Prediction of dissolved oxygen in marine ranch by Bayes algorithm based on linear sliding window selection

Jiajun Sun1,a

1the School of Computer Science and Technology, Shandong Business and Technology University, Yantai, Shandong.

a2019410040@sdtbu.edu.cn

Abstract. Accurate prediction of water quality in marine ranch can help to avoid the problem of marine mortality caused by changes in water quality and can help to improve the productivity of marine ranch. In order to make effective predictions of water quality parameters, this paper proposes a Bayes-based prediction algorithm based on linear sliding window selection method, and use the model to predict dissolved oxygen data. Comparative experiments with similar algorithms illustrate that the prediction accuracy of this model has improved considerably.

1. Introduction

Marine ranch is a modern form of marine fisheries production. Daoliang Li pointed out that at the current stage, one of the three major scientific challenges facing intensive marine ranching is the difficulty of precise control of marine ranching in real time[1]. Because aquaculture environments are diverse and require complex system integration in data collection, transmission and processing. For the time being, research on early warning prediction models is still in the exploratory stage.

Therefore, the key to intensive marine ranching lies in the development of predictive models for marine water quality. A great deal of work has been done to enable better prediction of water quality data. Daoliang Li integrates filtering and deep learning techniques to propose a shrimp body identification model and water quality prediction model[2–4]. Liu Shuangyi proposed a water environment prediction method based on rough set theory with multi-scale analysis[5,6]. By constructing a predictive early warning model, predictions of short-term water quality are made. However, machine learning algorithms are dependent on the size of the data set and the setting of parameters. When the data size is small or the parameters are not optimized, the models often do not give good predictions. Therefore, a low complexity model that does not rely on parameter optimization and can be more effectively applied in a practical water quality prediction environment.

Bayes algorithm is a prediction algorithm with a simple structure. And it has the advantages of low algorithm complexity, high prediction accuracy and resistance to data noise. Depending on the strength of the dependencies between attributes, they can be classified as naive Bayes models, semi-naive Bayes models and Bayes networks. Bayes algorithms are often used in areas such as text classification[7], address disaster assessment[8] and soil pollution analysis[9]. The traditional naive Bayes algorithm has shown good results in classification problems. However, since Bayes algorithms calculate the probability of occurrence of each category, this algorithm cannot be directly applied in numerical prediction problems.

In this paper, a water quality prediction model by Bayes algorithm based on linear sliding window selection was proposed, and then, we make predictions on a real dataset of a marine ranch in Shandong,
China. Finally, we compare this model with other time series prediction algorithms. The results of the comparison experiments show that the model has better prediction results.

The other sections of this paper are organized as follows: Section 2 introduces the data sources of the paper and the Bayes theory. Section 3 introduces the predicted results of the Bayes model for DO data prediction, and compare the result with the similar algorithms. Section 4 concludes the paper and describes the next phase of work.

2. Materials and Methods

2.1. Study area and data sources
Some water quality sensors were installed in the Qixia Marine Ranch in Shandong Province to obtain data on water quality. These data include information on dissolved oxygen, salinity, depth, and chlorophyll concentration. We collated the dissolved oxygen data from this marine ranch from 2016 to 2020 to form the training data-set for the model in this paper.

2.2. Bayes theory
In the Bayes formulation, the observed attributes are defined as X and the categories to be predicted are defined as Y. The proportion of each attribute among all attributes is defined as P(X). The proportion of each category in the overall sample is defined as P(Y). The maximum value of each label for a given set of attributes is calculated using Equation 1.

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)}$$ (1)

P (X|Y) is called the posterior probability and its calculation is more complex. In this paper, it is assumed that the attributes are independent of each other and Equation 2 is used to calculate this value. And N represents the total number of attributes.

$$P(X|Y = y_c) = \prod_{i=1}^{N}P(X_i = x_i|Y = y_c)$$ (2)

In practice, it may happen that a certain set of attributes does not find the same record in the model's data-set, thus making the probability of a certain item 0 and ultimately affecting the prediction accuracy of that category. To avoid this problem, it is necessary to use the Laplace formula as shown in Equation 3.

$$P(Y = y_c|X) = \frac{1+P_{c,x_i}}{|D|+N_y}$$ (3)

2.3. Linear sliding window selection
Sliding window is a traffic control technique used to process communication messages in computer networks. The two communicating parties negotiate the size of a window so that data is sent in a fixed number of bytes[10].

Following the sliding window method, the model in this paper generates a new data set for the original dissolved oxygen data series (Fig. 1). This is done by specifying a sliding window of size k and then sliding backwards from the start of the dissolved oxygen sequence, with the first k-1 records within the sliding window being the features and the last record being the label. All the data segments form the dissolved oxygen difference series data set.
In network communication, the sliding window is determined by negotiation based on the network conditions of the communicating parties. In this paper, the size of the window will have an impact on the prediction results.

In general, there are three ways to choose the size of the sliding window. The first method is to set the window size to a fixed value, and this method enables predictions to be calculated quickly. However, for different marine ranches, fixed parameter values can lead to low prediction accuracy. The second method is to traverse all sliding window sizes in the interval. This method can effectively improve the prediction accuracy of the algorithm, but in the case of large data set size, using this method will lead to long computation time and cannot meet the requirements of real-time prediction. The third method is to design an adaptive algorithm that quickly finds the parameter with the highest accuracy. This method both reduces the time required to calculate all window sizes and improves the prediction accuracy. For the purposes of this paper, the error function images are assumed to be convex under different sliding windows size. The final objective function of the program is therefore shown in Equation 4.

\[
\max i \{ \text{error}(i) > \text{error}(i + 1), i = 1, 2, 3, \ldots, n \} \quad (4)
\]

2.4. Algorithm description

Step 1: Data pre-processing. First calculate the first order difference series of the water quality data. And then initialise a sliding window size.

Step 2: Divide the data-set. 90% of all first-order difference data is selected as the training data set for the model. The remaining 10% is used as the test data set for the model. Then using the sliding window method, a sequence of differences is generated according to the initialized sliding window size.

Step 3: Make predictions. Firstly, record all occurrences of difference sequences in the training set. And then, for all samples on the test set, the probability of occurrence of various values at the next moment is calculated using Laplace's formula, and the value with the highest probability is selected as the prediction.

Step 4: Optimise prediction accuracy using the linear sliding window method. Calculate the error function corresponding to each window size and terminate the calculation when the error is no longer decreasing.

Step 5: Evaluation of prediction effect. Use the error function to make an evaluation of the predicted sequences.
3. Results & Discussion

3.1. Effect of sliding window size on error

In order to verify that the final objective function proposed in this paper holds in this marine ranch, the effect of different sliding window sizes on the error is calculated. As shown in the Table 1, the error functions were calculated, such as MAPE, RMSE and MAE [11,12]. The data in the table shows that there is a unique minimum value point for the sliding window size and therefore the model proposed in this paper can be used.

| Sliding window Size | Error evaluation function |
|---------------------|---------------------------|
|                     | MAE | RMSE | MAPE |
| 1                   | 0.003644 | 0.005476 | 0.002048 |
| 2                   | 0.003279 | 0.004934 | 0.001849 |
| 3                   | 0.003084 | 0.004728 | 0.001739 |
| 4                   | 0.002971 | 0.004599 | 0.001676 |
| 5                   | 0.002943 | 0.004602 | 0.001654 |
| 6                   | 0.002885 | 0.004525 | 0.001618 |
| 7                   | 0.002953 | 0.004616 | 0.001656 |
| 8                   | 0.002996 | 0.004692 | 0.001678 |
| 9                   | 0.003051 | 0.004765 | 0.001712 |
| 10                  | 0.003143 | 0.004809 | 0.001761 |
| 11                  | 0.003293 | 0.004956 | 0.001847 |
| 12                  | 0.003389 | 0.005066 | 0.001896 |
| 13                  | 0.003425 | 0.005083 | 0.001919 |
| 14                  | 0.003540 | 0.005216 | 0.001983 |
| 15                  | 0.003616 | 0.005293 | 0.002027 |
| 16                  | 0.003659 | 0.005331 | 0.002049 |
| 17                  | 0.003743 | 0.005411 | 0.002097 |
| 18                  | 0.003850 | 0.005555 | 0.002158 |
| 19                  | 0.003943 | 0.005628 | 0.002210 |

3.2. Model prediction effects and comparison experiments

Using the Bayes algorithm based on linear sliding window selection method proposed in this paper, the search process for the window size is shown in Fig. 2. The MAE is used as the error function and the red dot in Fig. 2. marks the calculation process. From the figure, we can see that the model performed a total of seven error calculations to find the minimum value of the error.
We compare the dissolved oxygen Bayes algorithm with linear regression, radial basis function network, long short-term memory, multilayer perceptual regression and support vector regression\[13–16\].

In table 2, the Bayes model was compared with similar time series prediction algorithm. The data in the table shows a improvement in the prediction accuracy of the dissolved oxygen Bayes model.

| Algorithm name | Error evaluation function |
|----------------|---------------------------|
|                | MAE | RMSE | MAPE     |
| Dissolved oxygen Bayes model based on linear sliding window selection | 0.002885 | 0.004525 | 0.001618 |
| Linear regression | 3.912484 | 3.916542 | 2.228264 |
| Radial basis function network | 0.030838 | 0.034020 | 0.017105 |
| Long short-term memory | 0.017707 | 0.019948 | 0.011849 |
| Multilayer perceptual regression | 0.011175 | 0.012384 | 0.006292 |
| Support vector regression | 0.027373 | 0.028499 | 0.015218 |

In order to be able to represent the prediction effect of the algorithm more visually, a comparative graph of similar algorithms is drawn as shown in Fig. 3.

**4. Conclusions**

At present, it is extremely difficult to achieve accurate control of marine pastures. In order to achieve accurate predictions of water quality, this paper establishes an effective water quality prediction model based on the analysis of water quality data from marine pastures.
The Bayes dissolved oxygen prediction model based on linear sliding window selection designed in this paper can quickly select the optimal sliding window size, reduce the calculation error and make the prediction accuracy higher than similar algorithms.

In the following work, the effectiveness of the model will be tested in more marine ranches, and the sliding window size finding algorithm will be further optimised to improve the prediction efficiency.

Acknowledgment
The authors would like to thank the editor-in-chief, the associate editor, and the reviewers for their insightful comments and suggestions.

References
[1] S, Liu. L, Xu. et al. (2014) Research on remote system for disease diagnosis of Penaeus vannamei based on internet of things. Journal of China Agricultural University 19: 189–195.
[2] L, Xu. S, Liu. D, Li. (2017) Prediction of water temperature in prawn cultures based on a mechanism model optimized by an improved artificial bee colony. Computers and Electronics in Agriculture 140: 397–408.
[3] Y, Chen. L, Song. Y, Liu. et al. (2020) A Review of the Artificial Neural Network Models for Water Quality Prediction. Applied Sciences 10: 5776.
[4] H, Yu. Y, Chen, S, Hassan. et al. (2016) Dissolved oxygen content prediction in crab culture using a hybrid intelligent method. Sci Rep 6: 27292.
[5] S, Liu. L, Xu. D, Li. (2016) Multi-scale prediction of water temperature using empirical mode decomposition with back-propagation neural networks. Computers & Electrical Engineering 49: 1–8.
[6] L, Xu. S, Liu. D, Li. (2017) Prediction of water temperature in prawn cultures based on a mechanism model optimized by an improved artificial bee colony. Computers and Electronics in Agriculture 140: 397–408.
[7] L, Jiang. L, Zhang. C, Li. et al. (2019) A Correlation-Based Feature Weighting Filter for Naive Bayes. IEEE Transactions on Knowledge and Data Engineering 31: 201–213.
[8] Q, He. H, Shahabi. A, Shirzadi. et al. (2019) Landslide spatial modelling using novel bivariate statistical based Naïve Bayes, RBF Classifier, and RBF Network machine learning algorithms. Science of The Total Environment 663: 1–15.
[9] M, Albuquerque. S, Gerassis. C, Sierra. et al. (2017) Developing a new Bayesian Risk Index for risk evaluation of soil contamination. Science of The Total Environment 603–604: 167–177.
[10] H, Kim. Y, Park. Y, Cho. M, Kim. et al. (2014) Time-slide window join over data streams. Journal of Intelligent Information Systems 43, 323–347.
[11] I, Castillo. R, Nickl. (2013) NONPARAMETRIC BERNSTEIN-VON MISES THEOREMS IN GAUSSIAN WHITE NOISE. Annals of Statistics 41: 1999–2028.
[12] C, Willmott. K, Matsuura. (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res 30: 79–82.
[13] J, Lupon. C, Diez-Lopez. M, Antonio. et al. (2017) Recovered heart failure with reduced ejection fraction and outcomes: a prospective study. European Journal of Heart Failure 19: 1615–1623.
[14] A, Bors. (2001) Introduction of the Radial Basis Function (RBF) Networks. Online Symposium for Electronics Engineers.
[15] K, Abe. J,A dam. H, Aihara. et al. (2015) Measurements of neutrino oscillation in appearance and disappearance channels by the T2K experiment with 6.6 x 10(20) protons on target. Physical Review D 91: 072010.
[16] S, Datta. S, Das. (2015) Near-Bayesian Support Vector Machines for imbalanced data classification with equal or unequal misclassification costs. Neural Networks 70: 39–52.