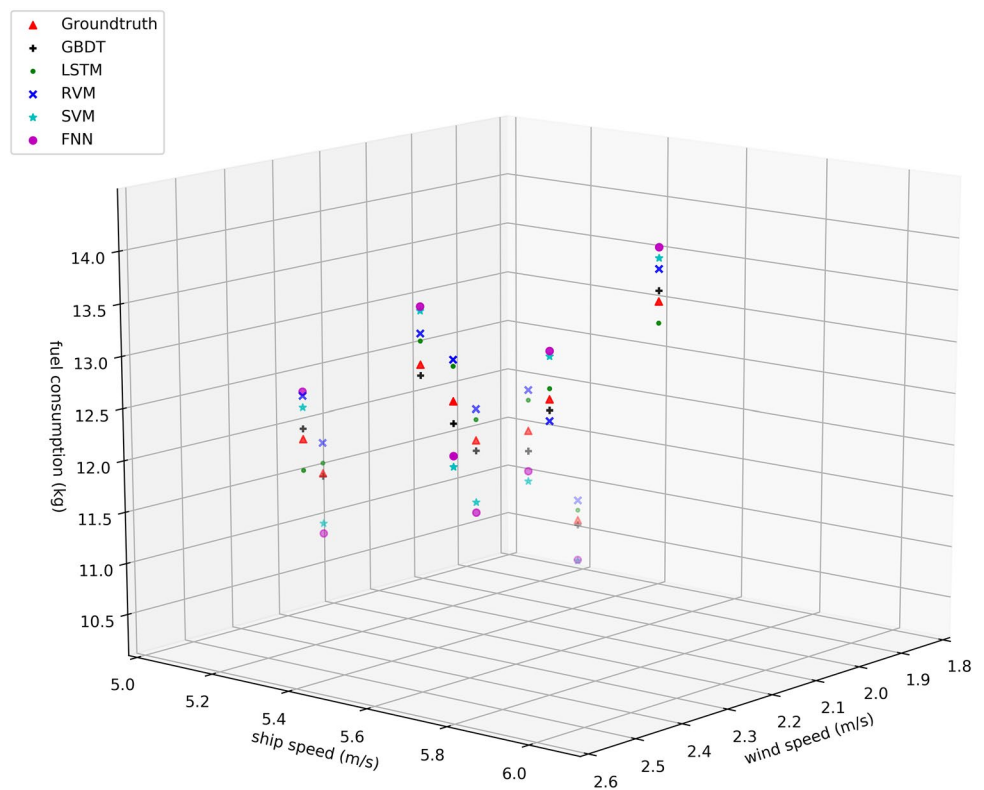
Fig. 2 Fuel consumption prediction results of different methods for every 20 min

Table 2 MAE prediction comparison of fuel consumption (kg) with different prediction periods

Method	20 min	40 min	60 min	Average
FNN	0.5056	0.5102	0.5234	0.5131
SVM	0.4689	0.4627	0.4628	0.4648
RVM	0.3089	0.3421	0.3506	0.3339
LSTM	0.2044	0.3842	0.5423	0.3770
GBDT	0.1089	0.2021	0.2018	0.1709

RVM and SVM. But for predictions with longer intervals, the error increases. This is mainly due to the structure of the cyclic network adopted by LSTM. When the interval is large, the prediction result still needs the support of small interval data, and the resultant error is accumulated. For the prediction of fuel consumption data, we can also observe results similar to the pH prediction.

The average MAE is also listed in the two tables, GBDT has the smallest average MAE, and the average error of FNN is greater than that of SVM and RVM. LSTM has the largest average error in pH prediction, and it is better than SVM and FNN in fuel consumption prediction. This is due to the different steps of multiple predictions and different cumulative errors. Its single-step forecasting effect is second only to GBDT. But for multi-step prediction, as the cumulative number of times increases, the error gradually increases.

Conclusions

An intelligent platform prototype is established for a coastal environment monitoring ship, LSTM and GBDT methods are developed for pH value and fuel consumption prediction. GBDT method has the smallest prediction error and gets the best prediction results. The intelligent platform can provide unified data support and general intelligent algorithms for data-driven applications, and it has the potential to be widely used in coastal environmental monitoring applications.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This research does not involve any human or animal.

Informed consent No informed consent was necessary for the study.

References

Ahmed U, Mumtaz R, Anwar H, Mumtaz S, Qamar AM (2020) Water quality monitoring: from conventional to emerging technologies. Water Supply 20(1):28–45

Gkerekos C, Lazakis I, Theotokatos G (2019) Machine learning models for predicting ship main engine fuel oil consumption: a comparative study. Ocean Eng 188:1–41

Jeon M, Noh Y, Shin Y, Lim OK, Lee I, Cho D (2018) Prediction of ship fuel consumption by using an artificial neural network. J Mech Sci Technol 32(12):5785–5796

Jia S, Ma L, Zhang S (2018). Big data prototype practice for unmanned surface vehicle. In: Proceedings of the 4th International Conference on Communication and Information Processing, pp 43–47

Le LT, Lee G, Park KS, Kim H (2020) Neural network-based fuel consumption estimation for container ships in Korea. Marit Policy Manag 47(5):615–632

Ma L, Ma C, Liu Y, Wang X (2019a) Thyroid diagnosis from SPECT images using convolutional neural network with optimization. Comput Intell Neurosci 2019:1–12

Ma LY, Xie W, Huang HB (2019b) Convolutional neural network based obstacle detection for unmanned surface vehicle. Math Biosci Eng MBE 17(1):845–861

Manigandan S, Atabani AE, Ponnusamy VK, Pugazhendhi A, Gunasekar P, Prakash S (2020) Effect of hydrogen and multiwall carbon nanotubes blends on combustion performance and emission of diesel engine using Taguchi approach. Fuel 276:1–15

Nie Z, Yuan Y, Xu D, Shen F (2019) Research on support vector regression model based on different kernels for short-term prediction of ship motion. In: 2019 12th International Symposium on Computational Intelligence and Design (ISCID), pp 61–64

Pauline JMN, Sivaramakrishnan R, Pugazhendhi A, Anbarasan T, Achary A (2021) Transesterification kinetics of waste cooking oil and its diesel engine performance. Fuel 285:1–11

Saravanan S, Kumar CR, Pugazhendhi A, Brindhadevi K (2020) Role of thermal barrier coating and porous medium combustor for a diesel engine: an experimental study. Fuel 280:1–7

Shahid F, Zameer A, Muneeb M (2021) A novel genetic LSTM model for wind power forecast. Energy 223:120069

Sun X, Wang X, Cai D, Li Z, Gao Y, Wang X (2020) Multivariate seawater quality prediction based on PCA-RVM supported by edge computing towards smart ocean. IEEE Access 8:54506–54513

Tiyasha, Tung TM, Yaseen ZM (2020) A survey on river water quality modelling using artificial intelligence models: 2000–2020. J Hydrol 585:124670. https://doi.org/10.1016/j.jhydrol.2020.124670

Wang C, Zhang X, Cong L, Li J, Zhang J (2019) Research on intelligent collision avoidance decision-making of unmanned ship in unknown environments. Evol Syst 10(4):649–658

Springer