PREDICTION AND MAPPING OF COCHLODINIUM POLYKRIKOIDES RED TIDE USING MACHINE LEARNING UNDER IMBALANCED DATA

S. H. Bak , D.H. Hwang, U. Enkhjargal, H. J. Yoon *

Pukyong National University, Division of Earth Environmental System Science, 48513 Nam-gu Busan, South Korea - yoonhj@pknu.ac.kr

KEY WORDS: Machine Learning, Deep Learning, Harmful Algal Bloom, Red tide

ABSTRACT:

Cochlodinium polykrikoides(*C. polykrikoides*) is a phytoplankton that causes red tides every year in the middle of the South Sea of Korea. *C. polykrikoides* is a harmful Algae that has migratory ability and causes the fisheries damage over a long period of wide sea area if it causes red tide once. To minimize red tide damage, it is important to anticipate and prepare the red tide occurrence timing and location in advance. In this study, we predicted the occurrence of red tide of *C. polykrikoides* using machine learning techniques and compared the results of each algorithm. Logistic regression model, decision tree model, and multilayer neural network model were used for prediction of red tide occurrence. To produce the data set for model learning, we used the red tide occurrence map provided by the National Institute of Fisheries Science, the Local Data Assimilation and Prediction System(LDAPS) provided by the Korea Meteorological Agency, and the G1SST provided by the National Oceanic and Atmospheric Administration(NOAA). The feature vectors used for modeling consisted of 59 elements, which were made by using temperature, water temperature, precipitation, solar radiation, wind direction and wind speed. Only a very small number of red tide cases can be collected compared to the case of no red tide cases. Thus, an imbalance data problem arises in the data set. To overcome this imbalanced data problem, we used adding noise after oversampling to data of red tide occurrence to solve the difference of data between two classes.The data set is divided into 8: 2 to prevent over-fitting and 80% is used as the learning data. The remaining 20% was used to evaluate the performance of each model. As a result of evaluating the prediction performance of each model, the multilayer neural network model showed the highest prediction accuracy.

1. INTRODUCTION

A red tide is a phenomenon that sea surface color changes by phytoplankton in a special environmental condition. About 200 species have been reported to cause of red tide occurrence(Yoon, 2012). In Korea 67 species have been reported to cause of red tide occurrence(Kim et al., 2005; Kim et al., 1998, Yoon, 2012). In Korean, after 1990s increasing frequency of red tide occurrence due to dinoflagellates[3]. *Cochlodinium polykrikoides*(*C. polykrikoides*) has been blooming in every summer and fall(July to September) for last 20 years in Mid-south sea of Korea. *C. polykrikoides* is kinds of dinoflagellates that had two flagellum and forming resting spore in parts of life cycle(Yoon,2012). In addition, C. polykrikoides can sustaine blooming to long time and wide areas because these had self-moving ability(Yoon, 2012).

Until now, Spraying of yellow clay is one of the proposed red tide prevention method in Korea(Oh et al., 2012). However, spraying yellow clay had problem that losing economic cost(Purchasing, storage and management costs) and ineffective after wide area blooming. If we can know where and when red tide occurs in early stage, damage can be reduced. It is possible to reduce the damage of aquaculture by sinking or isolating the farm before the red tide patch is come close into the area where the coastal farms are concentrated. In case of the creature which can be selling, early harvesting can reduce the economic damage amount. However, the mechanism of red tide phenomenon such as occurrence, spread and disappearance of red tide were not clarified. In particular, *C. polykrikoides* has no information on causality that causes large-scale red tide. so it is difficult to predict red tide. To predict the natural phenomenon, a model is needed. In order to develop a model, a causal relation must be derived by analyzing a large amount of data. Therefore, we can not develop a model for a natural phenomenon that causal relation can not derived.

In order to solve this problem, we propose a new red tide prediction model based on machine learning. Since the machine learning based model uses the correlation between data and natural phenomenon instead of causal relation, it is possible to develop the model without knowing the causal relation.

2. DATA AND METHOD

In this paper, we use Deep Neural Network as a model based on machine learning. The dataset for model training by using G1SST(GHRSST; Group for High Resolution Sea Surface Temperature) from NASA(National Aeronautics and Space Administration), and LDAPS(Local Data Assimilation and Prediction System) from KMO(Korea Meteorological Administration). We extracted 59 features(variables) from these data for training model(Table 1). We used red tide alert data from NIFS in order to distinguish red tide period and non red tide period. We obtained red tide occurred location data from map of red tide occurrence contained from red tide alert using georeferencing. Oceanic and meteorological data were extracted from G1SST and LDAPS using red tide occurrence location data (Figure 1).

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-4-W16-143-2019 | © Authors 2019. CC BY 4.0 License.

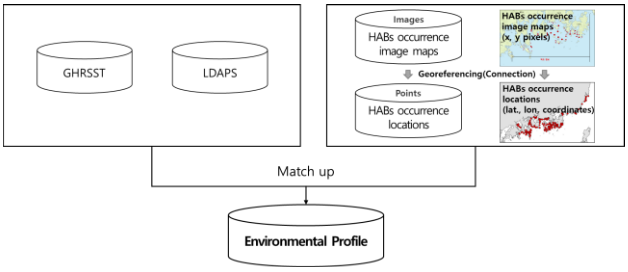
GHRSST
LDAPS
Images
HABs occurrence image maps
HABs occurrence image maps (x, y pixels)
Georeferencing(Connection)
Points
HABs occurrence locations
HABs occurrence locations (lat., lon, coordinates)
Match up
Environmental Profile

Figure 1. Dataset production process for modelling

Georeferencing is method that the internal coordinate system of a map or photo image can be related to a ground system of geographic coordinates. The oceanic environment data during non red tide occurred were randomly sampled in summer data when red tide not occurred and spring, winter data when red tide occurred. In order to prevent over-fitting, dataset were divided into 80:20 and 80% were used training data. Remaining 20% were used verification (Figure 2).

Only a very small number of red tide cases can be collected compared to the case of no red tide cases. Thus, an imbalance data problem arises in the data set. To overcome this imbalanced data problem, we used adding noise after oversampling to data of red tide occurrence to solve the difference of data between two classes.

Note
Precipitation in D-n days(Daily total value, D-4 to 9)
Solar radiation in D-n days(Daily total value, D-4 to 9)
Wind speed in D-n days(Daily mean, D-4 to 9)
Wind direction in D-n days(Daily mean, D-4 to 9)
Water temperature in D-n days(Daily mean, D-4 to 9)
Average amount of solar radiation change from D-4 to 9.
Average amount of water temperature change from D-4 to 9.
Difference between water temperature and air temperature of D-n days(D-4 to 9)
Minimum(maximum, mean) precipitation from D-4 to 9
Minimum(maximum, mean) solar radiation from D-4 to 9
Minimum(maximum, mean) wind speed from D-4 to 9
Minimum(maximum, mean) water temperature from D-4 to 9
Wind Chill Index in D-n days

Table 1. List of feature(variable) for prediction model training

In this study, logistic regression model, decision tree model, and deep neural network model were used to predict red tide occurrence. These machine learning models are known to be suitable for nonlinear model design when compared with regression or Threshhold Method(Bak et al., 2018).

The logistic regression model is a statistical model that expresses the relationship between independent and dependent variables as a function, like a general regression model. However, logistic regression models use categorical data, whereas regression models use continuous values as dependent variables. The input data is given as the probability that the results will belong to a particular class, although continuous data can be used.

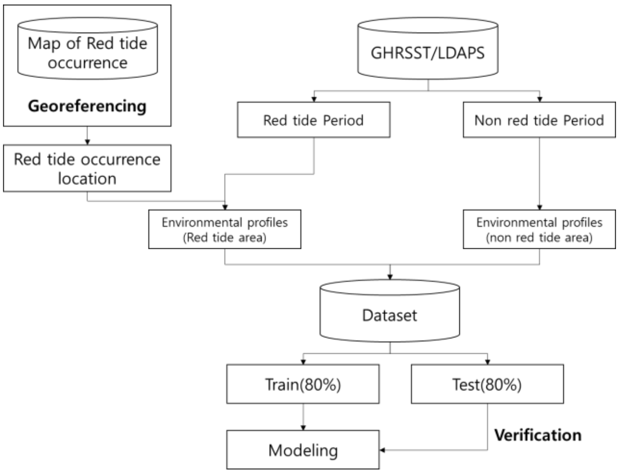
Map of Red tide occurrence
Georeferencing
Red tide occurrence location
GHRSST/LDAPS
Red tide Period
Non red tide Period
Environmental profiles (Red tide area)
Environmental profiles (non red tide area)
Dataset
Train(80%)
Test(80%)
Verification
Modeling

Figure 2. Process of modelling for red tide prediction using machine learning

Decision tree model is a machine learning algorithm that is used for classification or prediction by expressing decision rule as tree shape. Because the process appears as a tree structure, it has the advantage that the researcher can easily understand the process compared to other machine learning methods (Song and Chae, 2008). The decision tree consists of several nodes. Each node starts from the root node and creates child nodes by splitting criterion until each branch reaches the terminal node(Chae et al., 2014). At this time, the splitting criterion uses an input variable and is a criterion for determining which class of the variable to classify. Commonly used splitting criterion algorithms include CHAID (Chi-squared Automatic Interaction Detection) and CART (Classification and Regression Tree, C4.5).

The in depth neural network is a statistical learning algorithm developed by focusing on the signal transduction process of human neurons. The layered neural network is made up of three layers, the input layer, the hidden layer, and the output layer. Generally, the network is designed by adjusting the number of hidden layers and the number of neurons in the hidden layer. The neural network used in this study is a feed-forward neural network modeled in the form of iterative learning using the back-propagation algorithm (Figure 3).

The Deep Neural Network used in this study consists of 4 hidden layers and each hidden layers had 500 nodes with dropout layer. Each layer’s activation function were used ReLU(Rectifier Linear Unit), and last layer’s activation function was used Softmax. Cost function was used Cross entropy for efficiency of compute speed, and used sigmoid cross entropy for overflow due to ReLU.

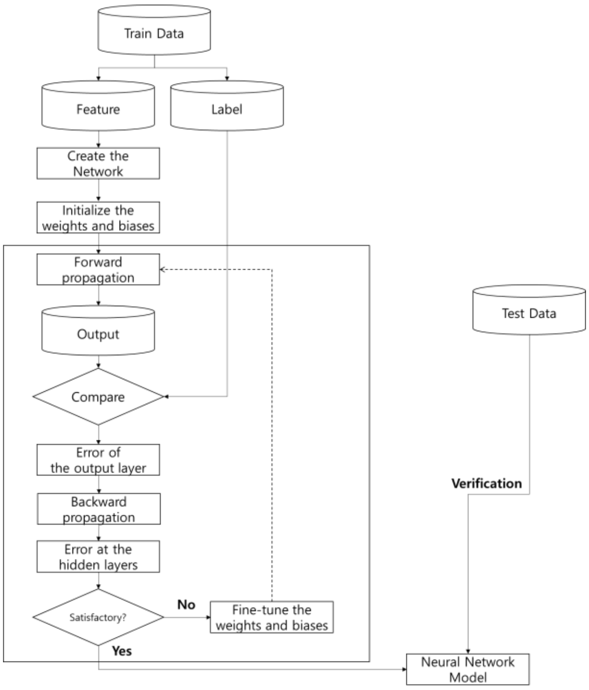

Figure 3. Modeling process of Deep neural network(Feed-forward neural network with Back-propagation algorithm)

3. RESULT AND DISCUSSION

Using the prediction model for *C. polykrikoides* red tide, 40,324 cases(14,476 cases of red tide occurrence and 25,848 cases of no red tide occurrence)were randomly selected from 2013 to 2017, and the result was 84%(Logistic Regression Model) to 99%(Decision Tree, Deep Neural Network) accuracy. In the previous study(Park et al., 2011) based on the neural network model, 75.2% classification accuracy was achieved using monolayer neural network and 78.1% classification accuracy using regression neural network.

However, in this study, about 20% Performance improvement compared previous study. In addition, classification accuracy is higher than SVM(Support Vector Machine) which showed the highest accuracy(86.9%) among the proposed models in previous study. There are two reasons why the proposed model has higher accuracy than the previous studies. First, we used more feature in dataset compared previous study. In previous research, only three characteristics such as water temperature, temperature and precipitation were used.

In this study, more factors such as solar radiation, wind direction and wind speed were used. In addition, the environment was more finely characterized at the time of red tide occurrence using the water temperature-air temperature difference and wind cooling index derived from ocean and weather information. Second, the deep neural network used in this study is capable of finer nonlinear modeling than the single layer neural network used in the previous study. In the previous research, only one hidden layer with 20 nodes was used. However, in this study, more detailed nonlinear modeling was possible than the previous studies using four hidden layers with 500 nodes.

In previous study, Gradiant vanishing problem may occur as the layer of neural network increases by using Sigmoid as an activation function. However, in this study, we could use more hidden layer by using ReLU as an activation function.

In this study, there is a limitation in evaluating the model performance based only on past data. The actual re-evaluation and limitations will be identified through actual operation during the actual red tide period.

ACKNOWLEDGEMENTS (OPTIONAL)

This research was a part of the project titled “Establishment and demonstration of red tide detection and prediction system for minimizing red tide damage” funded by the Ministry of Oceans and Fisheries, Korea (PM60650).

REFERENCES

Kim, H.J., Moon, C.H., Cho, H.J., 2005: Spacial-Temporal Characteristics of Dinoflagellate Cyst Distribution in Sediments of Busan Harbor. *The Sea*, 10(4), 196-203.

Kim, S.S., Go, W.J., Jo, Y.J., Jeon, K.A., 1998: Low Salinity Anomaly and Nutrient Distribution at Surface Waters of the South Sea of Korea during 1996 Summer. The Sea, 3(3), 165-169.

Yoon, Y.H.: Sea rebellion, Red tide. Jipmoondang, Paju, South Korea(2012)

Oh, S.Y., Park, J.M., Yoon, H.J., 2012: Prediction of Red Tide Occurrence by using Oceanic and Atmospheric Data by Satellite, *J. of the Korean institute of Electronic Communication Sciences*, 10(2), 311-318.

Park, S., Kim, K.J., Lee, J.S., Lee, S.R., 2011: Red Tide Prediction using Neural Network and SVM. *J. of the Institute of Electronics Engineers of Korea*, 48SP(5), 651-657.

Bak, S.H., Kim, H.M., Kim, B.K., Hwang, D.H., Unuzaya, E., Yoon, H.J., 2018: Study on Detection Technique for Cochlodinium polykrikoides Red tide using Logistic Regression Model and Decision Tree Model, *J. of the Korean institute of Electronic Communication Sciences*, 13(4), pp. 777-786.

Chae, B., Kim, W., Cho, Y., Kim, K., Lee, C., Choi, Y., 2014:Development of a Logistic Regression Model for Probabilistic Prediction of Debris Flow, *The J. of Engineering Geology*, 14(2), pp. 211-222.

Song, Y., Chae, B., 2008: Development to Prediction Technique of Slope Harzards in Geneiss Area using Decision Tree Model, *The J. of Engineering Geology*, 18(1), pp. 45-54.

Revised August 2019